OCTAVA EDICIÓN

PROBABILIDAD & ESTADÍSTICA

para ingeniería & ciencias

WALPOLE MYERS MYERS YE

PEARSON
Prentice Hall ®

Probabilidad y estadística para ingeniería y ciencias

www.LibrosEnPdf.org

www.LibrosEnPdf.org

Probabilidad y estadística para ingeniería y ciencias

OCTAVA EDICIÓN

Ronald E. Walpole
Roanoke College

Raymond H. Myers
Virginia Polytechnic Institute and State University

Sharon L. Myers
Radford University

Keying Ye
University of Texas at San Antonio

TRADUCCIÓN

Javier Enríquez Brito
Traductor profesional

Victoria Augusta Flores Flores
Traductora profesional

REVISIÓN TÉCNICA

Roberto Hernández Ramírez
Departamento de Matemáticas
Universidad de Monterrey

Marco Antonio Montúfar Benítez
Departamento de Matemáticas
Universidad La Salle Pachuca

José de Jesús Cabrera Chavarría
Julieta Carrasco García
Departamento de Matemáticas
Universidad de Guadalajara

Leticia del Pilar de la Torre
Departamento de Ingeniería Industrial
Instituto Tecnológico de Chihuahua

María de los Ángeles Ramírez Ambriz
Departamento de Ciencias Básicas
Instituto Tecnológico de Ciudad Juárez

Juan de Jesús Blanco Hernández
Facultad de Ingeniería Industrial
Pontificia Universidad Javeriana
Bogotá, Colombia

Eddy Herrera Daza
Departamento de Matemáticas
Pontificia Universidad Javeriana
Bogotá, Colombia

Luis Alejandro Másmela Caita
Departamento de Matemáticas
Universidad Autónoma de Colombia

México • Argentina • Brasil • Colombia • Costa Rica • Chile • Ecuador
España • Guatemala • Panamá • Perú • Puerto Rico • Uruguay • Venezuela

www.LibrosEnPdf.org

Datos de catalogación bibliográfica

Walpole, Ronald E.; Raymond H. Myers; Sharon L. Myers y Keying Ye

Probabilidad y estadística para ingeniería y ciencias
octava edición

Pearson Educación, México, 2007
ISBN: 978-970-26-0936-0
Área: Matemáticas

Formato: 20 × 25.5 cm Páginas: 840

Authorized translation from the English language edition, entitled *Probability and statistics for engineers and scientists* by *Ronald E. Walpole*, *Raymond H. Myers*, *Sharon L. Myers*, *Keying Ye*, published by Pearson Education, Inc., publishing as PRENTICE HALL, INC., Copyright ©2007. All rights reserved.
ISBN 0131877119

Traducción autorizada de la edición en idioma inglés *Probability and statistics for engineers and scientists* por *Ronald E. Walpole*, *Raymond H. Myers*, *Sharon L. Myers*, *Keying Ye*, publicada por Pearson Education, Inc., publicada como PRENTICE-HALL INC., Copyright ©2007. Todos los derechos reservados.

Esta edición en español es la única autorizada.

Edición en español
Editor: Luis Miguel Cruz Castillo
e-mail: luis.cruz@pearsoned.com
Editor de desarrollo: Felipe Hernández Carrasco
Supervisor de producción: Enrique Trejo Hernández

Edición en inglés
Editor en Chief: *Sally Yagan*
Production Editor: *Lynn Savino Wendel*
Senior Managing Editor: *Linda Mihatov Behrens*
Assistant Managing Editor: *Bayani Mendoza de Leon*
Executive Managing Editor: *Kathleen Schiaparelli*
Manufacturing Buyer: *Maura Zaldivar*
Manufacturing Manager: *Alexis Heydt-Long*
Marketing Manager: *Halee Dinsey*
Marketing Assistant: *Jennifer de Leewwerk*
Director of Marketing: *Patrice Jones*
Editorial Assistant/Print Supplements Editor: *Jennifer Urban*
Art Editor: *Thomas Benfatti*
Art Director: *Heather Scott*
Creative Director: *Juan R. López*
Director of Creative Services: *Paul Belfanti*
Cover Photo: *Corbis Royalty Free*
Art Studio: *Laserwords*

OCTAVA EDICIÓN, 2007

D.R. © 2007 por Pearson Educación de México, S.A. de C.V.
Atlacomulco 500-5to. piso
Industrial Atoto
53519, Naucalpan de Juárez, Edo. de México
E-mail: editorial.universidades@pearsoned.com

Cámara Nacional de la Industria Editorial Mexicana. Reg. Núm. 1031.

Prentice Hall es una marca registrada de Pearson Educación de México, S.A. de C.V.

Reservados todos los derechos. Ni la totalidad ni parte de esta publicación pueden reproducirse, registrarse o transmitirse, por un sistema de recuperación de información, en ninguna forma ni por ningún medio, sea electrónico, mecánico, fotoquímico, magnético o electroóptico, por fotocopia, grabación o cualquier otro, sin permiso previo por escrito del editor.

El préstamo, alquiler o cualquier otra forma de cesión de uso de este ejemplar requerirá también la autorización del editor o de sus representantes.

ISBN 10: 970-26-0936-4
ISBN 13: 978-970-26-0936-0

Impreso en México. *Printed in Mexico.*
1 2 3 4 5 6 7 8 9 0 - 10 09 08 07

www.LibrosEnPdf.org

Este libro está dedicado a

Billy y Julie
R.H.M. y S.L.M.

Limin
K.Y.

www.LibrosEnPdf.org

www.LibrosEnPdf.org

Contenido

Prefacio ..... **xv**

1 Introducción a la estadística y al análisis de datos **1**

1.1 Panorama general: Inferencia estadística, muestreo, poblaciones y diseño experimental 1
1.2 El papel de la probabilidad 4
1.3 Procedimientos de muestreo; acopio de los datos 7
1.4 Medidas de posición: La media y la mediana de una muestra 11
Ejercicios 13
1.5 Medidas de variabilidad 14
Ejercicios 17
1.6 Datos discretos y continuos 17
1.7 Modelado estadístico, inspección científica y diagnósticos gráficos 19
1.8 Métodos gráficos y descripción de datos 20
1.9 Tipos generales de estudios estadísticos: Diseño experimental, estudio observacional y estudio retrospectivo 25
Ejercicios 28

2 Probabilidad **31**

2.1 Espacio muestral 31
2.2 Eventos 34
Ejercicios 38
2.3 Conteo de puntos muestrales 40
Ejercicios 47
2.4 Probabilidad de un evento 48
2.5 Reglas aditivas 52
Ejercicios 55
2.6 Probabilidad condicional 58
2.7 Reglas multiplicativas 61
Ejercicios 65

www.LibrosEnPdf.org

2.8 Regla de Bayes ... 68
Ejercicios ... 72
Ejercicios de repaso ... 73

3 Variables aleatorias y distribuciones de probabilidad ... 77
3.1 Concepto de variable aleatoria ... 77
3.2 Distribuciones discretas de probabilidad ... 80
3.3 Distribuciones continuas de probabilidad ... 84
Ejercicios ... 88
3.4 Distribuciones de probabilidad conjunta ... 91
Ejercicios ... 101
Ejercicios de repaso ... 103
3.5 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos ... 106

4 Esperanza matemática ... 107
4.1 Media de una variable aleatoria ... 107
Ejercicios ... 113
4.2 Varianza y covarianza de variables aleatorias ... 115
Ejercicios ... 122
4.3 Medias y varianzas de combinaciones lineales de variables aleatorias ... 123
4.4 Teorema de Chebyshev ... 131
Ejercicios ... 134
Ejercicios de repaso ... 136
4.5 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos ... 138

5 Algunas distribuciones de probabilidad discreta ... 141
5.1 Introducción y motivación ... 141
5.2 Distribución uniforme discreta ... 141
5.3 Distribuciones binomial y multinomial ... 143
Ejercicios ... 150
5.4 Distribución hipergeométrica ... 152
Ejercicios ... 157
5.5 Distribuciones binomial negativa y geométrica ... 158
5.6 Distribución de Poisson y proceso de Poisson ... 161
Ejercicios ... 165
Ejercicios de repaso ... 167
5.7 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos ... 169

www.LibrosEnPdf.org

6 Algunas distribuciones continuas de probabilidad 171
6.1 Distribución uniforme continua 171
6.2 Distribución normal 172
6.3 Áreas bajo la curva normal 176
6.4 Aplicaciones de la distribución normal 182
Ejercicios 185
6.5 Aproximación normal a la binomial 187
Ejercicios 193
6.6 Distribuciones gamma y exponencial 194
6.7 Aplicaciones de las distribuciones exponencial y gamma 197
6.8 Distribución chi cuadrada 200
6.9 Distribución logarítmica normal 201
6.10 Distribución de Weibull (opcional) 202
Ejercicios 205
Ejercicios de repaso 206
6.11 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 209

7 Funciones de variables aleatorias (opcional) 211
7.1 Introducción 211
7.2 Transformaciones de variables 211
7.3 Momentos y funciones generadoras de momentos 219
Ejercicios 226

8 Distribuciones de muestreo fundamentales y descripciones de datos 229
8.1 Muestreo aleatorio 229
8.2 Algunos estadísticos importantes 231
Ejercicios 234
8.3 Presentación de datos y métodos gráficos 236
8.4 Distribuciones muestrales 243
8.5 Distribuciones muestrales de medias 244
Ejercicios 251
8.6 Distribución muestral de S^2 254
8.7 Distribución t 257
8.8 Distribución F 261
Ejercicios 265
Ejercicios de repaso 266
8.9 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 268

www.LibrosEnPdf.org

9 Problemas de estimación de una y dos muestras 269

9.1 Introducción 269
9.2 Inferencia estadística 269
9.3 Métodos clásicos de estimación 270
9.4 Una sola muestra: Estimación de la media 274
9.5 Error estándar de una estimación puntual 280
9.6 Intervalos de predicción 281
9.7 Límites de tolerancia 283
Ejercicios 285
9.8 Dos muestras: Estimación de la diferencia entre dos medias 288
9.9 Observaciones pareadas 294
Ejercicios 297
9.10 Una sola muestra: Estimación de una proporción 299
9.11 Dos muestras: Estimación de la diferencia entre dos proporciones 302
Ejercicios 304
9.12 Una sola muestra: Estimación de la varianza 306
9.13 Dos muestras: Estimación de la razón de dos varianzas 308
Ejercicios 310
9.14 Estimación de la probabilidad máxima (opcional) 310
Ejercicios 315
Ejercicios de repaso 315
9.15 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 319

10 Pruebas de hipótesis de una y dos muestras 321

10.1 Hipótesis estadísticas: Conceptos generales 321
10.2 Prueba de una hipótesis estadística 323
10.3 Pruebas de una y dos colas 332
10.4 Uso de valores P para la toma de decisiones en la prueba de hipótesis 334
Ejercicios 336
10.5 Una sola muestra: Pruebas con respecto a una sola media (varianza conocida) 338
10.6 Relación con la estimación del intervalo de confianza 341
10.7 Una sola muestra: Pruebas sobre una sola media (varianza desconocida) 342
10.8 Dos muestras: Pruebas sobre dos medias 345
10.9 Elección del tamaño de la muestra para probar medias 350
10.10 Métodos gráficos para comparar medias 355
Ejercicios 357
10.11 Una muestra: Prueba sobre una sola proporción 361
10.12 Dos muestras: Pruebas sobre dos proporciones 364
Ejercicios 366
10.13 Pruebas de una y dos muestras referentes a varianzas 367
Ejercicios 370

www.LibrosEnPdf.org

10.14 Prueba de la bondad de ajuste 371
10.15 Prueba de independencia (datos categóricos) 374
10.16 Prueba de homogeneidad 377
10.17 Prueba para varias proporciones 378
10.18 Estudio de caso de dos muestras 380
Ejercicios 383
Ejercicios de repaso 385
10.19 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 387

11 Regresión lineal simple y correlación 389
11.1 Introducción a la regresión lineal 389
11.2 El modelo de regresión lineal simple 390
11.3 Los mínimos cuadrados y el modelo ajustado 394
Ejercicios 397
11.4 Propiedades de los estimadores de los mínimos cuadrados 400
11.5 Inferencias que conciernen a los coeficientes de regresión 402
11.6 Predicción 409
Ejercicios 412
11.7 Selección de un modelo de regresión 414
11.8 El enfoque del análisis de varianza 415
11.9 Prueba para la linealidad de la regresión: Datos con observaciones repetidas 417
Ejercicios 423
11.10 Gráficas de datos y transformaciones 425
11.11 Caso de estudio de regresión lineal simple 430
11.12 Correlación 432
Ejercicios 438
Ejercicios de repaso 438
11.13 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 443

12 Regresión lineal múltiple y ciertos modelos de regresión no lineal 445
12.1 Introducción 445
12.2 Estimación de los coeficientes 446
12.3 Modelo de regresión lineal con el empleo de matrices (opcional) 449
Ejercicios 452
12.4 Propiedades de los estimadores de mínimos cuadrados 456
12.5 Inferencias en la regresión lineal múltiple 458
Ejercicios 464

www.LibrosEnPdf.org

12.6 Selección de un modelo ajustado mediante la prueba de hipótesis. 465
12.7 Caso especial de ortogonalidad (opcional) 469
Ejercicios 473
12.8 Variables categóricas o indicadoras 474
Ejercicios 478
12.9 Métodos secuenciales para la selección del modelo 479
12.10 Estudio de los residuos y trasgresión de las suposiciones (verificación del modelo) . . . 485
12.11 Validación cruzada, x_1, y otros criterios para la selección del modelo 490
Ejercicios 496
12.12 Modelos especiales no lineales para condiciones no ideales 499
Ejercicios de repaso 503
12.13 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 508

13 Experimentos con un solo factor: General 511

13.1 Técnica del análisis de varianza 511
13.2 La estrategia del diseño de experimentos 512
13.3 Análisis de varianza de un solo factor: Diseño completamente al azar (ANOVA de un solo factor) 513
13.4 Pruebas para la igualdad de diversas varianzas 518
Ejercicios 521
13.5 Comparaciones con un grado de libertad 523
13.6 Comparaciones múltiples 527
13.7 Comparación de los tratamientos con un control 531
Ejercicios 533
13.8 Comparación de un conjunto de tratamientos por bloques 535
13.9 Diseños por bloques completamente aleatorios 537
13.10 Métodos gráficos y comprobación del modelo 544
13.11 Transformaciones de los datos en el análisis de varianza 547
13.12 Cuadrados latinos (opcional) 549
Ejercicios 551
13.13 Modelos de efectos aleatorios 555
13.14 Potencia de las pruebas del análisis de varianza 559
13.15 Estudio de caso 563
Ejercicios 565
Ejercicios de repaso 567
13.16 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 571

www.LibrosEnPdf.org

14 Experimentos factoriales (dos o más factores) 573
14.1 Introducción 573
14.2 Interacción en el experimento de dos factores 574
14.3 Análisis de varianza de dos factores 577
Ejercicios 587
14.4 Experimentos con tres factores 590
Ejercicios 597
14.5 Experimentos factoriales de modelos II y III 600
14.6 Elección del tamaño de la muestra 603
Ejercicios 605
Ejercicios de repaso 607
14.7 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 609

15 Experimentos factoriales 2^k y fracciones 611
15.1 Introducción 611
15.2 El factorial 2^k: Cálculo de los efectos y análisis de varianza 612
15.3 Experimento factorial 2^k no replicado 618
15.4 Estudio de caso del moldeo por inyección 619
Ejercicios 622
15.5 Experimentos factoriales en la preparación de la regresión 625
15.6 El diseño ortogonal 631
15.7 Experimentos factoriales en bloques incompletos 639
Ejercicios 645
15.8 Experimentos factoriales fraccionarios 647
15.9 Análisis de los experimentos factoriales fraccionarios 653
Ejercicios 656
15.10 Fracciones superiores y diseños exploratorios 657
15.11 Construcción de diseños con resoluciones III y IV, con 8, 16 y 32 puntos de diseño 658
15.12 Otros diseños de resolución III con dos niveles; los diseños de Plackett-Burman 660
15.13 Diseño de parámetros robustos 661
Ejercicios 666
Ejercicios de repaso 667
15.14 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos 669

16 Estadística no paramétrica 671
16.1 Pruebas no paramétricas 671
16.2 Prueba de rango con signo 676
Ejercicios 679

www.LibrosEnPdf.org

16.3 Prueba de la suma de rangos de Wilcoxon 681
16.4 Prueba de Kruskal-Wallis 684
Ejercicios 686
16.5 Pruebas de corridas 687
16.6 Límites de tolerancia 690
16.7 Coeficiente de correlación de rango 690
Ejercicios 693
Ejercicios de repaso 695

17 Control estadístico de la calidad 697
17.1 Introducción 697
17.2 Naturaleza de los límites de control 699
17.3 Propósitos de la gráfica de control 699
17.4 Gráficas de control para variables 700
17.5 Gráficas de control para atributos 713
17.6 Gráficas de control de cusum 721
Ejercicios de repaso 722

18 Estadística bayesiana (opcional) 725
18.1 Conceptos bayesianos 725
18.2 Inferencias bayesianas 726
18.3 Estimación bayesiana utilizando el contexto de la teoría de decisión 732
Ejercicios 734

Bibliografía 737

A Tablas y pruebas estadísticas 741

B Respuesta a los ejercicios de repaso impares 795

Índice 811

www.LibrosEnPdf.org

Prefacio

Enfoque general y nivel matemático

Los objetivos generales de la octava edición son los mismos que los de las ediciones recientes. Consideramos que es importante conservar el equilibrio entre la teoría y las aplicaciones. Los ingenieros y los físicos, al igual que los especialistas en ciencias de la computación, están capacitados en cálculo, de manera que esta obra se apoya en las matemáticas cuando consideramos que esto enriquece la labor didáctica. Este enfoque impide que el material se convierta en una mera colección de herramientas sin fundamentos matemáticos. Seguramente, los estudiantes con ciertos conocimientos de cálculo y, en algunos casos, en álgebra lineal, tienen la capacidad de entender mejor los conceptos y de utilizar las herramientas resultantes de una forma más inteligente. De lo contrario, se correría el riesgo de que el estudiante sólo sea capaz de aplicar el material dentro de límites muy estrechos.

La nueva edición incluye abundantes ejercicios, los cuales desafían al estudiante a utilizar los conceptos del texto para resolver problemas relacionados con diversas situaciones del campo científico y de la ingeniería. Los datos de los ejercicios están disponibles para descargarse del *companion website* en http:www.pearsoneducacion.net/walpole. El aumento en la cantidad de ejercicios da como resultado un espectro más amplio de áreas de aplicación, que incluyen la ingeniería biomédica, la bioingeniería, los problemas de negocios, diversos temas de computación y muchos otros. Incluso los capítulos relacionados con la introducción a la teoría de la probabilidad contienen ejemplos y ejercicios que tienen un amplio rango de aplicaciones, cuya importancia reconocerán fácilmente los estudiantes de ciencias e ingeniería. Al igual que en ediciones previas, el uso del cálculo se restringe a la teoría elemental de la probabilidad y a las distribuciones de probabilidad. Estos temas se estudian en los capítulos 2, 3, 4, 6 y 7. El capítulo 7 es un capítulo opcional que incluye transformaciones de variables y funciones generadoras de momentos. El álgebra de matrices se utiliza sólo en los capítulos 11 y 12, dedicados a la regresión lineal. Para quienes desean un mayor apoyo en el tema de matrices, tienen a su disposición una sección opcional en el capítulo 12. El profesor que quiera reducir el uso de matrices podría omitir esta sección sin pérdida de continuidad. Los estudiantes que utilicen este texto deben haber completado el equivalente de un semestre de cálculo diferencial e integral. El conocimiento del álgebra de matrices sería útil, aunque no necesario si el contexto del curso excluye la sección opcional del capítulo 12 antes mencionada.

www.LibrosEnPdf.org

Contenido y planeación del curso

Este texto está diseñado para cursos tanto de uno como de dos semestres. Un programa de estudios razonable para un semestre incluiría los capítulos 1 al 10. Muchos profesores desean que los alumnos hayan estudiado en algún grado la regresión lineal simple en un curso de un semestre. En tal caso, podría incluirse una parte del capítulo 11. Por otro lado, algunos profesores desearán abarcar una parte del análisis de varianza, en cuyo caso podrían excluirse los capítulos 11 y 12 a favor de una parte del capítulo 13, que se refiere al análisis de varianza de un factor. Con la finalidad de tener tiempo suficiente para dedicar a uno de estos temas o quizás a los dos, el profesor tal vez quiera eliminar el capítulo 7 y/o ciertos temas especializados de los capítulos 5 y 6 (por ejemplo, las distribuciones gamma, logarítmicas normales y de Weibull, o el material sobre las distribuciones negativa binomial y geométrica). De hecho, algunos profesores consideran que en un curso de un semestre, donde el análisis de regresión y el análisis de varianza son de interés prioritario, deben eliminarse ciertos temas del capítulo 9, dedicado a la estimación (por ejemplo, probabilidad máxima, intervalos de predicción y/o límites de tolerancia). Pensamos que si hay flexibilidad, el profesor podrá establecer las prioridades en un curso de un semestre.

El capítulo 1 ofrece una panorámica elemental de la inferencia estadística diseñada para el principiante. Contiene material sobre el muestreo y el análisis de datos e incluye muchos ejemplos y ejercicios para motivar al alumno. De hecho, algunos aspectos muy rudimentarios del diseño experimental se incluyen junto con una apreciación de técnicas gráficas y ciertas características esenciales de la recolección de datos. Los capítulos 2, 3 y 4 se ocupan de la probabilidad básica, así como de las variables aleatorias discretas y continuas. Los capítulos 5 y 6 se ocupan de las distribuciones discretas y continuas específicas; además, se incluye un número importante de ejemplos y ejercicios con ilustraciones de su uso, destacando las relaciones que hay entre ellos. El capítulo 7 es opcional y se ocupa de la transformación de las variables aleatorias. Tal vez un profesor desee cubrir este material sólo si imparte un curso más teórico. Sin duda, este capítulo es el que incluye más matemáticas de todo el texto. El capítulo 8 contiene material adicional sobre métodos gráficos, así como una introducción de suma relevancia para el estudio de la distribución muestral. Se analizan las gráficas de probabilidad. El material sobre distribución muestral se refuerza con una explicación completa sobre el teorema del límite central, y sobre la distribución de una varianza muestral bajo muestreo normal, idéntica e independientemente distribuido (i.i.d.). Las distribuciones t y F y sus diversos usos se presentan en los capítulos que siguen. Los capítulos 9 y 10 incluyen material sobre uno y dos puntos muestrales, estimación del intervalo y prueba de hipótesis. El material sobre intervalos de confianza, intervalos de predicción, intervalos de tolerancia y estimación de probabilidad máxima en el capítulo 9 ofrece al usuario una flexibilidad considerable en relación con lo que se podría excluir en un curso de un semestre. Se eliminó una sección sobre la estimación de Bayes, que se incluía en el capítulo 9 de la séptima edición. Se prestará más atención a este tema en la sección "Lo nuevo en esta edición", que viene más adelante.

Los capítulos 11 a 17 incluyen abundante material para un segundo semestre. La regresión lineal simple y múltiple se presentan en los capítulos 8 y 12, respectivamente. El capítulo 12 contiene material sobre regresión logística, cuyas aplicaciones son abundantes en las áreas de ingeniería y ciencias biológicas. El material sobre regresión lineal múltiple es muy abundante y permite flexibilidad al profesor. Entre

www.LibrosEnPdf.org

los "temas especiales" a los que el profesor tiene acceso están el caso especial de variables regresoras y ortogonales, categóricas e indicadoras, métodos secuenciales para selección de modelos, estudio de residuos y transgresión de las suposiciones, validación cruzada y el uso de *PRESS* y C_p, y, por supuesto, regresión logística. Los capítulos 13 a 17 incluyen temas sobre análisis de varianza, diseño experimental, estadísticos no paramétricos y control de calidad. El capítulo 15 trata factoriales de dos niveles (con y sin bloqueo) y factoriales fraccionales; una vez más, la flexibilidad se hace presente en los múltiples "temas especiales" que se presentan en este capítulo. Los temas más allá de los diseños estándar 2^k y fraccional 2^k incluyen bloqueo y confusión parcial, fracciones especiales superiores, diseños de Plackett-Burman y diseño de paramétro robusto.

Todos los capítulos incluyen un gran número de ejercicios, muchos más de los que se incluían en la séptima edición. Se detalla más información sobre los ejercicios en la sección "Lo nuevo en esta edición".

Estudios de caso y software

El material sobre prueba de hipótesis de dos muestras, regresión lineal múltiple, análisis de varianza y el uso de experimentos factoriales de dos niveles se complementa con estudios de caso, que presentan las hojas de salida de computadoras y material gráfico. Se incluyen archivos de texto que pueden usarse tanto en *SAS* como *MINITAB*. El uso de hojas de salida de computadora refleja nuestra idea de que los estudiantes deberían tener la experiencia de leer e interpretar los resultados de computadora y las gráficas, incluso si el profesor no utiliza los que se presentan en el texto. La exposición a más de un tipo de software amplía la base de experiencia para el alumno. No hay razón para creer que el software en el curso será el mismo que el estudiante tendrá que usar en su práctica posterior a la graduación. Muchos ejemplos y estudios de caso en el texto se complementan, cuando resulta adecuado, con diversos tipos de gráficas residuales, gráficas de cuantiles, gráficas de probabilidad normal y algunas otras. Esto sucede, sobre todo, en el material utilizado en los capítulos 11 a 15.

Lo nuevo en esta edición

En general

1. Se agregó entre un 15 y 20% de problemas nuevos, con muchas aplicaciones recientes demostradas en ingeniería, así como en las ciencias biológicas, físicas y de la computación.
2. Hay material nuevo y de repaso al final de cada capítulo, donde resulte apropiado. Este material destaca las ideas clave, así como los riesgos y peligros de los que debe estar consciente el usuario del material que se estudia en el capítulo. Esta sección también brinda la demostración de cómo el material presentado se relaciona con el material de otros capítulos.
3. Se incorporó un nuevo (y opcional) minicapítulo sobre la estadística bayesiana. El capítulo presenta material práctico con aplicaciones en muchos campos.
4. Hay otros cambios importantes a lo largo de la obra, con base en lo que los autores y revisores percibieron. A continuación se describen de manera específica algunos de tales cambios.

www.LibrosEnPdf.org

Capítulo 1: Introducción a la estadística y al análisis de datos

El capítulo 1 presenta una cantidad significativa de material novedoso. Hay una nueva explicación sobre la diferencia entre medidas discretas y continuas. Muchos ejemplos se presentan con aplicaciones específicas de las medidas discretas en la vida real (por ejemplo, los números de partículas radiactivas, el número de personal responsable de una instalación portuaria particular y el número de buques petroleros que llegan cada día a un puerto). Se presta especial atención a las situaciones asociadas con datos binarios. Se dan ejemplos del campo biomédico y del control de calidad.

Se analizan nuevos conceptos (para este texto) en el capítulo 1, en relación con las propiedades de una distribución o una muestra, además de aquellas que caracterizan la tendencia central y la variabilidad. Se definen y analizan los cuartiles y, más generalmente, los cuantiles.

Con respecto a la séptima edición, se amplió la explicación sobre la importancia del diseño experimental y las ventajas que ofrece. En este desarrollo se tratan importantes nociones, que incluyen aleatorización, reducción de variabilidad en el proceso y la interacción entre factores.

Los lectores se enfrentan en este capítulo a diferentes tipos de estudios estadísticos: el diseño experimental, el estudio observacional y el estudio retrospectivo. Se dan ejemplos de cada tipo de estudio, y se analizan sus ventajas y desventajas. El capítulo continúa con énfasis en los procedimientos gráficos y sus campos de aplicación.

Se agregaron 19 nuevos ejercicios al capítulo 1. Algunos emplean datos de los estudios realizados en el centro de consulta del Tecnológico de Virginia Tech, otros se tomaron de publicaciones especializadas en ingeniería, y otros más incluyen datos históricos. Este capítulo contiene ahora 30 ejercicios.

Capítulo 2: Probabilidad

Hay nuevos ejemplos y una nueva explicación para ilustrar mejor la noción de la probabilidad condicional. El capítulo 2 ofrece un total de 136 ejercicios. Todos los ejercicios nuevos implican aplicaciones directas en ciencias y en ingeniería.

Capítulo 3: Variables aleatorias y distribuciones de probabilidad

Hay una nueva explicación sobre la noción de variables "dummy", que juegan un rol importante en las distribuciones de Bernoulli y binomial. Hay mucho más ejercicios con nuevas aplicaciones. La sección de repaso al final del capítulo destaca la relación entre el material del capítulo 3 con el concepto de parámetros de distribución y distribuciones de probabilidad específica, que se estudian en capítulos posteriores.

Los temas para los nuevos ejercicios incluyen la distribución del tamaño de las partículas para el combustible de misiles, errores de medición en sistemas científicos, estudios sobre el tiempo que tardan las lavadoras en presentar fallas, la producción de tubos de electrones en una línea de ensamble, problemas de tiempo de llegada a ciertas intersecciones en las grandes ciudades, la vida de un producto en el anaquel, problemas de congestionamiento de pasajeros en los aeropuertos, problemas con las impurezas en lotes de productos químicos, fallas en sistemas de componentes electrónicos que trabajan en paralelo, entre muchos otros. Ahora hay 82 ejercicios en este capítulo.

www.LibrosEnPdf.org

Capítulo 4: Esperanza matemática

Se agregaron varios ejercicios más al capítulo 4. Las reglas para las expectativas y las varianzas de funciones lineales se ampliaron para cubrir aproximaciones de funciones no lineales. Se ofrecen ejemplos para ilustrar el uso de estas reglas. El repaso al final del capítulo 4 revela posibles dificultades y riesgos con las aplicaciones prácticas del material, ya que la mayoría de los ejemplos y ejercicios suponen que los parámetros (media y varianza) se conocen; mientras que en las aplicaciones reales estos parámetros serían estimados. Se hace referencia al capítulo 9, donde se estudia la estimación. Ahora hay 103 ejercicios en este capítulo.

Capítulo 5: Algunas distribuciones de probabilidad discreta

Se agregaron nuevos ejercicios que representan las diversas aplicaciones de la distribución de Poisson. También se presenta una explicación adicional sobre la función de probabilidad de Poisson.

Se incluyen nuevos ejercicios de aplicaciones en la vida real de las distribuciones de Poisson, binomial e hipergeométrica. Los temas para los nuevos ejercicios se refieren a los defectos en cables de cobre, baches en las carreteras que requieren reparación, tráfico de pacientes en hospitales urbanos, inspección del equipaje en aeropuertos, sistemas de seguridad en tierra para detección de misiles y muchos otros. Además, se presentan gráficas que ofrecen al lector una clara indicación acerca de la naturaleza de las distribuciones de Poisson y binomial conforme cambian los parámetros. En este capítulo hay ahora 105 ejercicios.

Capítulo 6: Algunas distribuciones continuas de probabilidad

Se agregaron muchos más ejemplos y ejercicios referentes a la distribución exponencial y gamma. La propiedad de "falta de memoria" de la distribución exponencial ahora se explica de manera extensa y en relación con el vínculo entre las distribuciones exponencial y de Poisson. La sección sobre la distribución de Weibull se mejoró y amplió considerablemente. Las extensiones presentadas se enfocan en la medición e interpretación de la tasa de falla o "tasa de riesgo", y en cómo el conocimiento de los parámetros de Weibull permiten al usuario aprender la forma en que las máquinas se desgastan o incluso se vuelven más resistentes con el paso del tiempo. Se presentan más ejercicios en relación con las distribuciones de Weibull y logarítmica normal. Al igual que en el capítulo 5, en la sección de repaso se advierte que hay que tener cuidado en ciertos casos. En situaciones prácticas, las suposiciones o estimaciones de los parámetros de proceso de la distribución gamma en los problemas relacionados con la tasa de falla, por ejemplo, o en los parámetros de una distribución gamma o de Weibull, podrían ser inestables, lo que da lugar a errores en los cálculos. Ahora hay 84 ejercicios en total en este capítulo.

Capítulo 7: Funciones de variables aleatorias (opcional)

No se realizaron cambios fundamentales en este capítulo opcional.

Capítulo 8: Distribuciones de muestreo fundamentales y descripción de datos

Se incluye una explicación adicional sobre el teorema del límite central, así como sobre el concepto general de distribuciones de muestreo. Hay muchos nuevos ejercicios.

El resumen brinda información importante sobre t, χ^2 y F, incluyendo la forma como se emplean y las suposiciones implicadas.

En este capítulo se presta mayor atención a la elaboración de gráficas de probabilidad normal. Además, se explica el teorema del límite central con mayor detalle, de manera que el lector entienda mejor el tamaño que debe tener n antes de buscar la normalidad. Se presentan gráficas para ilustrar esta situación.

Se da una exposición adicional en relación con la aproximación normal a la distribución binomial y cómo opera en situaciones prácticas. La presentación incluye un argumento intuitivo que vincula la aproximación normal de la binomial con el teorema del límite central. El número de ejercicios en este capítulo ahora es de 75.

Capítulo 9: Problemas de estimación de una y dos muestras

En los nuevos ejercicios se presentan muchas aplicaciones recientes de este capítulo. El resumen explica la razón fundamental y los riesgos asociados con el llamado intervalo de confianza de muestra grande. Se explica la importancia de la suposición de normalidad y las condiciones en las cuales se realiza.

Al principio de este capítulo, el desarrollo de los intervalos de confianza ofrece una explicación pragmática acerca de por qué uno debe comenzar con el caso de "σ conocida". Se sugiere que este tipo de situaciones no ocurren verdaderamente en la práctica, pero la consideración del caso de σ conocida, en principio, ofrece una estructura que permite que los estudiantes comprendan el caso más útil de "σ desconocida".

Los límites unilaterales de todos los tipos se presentan aquí y se da una explicación sobre cuándo se les utiliza como opuestos a sus contrapartes bilaterales. Se presentan nuevos ejemplos que requieren del uso de intervalos unilaterales. Éstos incluyen los intervalos de confianza, de predicción y de tolerancia. Se explica el concepto de error cuadrado medio de un estimador. De esta forma, es posible concentrar la noción de sesgo y de varianza en la comparación general de los estimadores. Se incluyen 27 nuevos ejercicios en el capítulo 9, y en total se presentan 111.

Capítulo 10: Pruebas de hipótesis de una y dos muestras

Se presenta una exposición enteramente reestructurada sobre la introducción a la prueba de hipótesis. Se diseñó para ayudar al estudiante a tener una visión clara de qué es lo que se realiza y qué no en una prueba de hipótesis. La noción de que rara vez, si es que acaso, "aceptamos la hipótesis nula" se analiza con la ayuda de ilustraciones. También se presenta una explicación completa con ejemplos, acerca de cómo se deberían estructurar o establecer la hipótesis nula y la alternativa. La noción de que el rechazo implica que la "evidencia de muestra refuta H_0" y de que H_0 es en realidad el complemento lógico de H_1 se analiza de manera precisa con la ayuda de varios ejemplos. Se discute mucho acerca del concepto de "no rechace H_0" y sobre lo que significa en situaciones prácticas. El resumen se refiere a "concepciones errónes y riesgos", lo cual revela problemas en establecer conclusiones equivocadas cuando el analista "no rechaza" la hipótesis nula. Además, se analiza la "robustez", que tiene que ver con la naturaleza de la sensibilidad de diversas pruebas de hipótesis para la suposición de normalidad. Ahora se incluyen 115 ejercicios en este capítulo.

Capítulo 11: Regresión lineal simple y correlación

Se agregaron muchos nuevos ejercicios sobre regresión lineal simple. Se da una explicación especial sobre los errores en el uso de R^2, el coeficiente de determinación. Se pone un interés adicional en las gráficas y el diagnóstico en relación con la regresión. El resumen se ocupa de los riesgos que uno encuentra si no se utilizan los diagnósticos. Se destaca que estos últimos proveen "verificaciones" sobre la validez de las suposiciones. Los diagnósticos incluyen gráficas de datos, gráficas de residuos studentizados y gráficas de probabilidad normal de residuos.

Al principio del capítulo se hace una importante presentación acerca de la naturaleza de los modelos lineales en ciencias y en ingeniería. Se señala que éstos, con frecuencia, constituyen modelos empíricos que son simplificaciones de estructuras más complejas y desconocidas.

En este capítulo se pone mayor énfasis en la elaboración de gráficas de datos. La "regresión a través del origen" se explica en un ejercicio. Se amplía la explicación sobre lo que significa que H_0: $\beta = 0$ se rechace o no. Se emplean gráficas para ilustrar los casos. Ahora se incluyen 68 ejercicios en este capítulo.

Capítulo 12: Regresión lineal múltiple y ciertos modelos de regresión lineal

En este capítulo se da un tratamiento adicional a los problemas en el uso de R^2. La discusión se centra alrededor de la necesidad de transigir entre el intento por alcanzar un "buen ajuste" para los datos y la pérdida inevitable en grados de libertad del error que se experimenta cuando se "sobreajusta". Al respecto, la "R^2 ajustada" se define y explica mediante ejemplos. Además, el coeficiente de variación (*cv*) se analiza y se interpreta como una medida que resulta útil para comparar modelos en competencia. Se presentan varios nuevos ejercicios para brindar al lector experiencia en la comparación de modelos en competencia utilizando conjuntos de datos reales. Se da un tratamiento adicional al tema de "regresores categóricos" con herramientas gráficas utilizadas para apoyar los conceptos implicados. Se incluyen ejercicios adicionales para ilustrar los usos prácticos de la regresión logística, tanto en el área industrial como en la investigación biomédica. Ahora se tienen 72 ejercicios en este capítulo.

Capítulo 13: Experimentos de un solo factor: General

La explicación de la prueba de Tukey sobre las comparaciones múltiples se amplió considerablemente. Se estudia más material sobre la noción de tasa de error y los valores α en el contexto de los intervalos de confianza simultáneos.

Se presenta una nueva e importante sección sobre la "Transformación de los datos en el análisis de varianza". Se hace un contraste con la explicación en los capítulos 11 y 12 en relación con la transformación para producir un buen ajuste en la regresión. Se incluye una breve presentación sobre la robustez del análisis de varianza para la suposición de varianza homogénea. Esta explicación se relaciona con las secciones anteriores sobre las gráficas de diagnóstico para detectar violaciones en las suposiciones.

Se hace una mención adicional sobre las causas fundamentales de la transgresión de la suposición de varianza homogénea y sobre cómo a menudo es una ocurrencia natural, cuando la varianza es una función de la media. Las transformaciones se discuten de tal manera que pueden utilizarse para dar cabida al problema. Se dan ejemplos y ejercicios para ilustrar. Se agregaron varios nuevos ejercicios, para llegar a un total de 67.

Capítulo 14: Experimentos factoriales (dos o más factores)

Desde el inicio de este capítulo se da considerable atención al concepto de interacción y a las gráficas de interacción. Se presentan ejemplos donde las interpretaciones científicas de la interacción se dan utilizando gráficas. Nuevos ejercicios ilustran el uso de gráficas, incluidas las de diagnóstico de residuos. Varios nuevos ejercicios aparecen en este capítulo. Todos incluyen datos experimentales tomados de las ciencias químicas y biológicas, donde se destaca el análisis gráfico. Hay 43 ejercicios en total.

Capítulo 15: Experimentos factoriales 2^k y fracciones

Desde el inicio de este capítulo se agrega nuevo material para destacar e ilustrar el papel de los diseños de dos niveles como experimentos de investigación. Éstos a menudo son parte de un plan secuencial, en el cual el científico o ingeniero intenta aprender acerca del proceso, evaluar el papel de los factores implicados y generar conocimiento que ayude a determinar la región más fructífera de experimentación. La noción de los diseños fraccionales factoriales se desarrolla desde el principio del capítulo.

La noción de "efectos" y los procedimientos gráficos que se utilizan para determinar los "efectos activos" se estudian con mayor detalle usando ejemplos. El capítulo utiliza considerablemente más ilustraciones gráficas y demostraciones geométricas para generar los conceptos tanto para los factoriales enteros como para los fraccionales. Además, los gráficos se utilizan para ilustrar la información disponible sobre falta de ajuste, cuando uno aumenta el diseño de dos niveles con corridas centrales.

En el desarrollo y discusión de los diseños factoriales fraccionales, el procedimiento para construir la fracción se simplificó de forma considerable y se diseñó de tal forma que apela mucho más a la intuición. Las "columnas agregadas" que se seleccionan de acuerdo con la estructura deseada se utilizan con varios ejemplos. Pensamos que el lector ahora logrará obtener una mejor comprensión de lo que se gana (y se pierde) con el uso de las fracciones. Esto representa una simplificación fundamental con respecto a la edición anterior. Por primera vez, se presenta una tabla sustancial que permite al lector construir diseños de dos niveles con resolución III y IV. Se agregaron 18 nuevos ejercicios a este capítulo, para dar un total de 50.

Capítulo 16: Estadística no paramétrica

No se realizaron cambios fundamentales. El número total de ejercicios es de 41.

Capítulo 17: Control estadístico de la calidad

No se realizaron cambios fundamentales. El número total de ejercicios es de 10.

Capítulo 18: Estadística bayesiana (opcional)

Este capítulo es completamente nuevo en la octava edición. El material sobre estadística bayesiana en la séptima edición (que se incluía en el capítulo 9) se eliminó para presentar este tema en un capítulo especial.

El capítulo trata los elementos pragmáticos y sumamente útiles de la estadística bayesiana, sobre los que los estudiantes de ciencias e ingeniería deberían tener conocimiento. El capítulo presenta el importante concepto de la probabilidad subjetiva

en conjunción con la noción de que, en muchas aplicaciones, los parámetros poblacionales son verdaderamente inconstantes, aunque deben tratarse como variables aleatorias. La estimación puntal y por intervalos se estudia desde un punto de vista bayesiano, y se presentan ejemplos prácticos. Este capítulo es relativamente corto (10 páginas) y contiene 9 ejemplos y 11 ejercicios.

Agradecimientos

Nos sentimos en deuda con aquellos colegas que revisaron las ediciones anteriores de este libro y que hicieron muchas sugerencias útiles para esta edición. Ellos son: Andre Adler, *Illinois Institute of Technology*; Georgiana Baker, *University of South Carolina*; Barbara Bennie, *University of Minnesota*; Nirmal Devi, *Embry Riddle*; Ruxu Du, *University of Miami*; Stephanie Edwards, *Bemidji State University*; Charles McAllister, *Louisiana State University*; Judith Miller, *Georgetown University*; Timothy Raymond, *Bucknell University*; Dennis Webster, *Louisiana State University*; Blake Whitten, *University of Iowa*; Michael Zabarankin, *Stevens Institute of Technology*.

Queremos agradecer los servicios editoriales y de producción que brindaron numerosas personas en Prentice Hall, especialmente la editora en jefe, Sally Yagan, la editora de producción, Lynn Savino Wendel, y la editora de publicaciones, Patricia Daly. Apreciamos profundamente los numerosos y útiles comentarios, sugerencias y lecturas de pruebas de Richard Charnigo, Jr., Michael Anderson, Joleen Beltrami y George Lobell. Agradecemos al Virginia Tech Statistical Consulting Center, por ser la fuente de muchos conjuntos de datos de la vida real. Además, agradecemos a Linda Douglas, quien trabajó arduamente para ayudarnos en la preparación del manuscrito.

R.H.M.
S.L.M.
K.Y.

Capítulo 1

Introducción a la estadística y al análisis de datos

1.1 Panorama general: Inferencia estadística, muestreo, poblaciones y diseño experimental

Desde inicios de la década de 1980 y hasta la actualidad, se ha puesto un interés especial en el *mejoramiento de la calidad* en la industria estadounidense y de todo el mundo. Se ha dicho y escrito mucho acerca del "milagro industrial" japonés que comenzó a mediados del siglo XX. Los nipones fueron capaces de tener éxito donde otras naciones fallaron; a saber, en la creación de un entorno que permita la manufactura de productos de alta calidad. Gran parte del éxito japonés se atribuye al uso de *métodos estadísticos* y del pensamiento estadístico entre el personal gerencial.

Empleo de datos científicos

El uso de métodos estadísticos en la manufactura, el desarrollo de productos alimenticios, el software para computadoras, los medicamentos y muchas otras áreas implican el acopio de información o **datos científicos**. Por supuesto que la obtención de datos no es algo nuevo, ya que se ha realizado por más de mil años. Los datos se han recabado, resumido, reportado y almacenado para su examen cuidadoso. Sin embargo, hay una diferencia profunda entre recabar información científica y la **estadística inferencial**. Esta última ha recibido atención legítima durante las últimas décadas.

La estadística inferencial generó un número enorme de "herramientas" de métodos estadísticos que utilizan los profesionales de la estadística. Los métodos estadísticos se diseñan para contribuir al proceso de realizar juicios científicos frente a la **incertidumbre** y a la **variación**. Dentro del proceso de manufactura la densidad de producto de un material específico no siempre será la misma. De hecho, si se trata de un proceso discontinuo en vez de uno continuo, habrá variación en la densidad de material no sólo entre los lotes (variación de un lote a otro) que salen de la línea de producción, sino también dentro de ellos. Los métodos estadísticos se utilizan para analizar datos de procesos como el anterior, para tener una mejor orientación respecto de dónde realizar mejoras a la **calidad** del proceso mismo. Aquí la calidad

podría definirse según su cercanía con el valor de la densidad meta en relación con *la proporción de las veces* que se cumple tal criterio de cercanía. A un ingeniero podría interesarle un instrumento específico que se utilice para medición del monóxido de azufre en estudios sobre la contaminación atmosférica. Si el ingeniero tiene duda respecto de la eficacia del instrumento, hay dos **fuentes de variación** con las cuales debe despejarla. La primera es la variación en los valores de monóxido de azufre que se encuentran en el mismo lugar el mismo día. La segunda es la variación entre los valores observados y el monóxido de azufre **real** que haya en el aire en ese momento. Si cualquiera de ambas fuentes de variación es extraordinariamente grande (según algún estándar determinado por el ingeniero), quizá se necesite reemplazar el instrumento. En un estudio biomédico de un nuevo fármaco que reduce la hipertensión, 85% de los pacientes experimentaron alivio; mientras que se reconoce que, por lo general, el medicamento "viejo" o actual alivia a 80% de los pacientes que sufren hipertensión crónica. No obstante, el nuevo fármaco es más caro de elaborar y quizás ocasione algunos efectos colaterales. ¿Debería adoptarse el nuevo medicamento? Se trata de un problema que a menudo se encuentra (a veces con mucha mayor complejidad) en la relación entre las empresas farmacéuticas y la FDA (Federal Drug Administration). De nuevo, necesita tomarse en cuenta la variación. El valor de 85% se basa en cierto número de pacientes seleccionados para el estudio. Tal vez si se repitiera el estudio con nuevos pacientes ¡el número observado de "éxitos" sería de 75%! Se trata de una variación natural de un estudio a otro que debe tomarse en cuenta para el proceso de toma de decisiones. Es evidente que tal variación es importante porque una variación de un paciente a otro es endémica al problema.

Variabilidad en los datos científicos

En los problemas discutidos anteriormente los métodos estadísticos empleados tienen que ver con la variabilidad y en cada caso la variabilidad que se estudia se encuentra en datos científicos. Si la densidad del producto observada en el proceso es siempre la misma y siempre es la esperada, no habría necesidad de métodos estadísticos. Si el dispositivo para medir el monóxido de azufre siempre diera el mismo valor y éste fuera exacto (es decir, correcto), no se requeriría análisis estadístico. Si no hubiera variabilidad de un paciente a otro inherente a la respuesta al medicamento (es decir, si siempre el fármaco causara alivio o no), la vida sería muy sencilla para los científicos de la industria farmacéutica y para la FDA y los estadísticos no serían necesarios en el proceso de toma de decisiones. La estadística inferencial ha originado un gran número de métodos analíticos que permiten efectuar análisis de datos obtenidos de sistemas como los que se describen anteriormente, lo cual refleja la verdadera naturaleza de la ciencia que conocemos como estadística inferencial; a saber, el uso de técnicas que nos permiten ir más allá de sólo reportar datos, ya que nos permiten obtener conclusiones (o inferencias) sobre el sistema científico. Los estadísticos usan leyes fundamentales de probabilidad e inferencia estadística para sacar conclusiones respecto de los sistemas científicos. La información se colecta en forma de **muestras**, o agrupaciones de **observaciones**. En el capítulo 2 se introduce el proceso de muestreo, cuyo estudio continúa a lo largo de todo el libro.

Las muestras se reúnen a partir de **poblaciones**, que son agrupaciones de todos los individuos o elementos individuales de un tipo específico. A veces una población representa un sistema científico. Por ejemplo, un fabricante de tarjetas para computadora quizá desee eliminar defectos. Un proceso de muestreo implicaría la recolección de información de 50 tarjetas de computadora tomadas aleatoriamente durante el proceso. Aquí, la población serían todas las tarjetas de computadora pro-

ducidas por la empresa en un periodo específico. En un experimento con fármacos, se toma una muestra de pacientes y a cada uno se le administra un medicamento específico para reducir la presión sanguínea. El interés se enfoca en la obtención de conclusiones sobre la población de quienes sufren hipertensión. Si se logra una mejoría en el proceso de producción de las tarjetas para computadora y se reúne una segunda muestra de tarjetas, cualesquiera conclusiones que se obtengan respecto de la efectividad del cambio en el proceso debería extenderse a toda la población de tarjetas para computadora que se produzcan bajo el "proceso mejorado".

A menudo, es muy importante el acopio de datos científicos en forma sistemática, cuando la planeación ocupa un lugar importante en la agenda. En ocasiones la planeación está, por necesidad, bastante limitada. Con frecuencia nos enfocamos en ciertas propiedades o características de los elementos u objetos de la población. Tal característica tiene importancia de ingeniería específica o, digamos, biológica para el "cliente": el científico o el ingeniero que busca aprender algo acerca de la población. Por ejemplo, en uno de los casos anteriores, la calidad del proceso tenía relación con la densidad del producto cuando sale del proceso. Un ingeniero podría necesitar estudiar el efecto de las condiciones del proceso, la temperatura, la humedad, la cantidad de un ingrediente particular, etcétera. Él o ella quizá muevan de manera sistemática estos **factores** a cualesquiera niveles que se sugieran, de acuerdo con cualquier prescripción o **diseño experimental** que se desee. Sin embargo, un científico silvicultor que está interesado en un estudio de los factores que influyen en la densidad de la madera en cierta clase de árbol no necesariamente tiene que diseñar un experimento. En este caso quizá requiera un **estudio observacional**, en el cual los datos se acopien en el campo, pero no se pueden seleccionar de antemano los **niveles de los factores**. Ambos tipos de estudios se prestan a los métodos de la inferencia estadística. En el primero, la calidad de las inferencias dependerá de la planeación adecuada del experimento. En el último, el científico está a expensas de lo que pueda recopilar. Por ejemplo, resulta inadecuado si un agrónomo se interesa en estudiar el efecto de la lluvia sobre la producción de plantas y los datos se obtienen durante una sequía.

Es necesario entender la importancia del pensamiento estadístico para los administradores y el uso de la inferencia estadística para el personal científico. Los investigadores obtienen mucho de los datos científicos. Los datos brindan una comprensión del fenómeno científico. Los ingenieros de producto y de procesos aprenden más en sus esfuerzos fuera de línea para mejorar el proceso. También logran una comprensión valiosa al reunir datos de producción (monitoreo *on line*) con una base regular, lo cual permite la determinación de las modificaciones necesarias con la finalidad de mantener el proceso en el nivel de calidad deseado.

En ocasiones un científico sólo desea obtener alguna clase de resumen del conjunto de datos representados en la muestra. En otras palabras, no utiliza la estadística inferencial. En cambio, le serían útiles un conjunto de estadísticos o **estadística descriptiva**. Tales números ofrecen un sentido del centro de ubicación de los datos, de la variabilidad en los datos y de la naturaleza general de la distribución de observaciones en la muestra. Aunque no se incorporen métodos estadísticos específicos que lleven a la **inferencia estadística**, se puede aprender mucho. A veces la estadística descriptiva va acompañada por gráficas. El software estadístico moderno permite el cálculo de **medias**, **medianas**, **desviaciones estándar** y otros estadísticos, así como el desarrollo de gráficas que presenten una "huella digital" de la naturaleza de la muestra. En las secciones siguientes veremos definiciones e ilustraciones de los estadísticos y descripciones de recursos gráficos como histogramas, diagramas de tallo y hojas, y diagramas de punto y de caja.

1.2 El papel de la probabilidad

En este libro, los capítulos 2 a 6 tratan de las nociones fundamentales de la probabilidad. Un estudio esmerado de las bases de tales conceptos permitirá al lector lograr una mejor comprensión de la inferencia estadística. Sin algo de formalismo en probabilidad, el estudiante no sería capaz de apreciar la verdadera interpretación del análisis de datos a través de los métodos estadísticos modernos. Es completamente natural estudiar probabilidad antes de estudiar inferencia estadística. Los elementos de probabilidad nos permiten cuantificar la fortaleza o "confianza" de nuestras conclusiones. Entonces, los conceptos de probabilidad forman un componente significativo que complementa los métodos estadísticos y ayuda a evaluar la consistencia de la inferencia estadística. Por consiguiente, la disciplina de la probabilidad brinda la transición entre la estadística descriptiva y los métodos inferenciales. Los elementos de la probabilidad permiten que la conclusión se exprese en un lenguaje que requieren los científicos y los ingenieros. El ejemplo que sigue permite al lector comprender la noción de un valor-P, el cual a menudo da el "fundamento" de la interpretación de los resultados a partir del uso de los métodos estadísticos.

Ejemplo 1.1: Suponga que un ingeniero se encuentra con datos de un proceso de producción donde se muestrean 100 artículos y se obtienen 10 defectuosos. Se espera que de cuando en cuando haya artículos defectuosos. En efecto, los 100 artículos representan la muestra. Sin embargo, se determina que, a largo plazo, la empresa sólo puede tolerar 5% de artículos defectuosos en el proceso. Entonces, los elementos de probabilidad permiten al ingeniero determinar qué tan concluyente es la información muestral respecto de la naturaleza del proceso. En este caso, la **población** representa conceptualmente todos los artículos posibles en el proceso. Suponga que averiguamos que *si el proceso es aceptable*, es decir, si produce artículos con sólo 5% defectuosos, hay una probabilidad de 0.0282 de obtener 10 o más artículos defectuosos en una muestra aleatoria de 100 artículos del proceso. Esta pequeña probabilidad sugiere que el proceso, en realidad, tiene un porcentaje de artículos defectuosos en el largo plazo que excede 5%. En otras palabras, en condiciones de un proceso aceptable, la información muestral que se obtuvo casi nunca ocurriría. No obstante, ¡en verdad ocurrió! Claramente, sin embargo, ocurriría con una probabilidad mucho mayor si la tasa de artículos defectuosos del proceso excediera 5% por un monto significativo.

De este ejemplo es evidente que los elementos de probabilidad ayudan en la traducción de información muestral en algo concluyente o no concluyente acerca del sistema científico. De hecho, probablemente lo que se aprendió constituye información inquietante para el ingeniero o administrador. Los métodos estadísticos (que examinaremos con más detalle en el capítulo 10) produjeron un valor-P de 0.0282. El resultado sugiere que el proceso **muy probablemente no sea aceptable**. En los capítulos siguientes se trata detenidamente el concepto de **valor-*P***. El ejemplo que sigue brinda una segunda ilustración.

Ejemplo 1.2: Con frecuencia la naturaleza del estudio científico señalará el papel que juegan la probabilidad y el razonamiento deductivo en la inferencia estadística. El ejercicio 9.40 en la página 297 proporciona datos asociados con un estudio que se llevó a cabo en el Instituto Politécnico y Universidad Estatal de Virginia, acerca del desarrollo de una relación entre las raíces de los árboles y la acción de un hongo. Se transfirieron minerales de los hongos a los árboles, y azúcares de los árboles al hongo. Se plantaron dos muestras de 10 plantones de roble rojo norteño en un invernadero: una que

contenía plantones tratados con nitrógeno y una muestra de plantones sin tratamiento. Todas las demás condiciones ambientales se mantuvieron constantes. Todos los plantones contenían el hongo *Pisolithus tinctorus*. En el capítulo 9 se incluyen más detalles. Los pesos en gramos de los tallos se registraron al finalizar 140 días. Los datos se presentan en la tabla 1.1.

Tabla 1.1: Conjunto de datos del ejemplo 1.2

Sin nitrógeno	**Con nitrógeno**
0.35	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

En este ejemplo hay dos muestras tomadas de dos **poblaciones distintas**. La finalidad del experimento consiste en determinar si el uso del nitrógeno tiene influencia sobre el crecimiento de las raíces. Se trata de un estudio comparativo (es decir, se busca comparar las dos poblaciones en cuanto a ciertas características importantes). Es conveniente graficar los datos como se indica en la figura 1.1. Los valores $\circ$ representan los datos "con nitrógeno" y los valores $\times$ representan los datos "sin nitrógeno". Así, el propósito de este experimento es determinar si el uso de nitrógeno tiene influencia en el crecimiento de las raíces. Note que la apariencia general de los datos podría sugerir al lector que, en promedio, el uso del nitrógeno aumenta el peso del tallo. Cuatro observaciones con nitrógeno son considerablemente más grandes que cualquiera de las observaciones sin nitrógeno. La mayoría de las observaciones sin nitrógeno parece estar por debajo del centro de los datos. La apariencia del conjunto de datos parecería indicar que el nitrógeno es efectivo. Pero, ¿cómo se cuantifica esto? ¿Cómo se resume toda la evidencia visual aparente con algún significado? Como en el ejemplo anterior, se pueden utilizar los fundamentos de la probabilidad. Las conclusiones se resumen en una declaración de probabilidad o valor-*P*. Aquí no demostraremos la inferencia estadística que produce la probabilidad resumida. Como en el ejemplo 1.1, tales métodos se estudiarán en el capítulo 10. El problema gira alrededor de la "probabilidad de que datos como éstos se puedan observar", *dado que el nitrógeno no tiene efecto*; en otras palabras, puesto que ambas muestras se generaron a partir de la misma población. Suponga que esta probabilidad es pequeña, digamos de 0.03; ésta sería con certeza suficiente evidencia de que el uso del nitrógeno en realidad influye (aparentemente lo aumenta) en el peso promedio del tallo en los plantones de roble rojo.

¿Cómo trabajan juntas la probabilidad y la inferencia estadística?

Para el lector es importante distinguir claramente entre la disciplina de la probabilidad, una ciencia por derecho propio, y la disciplina de la estadística inferencial.

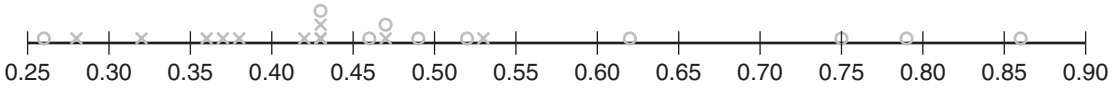


Figura 1.1: Datos de peso del tallo.

Como señalamos, el uso o la aplicación de conceptos de probabilidad permiten una interpretación de la vida cotidiana de los resultados de la inferencia estadística. Entonces, se afirma que la inferencia estadística emplea los conceptos de probabilidad. A partir de los dos ejemplos anteriores, se puede saber que la información muestral está disponible para el analista y, con la ayuda de métodos estadísticos y elementos de probabilidad, se obtienen conclusiones acerca de alguna característica de la población. (El proceso no parece ser aceptable en el ejemplo 1.1 y el nitrógeno en verdad influye en el peso promedio de los tallos del ejemplo 1.2.) Así, para un problema estadístico, **tanto la muestra como la estadística inferencial nos permiten obtener conclusiones acerca de la población, de manera que la estadística inferencial utiliza ampliamente los elementos de probabilidad**. Tal razonamiento es *inductivo* por naturaleza. Ahora conforme avancemos hacia el capítulo 2 y más adelante, el lector encontrará que a diferencia de nuestros dos ejemplos actuales, no nos enfocaremos en resolver problemas estadísticos. En muchos de los ejemplos que estudiaremos no se utilizarán muestras. Se describirá claramente una población con todas sus características. Luego las preguntas importantes se enfocarán en la naturaleza de los datos que hipotéticamente podrían obtenerse a partir de la población. Entonces, los **problemas de probabilidad nos permiten obtener conclusiones acerca de las características de los datos hipotéticos que se tomen de la población con base en las características conocidas de la población**. Esta clase de razonamiento es *deductivo* por naturaleza. La figura 1.2 muestra las relaciones básicas entre la probabilidad y la estadística inferencial.

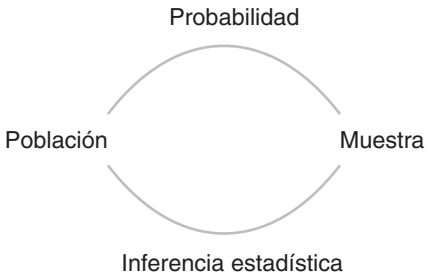


Figura 1.2: Relaciones básicas entre la probabilidad y la estadística inferencial.

Ahora, en términos generales, ¿cuál es más importante, el campo de la probabilidad o el de la estadística? Ambos son muy importantes y evidentemente se complementan. La única certeza respecto de la didáctica de ambas disciplinas reside en el hecho de que si la estadística debe enseñarse con un nivel mayor que el de un simple "libro de cocina", entonces tiene que enseñarse primero la disciplina de la probabilidad. Esta regla se deriva de la noción de que nada puede aprenderse sobre una población a partir de una muestra, hasta que el analista aprenda los rudimentos

de incertidumbre en esa muestra. Considere el ejemplo 1.1, la pregunta se centra en torno de si la población, definida por el proceso, tiene o no más de 5% elementos defectuosos. En otras palabras, la suposición es que **en promedio** 5 de cada 100 artículos salen defectuosos. Ahora la muestra contiene 100 artículos y 10 están defectuosos. ¿Esto apoya la suposición o la refuta? Aparentemente se trataría de una refutación de la suposición, pues 10 de cada 100 parecería ser "bastante". Pero sin nociones de probabilidad, ¿cómo lo sabríamos? Sólo mediante el estudio del material de los siguientes capítulos aprenderemos que a condición de que el proceso sea aceptable (5% de defectuosos), la probabilidad de obtener 10 o más artículos defectuosos en una muestra de 100 es de 0.0282.

Dimos dos ejemplos donde los elementos de probabilidad ofrecen un resumen que el científico o el ingeniero pueden usar como evidencia sobre la cual basar una decisión. El puente entre los datos y la conclusión está, por supuesto, basado en los fundamentos de la inferencia estadística, la teoría de la distribución y las distribuciones de muestreos que se examinan en futuros capítulos.

1.3 Procedimientos de muestreo; acopio de los datos

En la sección 1.1 estudiamos muy brevemente la noción de muestreo y del proceso de muestreo. Mientras que el muestreo aparece como un concepto simple, la complejidad de las preguntas que deben contestarse acerca de la población o las poblaciones, en ocasiones requiere que el proceso de muestreo sea muy complejo. Mientras que la noción de muestreo se examina con detalles en el capítulo 8, aquí nos esforzaremos por dar algunas nociones de sentido común sobre el muestreo. Se trata de una transición natural hacia el análisis del concepto de variabilidad.

Muestreo aleatorio simple

La importancia del muestreo adecuado gira en torno del grado de confianza con que el analista es capaz de responder las preguntas que se le formulan. Supongamos que sólo hay una población en el problema. Recuerde que en el ejemplo 1.2 había dos poblaciones implicadas. El **muestreo aleatorio simple** significa que cualquier muestra dada de un *tamaño muestral* específico tiene la misma probabilidad de ser seleccionada que cualquier otra muestra del mismo tamaño. El termino **tamaño muestral** simplemente indica el número de elementos en la muestra**.** Evidentemente en muchos casos es posible utilizar una tabla de números aleatorios al seleccionar la muestra. La ventaja del muestreo aleatorio simple radica en que ayuda en la eliminación del problema de tener una muestra que refleje una población diferente (quizá más restringida) de aquella sobre la cual se necesitan realizar las inferencias. Por ejemplo, se elige una muestra para contestar diferentes preguntas respecto de las preferencias políticas en cierta entidad del país. La muestra implica la elección, digamos, de 1000 familias a las cuales aplicar una encuesta. Ahora suponga que resulta que no se utiliza el muestreo aleatorio. Más bien, todas o casi todas las 1000 familias se eligen de una zona urbana. Se considera que las preferencias políticas en las áreas rurales difieren de las de las áreas urbanas. En otras palabras, la muestra obtenida en realidad limitó a la población y, por lo tanto, las inferencias también tendrán que restringirse a la "población limitada", por lo que en este caso tal confinamiento podría volverse indeseable. Si, de hecho, las inferencias necesitan hacerse respecto de la entidad en su conjunto, la muestra cuyo tamaño son 1000 familias que se utiliza aquí a menudo se conoce como **muestra sesgada**.

Como sugerimos anteriormente, el muestreo aleatorio simple no siempre resulta adecuado. El enfoque alternativo que se utilice dependerá de la complejidad del problema. Con frecuencia, por ejemplo, las unidades muestrales no son homogéneas y naturalmente se dividen en grupos que no se traslapan que son homogéneos. Tales grupos se llaman estratos, y un procedimiento llamado *muestreo aleatorio estratificado* implica la selección al azar de una muestra *dentro* de cada estrato. El propósito consiste en asegurarse que cada uno de los estratos no esté ni sobrerrepresentado ni subrepresentado. Por ejemplo, suponga que se encuesta a una muestra para reunir información preliminar sobre un referéndum que se piensa realizar en determinada ciudad. La ciudad se subdivide en varios grupos étnicos que representan estratos naturales y, para no excluir ni sobrerrepresentar a algún grupo de cada uno de ellos, podrían elegirse muestras aleatorias separadas de cada grupo.

Diseño experimental

El concepto de aleatoriedad o asignación aleatoria juega un papel muy importante en el área del **diseño experimental**, el cual se introdujo brevemente en la sección 1.1 y es un fundamento muy importante en casi cualquier área de la ingeniería y de la ciencia experimental. Lo estudiaremos con detenimiento en los capítulos 13 a 15. No obstante, sería útil dar aquí una breve introducción en el contexto del muestreo aleatorio. Un conjunto de **tratamientos** o **combinaciones de tratamientos** se vuelven las poblaciones que van a estudiarse o a compararse en algún sentido. Un ejemplo es el tratamiento "con nitrógeno" versus "sin nitrógeno" del ejemplo 1.2. Otro ejemplo sencillo sería el "placebo" versus "medicamento activo"; o en un estudio sobre la fatiga por corrosión, tendríamos combinaciones de tratamientos que impliquen espécimen con recubrimiento o sin recubrimiento, así como condiciones de alta o de baja humedad, a las cuales se somete el espécimen. De hecho, hay cuatro combinaciones de factores o de tratamientos (es decir, 4 poblaciones), y quizá se formulen y se respondan muchas preguntas usando los métodos estadísticos e inferenciales. Considere primero la situación del ejemplo 1.2. Hay 20 plantones enfermos implicados en el experimento. A partir de los datos es fácil observar que los plantones son diferentes entre sí. Dentro del grupo con nitrógeno (o del grupo sin nitrógeno) hay **variabilidad** considerable en el peso de los tallos, la cual se debe a lo que, por lo general, se denomina **unidad experimental**. Éste es un concepto muy importante en la estadística inferencial, cuya descripción no termina en este capítulo. La naturaleza de la variabilidad es muy importante. Si es demasiado grande, derivada de una condición de falta de homogeneidad excesiva en las unidades experimentales, la variabilidad "eliminará" cualquier diferencia detectable entre ambas poblaciones. Recuerde que en este caso eso no ocurrió.

La gráfica de puntos de la figura 1.1 y el valor-P indican una clara distinción entre esas dos condiciones. Pero ¿qué papel juegan tales unidades experimentales en el proceso mismo de acopio de los datos? El enfoque por sentido común y, de hecho, estándar es asignar los 20 plantones o unidades experimentales **aleatoriamente a las dos condiciones o tratamientos**. En el estudio del medicamento quizá decidamos utilizar un total de 200 pacientes disponibles, quienes serán claramente distinguibles en algún sentido. Ellos son las unidades experimentales. No obstante, tal vez todos tengan una condición crónica para la cual el fármaco sea un tratamiento potencial. Así en el denominado **diseño completamente aleatorio**, se asignan al azar 100 pacientes al placebo y 100 al medicamento activo. De nuevo, son estas unidades experimentales en el grupo o tratamiento las que producen la variabilidad en el resultado de los datos (es decir, la variabilidad en el resultado medido), digamos,

la presión sanguínea; o cualquier valor de la eficacia de un medicamento que sea importante. En el estudio de la fatiga por corrosión, las unidades experimentales son los especímenes que se someten a la corrosión.

¿Por qué las unidades experimentales se asignan aleatoriamente?

¿Cuál es la posible influencia negativa de no asignar aleatoriamente las unidades experimentales a los tratamientos o a las combinaciones de tratamientos? Esto se observa más claramente en el caso del estudio del medicamento. Entre las características de los pacientes que producen variabilidad en los resultados están la edad, el género, el peso, etcétera. Tan sólo suponga que por casualidad el grupo del placebo contiene una muestra de personas que son predominantemente más obesas que las del grupo del tratamiento. Quizá los individuos más obesos muestren una tendencia a tener mayor presión sanguínea, lo cual evidentemente sesga el resultado y, por lo tanto, cualquier resultado que se obtenga mediante la aplicación de la inferencia estadística podría tener poco que ver con el efecto del medicamento, pero mucho con las diferencias en el peso de ambas muestras de pacientes.

Deberíamos enfatizar la importancia del término **variabilidad**. La variabilidad excesiva entre las unidades experimentales "disfraza" los hallazgos científicos. En secciones posteriores intentaremos clasificar y cuantificar las medidas de variabilidad. En las siguientes secciones presentaremos y estudiaremos cantidades específicas que se calculan a partir de las muestras; las cantidades dan un sentido de la naturaleza de la muestra respecto del centro de ubicación de los datos y la variabilidad de los mismos. Un análisis de varias de tales medidas de un solo número ofrece un preámbulo de los componentes importantes de la información estadística en los métodos estadísticos que se utilizan en los capítulos 8 a 15. Se trata de medidas que ayudan a clasificar la naturaleza del conjunto de datos que caen en la categoría de **estadística descriptiva**. Este material es una introducción a una presentación breve de los métodos pictóricos y gráficos que van incluso más allá en la caracterización del conjunto de datos. El lector debería entender que los métodos estadísticos que se presentan aquí se utilizarán a lo largo de todo el texto. Para tener una imagen más clara de lo que implican los estudios de diseño experimental, tenemos el siguiente ejemplo.

Ejemplo 1.3: Se realizó un estudio sobre la corrosión con la finalidad de determinar si un metal de aluminio recubierto con una sustancia retardadora de la corrosión reducía la cantidad de la corrosión. El recubrimiento es un protector que se publicita como que minimiza el daño por fatiga en esta clase de material. La influencia de la humedad sobre la magnitud de la corrosión también es de interés. Una medición de la corrosión puede expresarse en millares de ciclos hasta ruptura. Se utilizaron dos niveles de recubrimiento: sin recubrimiento y con recubrimiento químico contra la corrosión. Además, los dos niveles de humedad relativa son de 20 y 80%, respectivamente.

El experimento implica cuatro combinaciones de tratamientos que se listan en la siguiente tabla. Hay ocho unidades experimentales que se usarán y son especímenes de aluminio preparados, de los cuales dos se asignan aleatoriamente a cada una de las cuatro combinaciones de tratamiento. Los datos se presentan en la tabla 1.2.

Los datos de la corrosión son promedios de los dos especímenes. En la figura 1.3 se presenta una grafica con los promedios. Un valor relativamente grande de ciclos hasta ruptura representa una cantidad pequeña de corrosión. Como podría esperarse, parece que un incremento en la humedad hace que empeore la corrosión. Además,

Tabla 1.2: Datos para el ejemplo 1.3

Recubrimiento	**Humedad**	**Promedio de corrosi?n en miles de ciclos hasta ruptura**
Sin recubrimiento	20%	975
	80%	350
Con recubrimiento químico contra la corrosión	20%	1750
	80%	1550

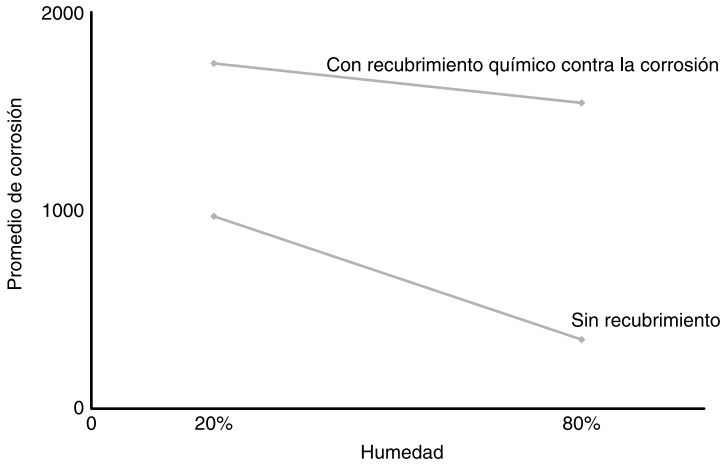


Figura 1.3: Resultados de corrosión para el ejemplo 1.3.

parece que el uso del procedimiento de recubrimiento químico contra la corrosión reduce la corrosión.

En este caso de diseño experimental, el ingeniero eligió sistemáticamente las cuatro combinaciones de tratamiento. Para vincular esta situación con los conceptos con los cuales el lector ha estado familiarizado hasta aquí, deberíamos suponer que las condiciones que representan las cuatro combinaciones de tratamientos son cuatro poblaciones separadas y que los dos valores de corrosión observados en cada una de las poblaciones constituyen importantes piezas de información. La importancia del promedio al captar y resumir ciertas característica en la población se destacará en la sección 1.4. Mientras seamos capaces de obtener conclusiones acerca del papel de la humedad y del impacto de recubrir el espécimen a partir de la figura, no podremos evaluar en realidad los resultados a partir de cualquier punto de vista analítico sin tomar en cuenta la *variabilidad alrededor* del promedio. De nuevo, como señalamos anteriormente, si los dos valores de corrosión en cada una de las combinaciones de tratamientos son muy cercanos, la imagen de la figura 1.3 podría ser una descripción precisa. Pero si cada valor de la corrosión en la figura es un promedio de dos valores que están ampliamente dispersos, entonces esta variabilidad podría, de hecho, verdaderamente "eliminar" cualquier información que parezca difundirse cuando uno tan sólo observa los promedios. Los siguientes ejemplos ilustran los conceptos:

1. La asignación aleatoria a las combinaciones de tratamientos (recubrimiento/humedad) de las unidades experimentales (especímenes)
2. El uso de promedios muestrales (valores de corrosión promedio) para resumir la información muestral
3. La necesidad de considerar las medidas de variabilidad en el análisis de cualquier nuestra o conjunto de muestras

Este ejemplo sugiere la necesidad del tema de las secciones 1.4 y 1.5, es decir, la estadística descriptiva que indica las medidas del centro de ubicación en un conjunto de datos, y aquellas que miden la variabilidad.

1.4 Medidas de posición: La media y la mediana de una muestra

En un conjunto de datos las medidas de posición están diseñadas para brindar al analista alguna medida cuantitativa de dónde está el centro de los datos en una muestra. En el ejemplo 1.2 parece como si el centro de la muestra con nitrógeno claramente excediera al de la muestra sin nitrógeno. Una medida obvia y muy útil es la **media de la muestra**. La media es simplemente un promedio numérico.

Definición 1.1: Suponga que las observaciones en una muestra son $x_1, x_2, \ldots, x_n$. La **media de la muestra**, que se denota con $\bar{x}$, es

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Hay otras medidas de tendencia central que se explican con detalle en capítulos posteriores. Una medida importante es la **mediana de la muestra**. El propósito de la mediana de la muestra es reflejar la tendencia central de la muestra, de manera que no esté influida por los valores extremos. Dado que las observaciones en una muestra son $x_1, x_2, \ldots, x_n$, acomodados en orden de magnitud creciente, la mediana de la muestra es

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{si } n \text{ es impar,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{si } n \text{ es par.} \end{cases}$$

Por ejemplo, supongamos que el conjunto de datos es el siguiente: 1.7, 2.2, 3.9, 3.11 y 14.7. La media y la mediana de la muestra son, respectivamente,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9.$$

Es evidente que la media está influida de manera considerable por la presencia de la observación extrema, 14.7; en tanto que el lugar de la mediana hace énfasis en el verdadero "centro" del conjunto de datos. En el caso del conjunto de datos de dos

muestras del ejemplo 1.2, las dos medidas de tendencia central para las muestras individuales son

$$\begin{aligned} \bar{x} \text{ (sin nitrógeno)} &= 0.399 \text{ gramos}, \\ \tilde{x} \text{ (sin nitrógeno)} &= \frac{0.38 + 0.42}{2} = 0.400 \text{ gramos}, \\ \bar{x} \text{ (con nitrógeno)} &= 0.565 \text{ gramos}, \\ \tilde{x} \text{ (con nitrógeno)} &= \frac{0.49 + 0.52}{2} = 0.505 \text{ gramos}. \end{aligned}$$

Hay una diferencia de concepto evidente entre la media y la mediana. Para el lector con ciertas nociones de ingeniería quizá sea de interés que la media de la muestra es el **centroide de los datos** en una muestra. En cierto sentido es el punto donde se puede colocar un fulcro para equilibrar un sistema de "pesos", que son las posiciones de los datos individuales. Esto se muestra en la figura 1.4 respecto de la muestra "con nitrógeno".

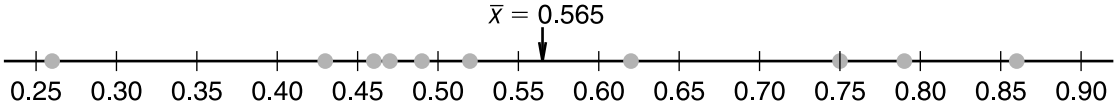


Figura 1.4: Media de la muestra como centroide del peso del tallo "con nitrógeno".

En capítulos futuros, la base para el cálculo de $\bar{x}$ es un **estimado** de la **media de la población**. Como señalamos anteriormente, el propósito de la inferencia estadística es obtener conclusiones acerca de las características o **parámetros** de la población, y la **estimación** es una característica muy importante de la inferencia estadística.

La mediana y la media llegan a ser muy diferentes entre sí. Note, sin embargo, que en el caso de los datos del peso de los tallos, el valor de la media de la muestra para "sin nitrógeno" es bastante similar al valor de la mediana.

Otras medidas de posición

Hay otros métodos para calcular el centro de ubicación de los datos en la muestra. No los trataremos en este momento. Por lo general, las alternativas para la media de la muestra se diseñan para generar valores que representen relación entre la media y la mediana. Rara vez utilizamos alguna de tales medidas. No obstante, es aleccionador estudiar una clase de estimadores conocida como **media recortada**, la cual se calcula "quitando" cierto porcentaje de los valores mayores y menores del conjunto. Por ejemplo, la media recortada 10% se encuentra eliminando tanto el 10% de los valores mayores como de los menores, y calculando el promedio de los valores restantes. Por ejemplo, en el caso de los datos del peso de los tallos eliminaríamos el valor más alto y el más bajo, ya que el tamaño de la muestra es 10 en cada caso. De manera que para el grupo sin nitrógeno la media recortada 10% está dado por

$$\bar{x}_{\text{tr}(10)} = \frac{0.32 + 0.37 + 0.47 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

y para la media recortada 10% del grupo con nitrógeno tenemos

$$\bar{x}_{\mathrm{tr}(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Observe que en este caso, como se esperaba, las medias recortadas están cerca tanto de la media como de la mediana para las muestras individuales. Desde luego, el enfoque de la media recortada es menos sensible a los valores extremos que la media de la muestra; pero no tan insensible como la mediana. Por otro lado, el enfoque de la media recortada utiliza mayor información. Note que la mediana de la muestra es, de hecho, un caso especial de la media recortada, en el cual se eliminan todos los datos de la muestra y queda solo el central o dos observaciones.

Ejercicios

1.1 Se registran las siguientes mediciones para el tiempo de secado (en horas) de cierta marca de pintura esmaltada.

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponga que las mediciones constituyen una muestra aleatoria simple.

a) ¿Cuál es el tamaño de la muestra anterior?

b) Calcule la media de la muestra para estos datos.

c) Calcule la mediana de la muestra.

d) Grafique los datos utilizando una gráfica de puntos.

e) Calcule la media recortada 20% para el conjunto de datos anterior.

1.2 Según la publicación *Chemical Engineering*, una propiedad importante de una fibra es su absorción del agua. Se toma una muestra aleatoria de 20 piezas de fibra de algodón y se mide la impermeabilidad de cada una. Los valores de absorción son los siguientes:

18.71	21.41	20.72	21.81	19.29	22.43	20.17
23.71	19.44	20.50	18.92	20.33	23.00	22.85
19.25	21.77	22.11	19.77	18.04	21.12	

a) Calcule la media y la mediana de la muestra para los valores de la muestra anterior.

b) Calcule la media recortada 10%.

c) Elabore una gráfica de puntos con los datos de la absorción.

1.3 Se utiliza cierto polímero para los sistemas de evacuación de los aviones. Es importante que el polímero sea resistente al proceso de envejecimiento. Se utilizaron veinte especímenes del polímero en un experimento. Diez se asignaron aleatoriamente para exponerse al proceso de acelerado, el cual implica la exposición a altas temperaturas durante 10 días. Se hicieron las mediciones de resistencia a la tensión de los especímenes y se registraron los siguientes datos sobre resistencia a la tensión en psi.

Sin envejecimiento acelerado:	227	222	218	217	225
	218	216	229	228	221
Con envejecimiento acelerado:	219	214	215	211	209
	218	203	204	201	205

a) Elabore la gráfica de puntos de los datos.

b) A partir de la gráfica, ¿parecería que el proceso de envejecimiento tuvo un efecto en la resistencia a la tensión de este polímero?

c) Calcule la resistencia a la tensión de la media de la muestra en ambas muestras.

d) Calcule la mediana de ambas. Discuta la similitud o falta de similitud entre la media y la mediana de cada grupo.

1.4 En un estudio realizado por el Departamento de Ingeniería Mecánica del Tecnológico de Virginia, se compararon las varillas de acero que abastecen dos compañías diferentes. Se fabricaron diez resortes de muestra con las varillas de metal proporcionadas por cada una de las compañías y se registraron sus medidas de flexibilidad. A continuación se presentan los datos.

Compañía A:	9.3	8.8	6.8	8.7	8.5
	6.7	8.0	6.5	9.2	7.0
Compañía B:	11.0	9.8	9.9	10.2	10.1
	9.7	11.0	11.1	10.2	9.6

a) Calcule la media y la mediana de la muestra para los datos de ambas compañías.

b) Grafique los datos para las dos compañías en la misma línea y explique su conclusión.

1.5 Veinte adultos hombres de entre 30 y 40 años de edad participaron en un estudio para evaluar el efecto de cierto régimen de salud, que incluye dieta y ejercicio, en el colesterol sanguíneo. Se eligieron diez aleatoriamente para el grupo de control y los otros diez se asignaron

para tomar parte en el régimen como grupo de tratamiento durante un periodo de 6 meses. Los siguientes datos muestran la reducción en el colesterol que experimentaron en ese periodo los 20 sujetos:

Grupo de control	7	3	−4	14	2
	5	22	−7	9	5
Grupo de tratamiento:	−6	5	9	4	4
	12	37	5	3	3

a) Elabore una gráfica de puntos, con los datos de ambos grupos en una misma gráfica.

b) Calcule la media, la mediana y la media recortada 10% para ambos grupos.

c) Explique por qué la diferencia en la media sugiere una conclusión acerca del efecto del régimen, en tanto que la diferencia en las medianas o las medias recortadas sugiere una conclusión diferente.

1.6 La resistencia a la tensión del caucho de silicón se considera una función de la temperatura de vulcanizado. Se llevó a cabo un estudio donde muestras de 12 especímenes del caucho se prepararon utilizando temperaturas de vulcanizado de 20 °C y 40 °C. Los siguientes datos presentan los valores de resistencia a la tensión en megapascales.

20 °C:	2.07	2.14	2.22	2.03	2.21	2.03
	2.05	2.18	2.09	2.14	2.11	2.02
45 °C:	2.52	2.15	2.49	2.03	2.37	2.05
	1.99	2.42	2.08	2.42	2.29	2.01

a) Elabore una gráfica de puntos con los datos tanto de los valores de resistencia a la tensión a temperatura alto como los de a temperatura baja.

b) Calcule la resistencia a la tensión de la media de la muestra para ambas muestras.

c) ¿Parece que la temperatura de vulcanizado tiene influencia en la resistencia a la tensión según la gráfica? Argumente.

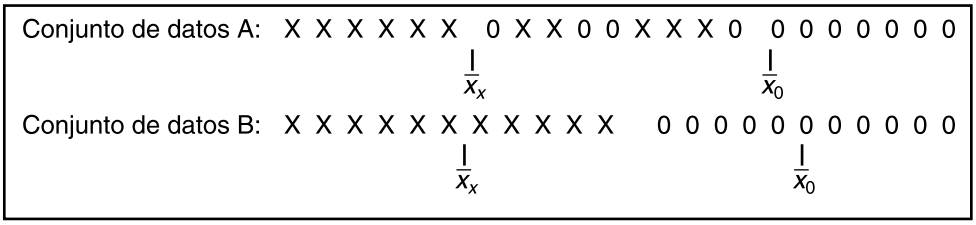
d) ¿Qué parece estar influido por un incremento en la temperatura de vulcanizado?

1.5 Medidas de variabilidad

La variabilidad de una muestra juega un papel importante en el análisis de datos. La variabilidad de un proceso y un producto es un hecho real en los sistemas científicos y de ingeniería: el control o la reducción de la variabilidad de un proceso a menudo es una fuente de mayores dificultades. Cada vez con mayor frecuencia, los ingenieros y administradores de procesos aprenden que la calidad del producto, y como resultado, las ganancias que se derivan de productos manufacturados son, con mucho, una función de la **variabilidad del proceso**. De esta manera, gran parte de los capítulos 9 a 15 tiene que ver con el análisis de datos y con los procedimientos de modelado, en los cuales la variabilidad de la muestra juega un papel significativo. Incluso en problemas de análisis de datos pequeños, el éxito de un método estadístico específico podría depender de la magnitud de la variabilidad entre las observaciones en la muestra. Las medidas de posición en una muestra no brindan un resumen adecuado de la naturaleza de un conjunto de datos. Es decir, en el ejemplo 1.2 no podemos concluir que el uso del nitrógeno realza el crecimiento sin tomar en cuenta la variabilidad de la muestra.

Mientras que los detalles del análisis de este tipo de conjuntos de datos se deja para estudiar en el capítulo 9, a partir de la figura 1.1 debería quedar claro que la variabilidad entre las observaciones "sin nitrógeno" y la variabilidad entre las observaciones "con nitrógeno", desde luego, tienen alguna consecuencia. De hecho, parece que la variabilidad dentro de la muestra con nitrógeno es mayor que la de la muestra sin nitrógeno. Quizás haya algo acerca de la inclusión del nitrógeno que no tan sólo incrementa el peso de los tallos ($\bar{x}$ de 0.565 gramos en comparación con una $\bar{x}$ de 0.399 gramos para la muestra sin nitrógeno), aunque también incrementa la variabilidad en el peso de los tallos (es decir, hace que el peso de los tallos sea más inconsistente).

Por ejemplo, compare los dos conjuntos de datos de abajo. Cada uno contiene dos muestras y la diferencia en las medias es aproximadamente la misma para las dos muestras: el conjunto de datos B parece proporcionar un contraste mucho más claro entre las dos poblaciones de las que se tomaron las muestras. Si el propósito de tal experimento es detectar la diferencia entre las dos poblaciones, la tarea se lleva a cabo en el caso del conjunto de datos B. Sin embargo, en el conjunto de datos A la



amplia variabilidad *dentro* de las dos muestras ocasiona dificultad. De hecho, no es claro que haya una diferencia *entre* las dos poblaciones.

Rango y desviación estándar de la muestra

Así como hay muchas medidas de tendencia central o de posición, hay muchas medidas de dispersión o variabilidad. Quizá la más simple sea el **rango de la muestra** $X_{\text{máx}} - X_{\text{mín}}$. El rango puede ser muy útil y se discute con amplitud en el capítulo 17 sobre *control estadístico de calidad*. La medida muestral de dispersión que se utiliza más a menudo es la **desviación estándar de la muestra**. Nuevamente denotemos con $x_1, x_2, \ldots, x_n$ los valores de la muestra;

Definición 1.2:

La **varianza de la muestra**, denotada con s^2, está dada por

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

La **desviación estándar de la muestra**, denotada con s, es la raíz cuadrada positiva de s^2, es decir,

$$s = \sqrt{s^2}.$$

Para el lector debería quedar claro que la desviación estándar de la muestra es, de hecho, una medida de variabilidad. Una variabilidad grande en un conjunto de datos produce valores relativamente grandes de $(x - \bar{x})^2$ y por ello una varianza de la muestra grande. La cantidad $n - 1$ a menudo se denomina **grados de libertad asociados con la varianza** estimada. En este ejemplo simple, los grados de libertad representan el número de piezas de información independientes disponibles para calcular la variabilidad. Por ejemplo, suponga que deseamos calcular la varianza de la muestra y la desviación estándar del conjunto de datos (5, 17, 6, 4). El promedio de la muestra es $\bar{x} = 8$. El cálculo de la varianza implica:

$$(5-8)^2 + (17-8)^2 + (6-8)^2 + (4-8)^2 = (-3)^2 + 9^2 + (-2)^2 + (-4)^2.$$

Las cantidades dentro de los paréntesis suman cero. En general, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (véase el ejercicio 1.16 de la página 28). Entonces, el cálculo de la varianza de una muestra no implica n **desviaciones cuadradas independientes** de la media $\bar{x}$. De hecho, como el último valor de $x - \bar{x}$ está determinado por los primeros $n - 1$ valores, decimos que éstas son $n - 1$ "piezas de información" que producen s^2. Por ello hay $n - 1$ grados de libertad, en vez de n grados de libertad para calcular la varianza de una muestra.

Ejemplo 1.4: En un caso que se estudia ampliamente en el capítulo 10, un ingeniero se interesa en probar el "sesgo" en un medidor de pH. Se recaban los datos utilizándolo para medir el pH de una sustancia neutral (pH = 7.0). Se toma una muestra de tamaño 10 y se obtienen los siguientes resultados:

$$7.07 \quad 7.00 \quad 7.10 \quad 6.97 \quad 7.00 \quad 7.03 \quad 7.01 \quad 7.01 \quad 6.98 \quad 7.08.$$

La media de la muestra $\bar{x}$ está dada por

$$\bar{x} = \frac{7.07 + 7.00 + 7.10 + \cdots + 7.08}{10} = 7.0250.$$

La varianza de la muestra s^2 está dada por

$$\begin{aligned} s^2 =& \frac{1}{9}[(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + (7.10 - 7.025)^2 \\ &+ \cdots + (7.08 - 7.025)^2] = 0\ 001939. \end{aligned}$$

Como resultado, la desviación estándar de la muestra está dada por

$$s = \sqrt{0.00193} = 0.044.$$

De manera que la desviación estándar de la muestra es 0.0440 con $n - 1 = 9$ grados de libertad.

Unidades para la desviación estándar y la varianza

A partir de la definición 1.2 debería ser evidente que la varianza es una medida de la desviación cuadrática promedio a partir de la media $\bar{x}$. Empleamos el término *desviación cuadrática promedio* aun cuando la definición utilice una división entre $n - 1$ grados de libertad, en vez de n. Desde luego, si n es grande la diferencia en el denominador es inconsecuente. Por lo tanto, la varianza de la muestra tiene unidades que son el cuadrado de las unidades en los datos observados; mientras que la desviación estándar de la muestra se encuentra en unidades lineales. Considere los datos del ejemplo 1.2. Los pesos del tallo se miden en gramos. Como resultado, las desviaciones estándar de la muestra están en gramos y las varianzas se miden en gramos^2. De hecho, las desviaciones estándar individuales son 0.0728 gramos para el caso sin nitrógeno y 0.1867 gramos para el grupo con nitrógeno. Observe que la variabilidad caracterizada por la desviación estándar en verdad indica una variabilidad significativamente más grande en la muestra con nitrógeno. Esta condición se destaca en la figura 1.1.

¿Cuál es la medida de variabilidad más importante?

Como indicamos antes, el rango de la muestra tiene aplicaciones en el área del control estadístico de la calidad. Quizás el lector considere que es redundante el uso tanto de la varianza de la muestra como de la desviación estándar de la muestra. Ambas medidas reflejan el mismo concepto en la variabilidad de la medición; pero la desviación estándar de la muestra mide la variabilidad en unidades lineales; en tanto que la varianza de la muestra se mide en unidades cuadradas. Ambas juegan papeles importantes en el uso de los métodos estadísticos. Mucho de lo que se logra en el contexto de la inferencia estadística implica la obtención de conclusiones acerca de

las características de poblaciones. Entre tales características son constantes los denominados **parámetros de la población**. Dos parámetros importantes son la **media de la población** y la **varianza de la población**. La varianza de la muestra juega un papel explícito en los métodos estadísticos que se utilizan para obtener inferencias sobre la varianza de la población. La desviación estándar de la muestra tiene un papel importante, junto con la media de la muestra, en las inferencias que se realizan acerca de la media de la población. En general, la varianza se considera más en la teoría inferencial; mientras que la desviación estándar se utiliza más en aplicaciones.

Ejercicios

1.7 Considere los datos del tiempo de secado del ejercicio 1.1 de la página 13. Calcule la varianza de la muestra y la desviación estándar de la muestra.

1.8 Calcule la varianza de la muestra y la desviación estándar para los datos de absorción del agua del ejercicio 1.2 de la página 13.

1.9 El ejercicio 1.3 de la página 13 presentó muestras de datos de resistencia a la tensión, unos para especímenes que se expusieron a un proceso de envejecimiento, y otros donde no hubo tal proceso en los especímenes. Calcule la varianza de la muestra y su desviación estándar en cuanto a la resistencia a la tensión en ambas muestras.

1.10 Para los datos del ejercicio 1.4 de la página 13, calcule tanto la media como la varianza de la "flexibilidad" para las compañías A y B.

1.11 Considere los datos del ejercicio 1.5 de la página 13. Calcule la varianza de la muestra y la desviación estándar de la muestra para ambos grupos: el de tratamiento y el de control.

1.12 Para el ejercicio 1.6 de la página 14, calcule la desviación estándar de la muestra en la resistencia a la tensión para las muestras, separadamente para ambas temperaturas. ¿Parece que un incremento en la temperatura influye en la variabilidad de la resistencia a la tensión? Explique.

1.6 Datos discretos y continuos

La inferencia estadística a través del análisis de estudios observacionales o de experimentos diseñados se utiliza en muchas áreas científicas. Los datos reunidos pueden ser **discretos** o **continuos**, según el área de aplicación. Por ejemplo, un ingeniero químico podría interesarse en un experimento que lo lleve a condiciones en que se maximice la producción. Aquí, por supuesto, la producción estaría en porcentaje, o gramos/libra, medida en un continuo. Por otro lado, un toxicólogo que realice un experimento de combinación de fármacos quizás encuentre datos que son binarios por naturaleza (es decir, el paciente responde o no).

Distinciones importantes se realizan entre datos discretos y continuos en la teoría de la probabilidad que nos permiten obtener inferencias estadísticas. Con frecuencia las aplicaciones de la inferencia estadística se encuentran cuando se trata de *datos por conteo*. Por ejemplo, un ingeniero que se interese en estudiar el número de partículas radiactivas que pasan a través de un contador en, digamos, 1 milisegundo. El personal responsable por la eficiencia de una instalación portuaria quizá se interese en las características del número de buques petroleros que llegan diariamente a cierta ciudad portuaria. En el capítulo 5, varios escenarios distintos, al mostrar varias formas de manejar los datos, se examinan para situaciones de datos por conteo.

Incluso en esta fase inicial del texto, debería ponerse especial atención a algunos detalles que se asocian con datos binarios. Son muchas las aplicaciones que requieren el análisis estadístico de datos binarios. Con frecuencia la medición que se utiliza

en el análisis es la *proporción muestral*. En efecto, la situación binaria implica dos categorías. Si en los datos hay n unidades y x se define como el número que cae en la categoría 1, entonces $n - x$ cae en la categoría 2. Así, x/n es la proporción muestral en la categoría 1 y $1 - x/n$ es la proporción muestral en la categoría 2. En la aplicación biomédica, por ejemplo, 50 pacientes representarían las unidades de la muestra y si, después de que se les suministra el medicamento, 20 de 50 experimentan mejoría en malestares estomacales (que son comunes en los 50), entonces $\frac{20}{50} = 0.4$ es la proporción muestral para la cual el medicamento tuvo éxito, y $1 - 0.4 = 0.6$ es la proporción muestral para la cual el fármaco no tuvo éxito. En realidad la medición numérica fundamental para datos binarios, por lo general, se denota con 0 o con 1. Por ejemplo, en nuestro ejemplo médico, un resultado exitoso se denota con un 1 y uno no exitoso con un 0. Entonces, realmente la proporción muestral es una media de la muestra de unos y ceros. Para la categoría de éxitos,

$$\frac{x_1 + x_2 + \cdots + x_{50}}{50} = \frac{1 + 1 + 0 + \cdots + 0 + 1}{50} = \frac{20}{50} = 0.4.$$

¿Qué clases de problemas se resuelven en situaciones con datos binarios?

Los tipos de problemas que enfrentan científicos e ingenieros que tratan con datos binarios no son muy difíciles, a diferencia de aquellos donde las mediciones continuas son de interés. No obstante, se utilizan técnicas diferentes, pues las propiedades estadísticas de las proporciones muestrales son bastante diferentes de las medias de la muestra que resultan de los promedios tomados a partir de poblaciones continuas. Considere los datos del ejemplo en el ejercicio 1.6 de la página 14. El problema estadístico que subyace a este caso se enfoca en si una intervención, digamos un incremento en la temperatura de vulcanizado, alterará la resistencia a la tensión de la media de la población que se asocia con el proceso del caucho de silicón. Por otro lado, en el área del control de la calidad, suponga que el fabricante de neumáticos para automóvil informa que en un embarque con 5000 neumáticos, seleccionados aleatoriamente del proceso, hay 100 defectuosos. Aquí la proporción muestral es $\frac{100}{5000} = 0.02$. Luego de realizar un cambio en el proceso para reducir los neumáticos defectuosos, se toma una segunda muestra de 5000 y se encuentran 90 defectuosos. La proporción muestral se redujo a $\frac{90}{5000} = 0.018$. Entonces, surge una pregunta: "¿La disminución en la proporción muestral de 0.02 a 0.018 es en verdad suficiente como para sugerir una mejoría real en la proporción de la población?" En ambos casos se requiere el uso de las propiedades estadísticas de los promedios de la muestra: en uno a partir de las muestras de poblaciones continuas, y en el otro a partir de las muestras de poblaciones discretas (binarias). Además, en ambos la media de la muestra es un **estimado** de un parámetro de la población: una media de la población en el primer caso (la resistencia media a la tensión), y una proporción de la población (la proporción de neumáticos defectuosos en la población) en el segundo caso. De manera que aquí tenemos estimados de la muestra que se utilizan para obtener conclusiones científicas respecto de los parámetros de la población. Como indicamos en la sección 1.4, se trata del tema general en muchos problemas prácticos donde se usa la inferencia estadística.

1.7 Modelado estadístico, inspección científica y diagnósticos gráficos

A menudo el resultado final de un análisis estadístico es la estimación de los parámetros de un **modelo postulado**. Esto es por completo natural para los científicos y los ingenieros, pues con frecuencia tratan con el modelado. Un modelo estadístico no es determinista sino, más bien, debe implicar algunos aspectos probabilistas. Por lo general, una forma de modelo es la fundamentación de las **suposiciones** que hace el analista. En nuestro ejemplo 1.2, quizás el científico desee extraer algún nivel de distinción entre las poblaciones "con nitrógeno" y "sin nitrógeno" a través de información de la muestra. El análisis puede requerir cierto modelo para los datos; por ejemplo, que las dos muestras provengan de **distribuciones normales** o **gaussianas**. Véase el capítulo 6 para el estudio de una distribución normal.

A veces el modelo postulado adquiere una forma algo más compleja. Por ejemplo, considere un fabricante de textiles que diseña un experimento donde los especímenes de tela se producen de manera que contengan diferentes porcentajes de algodón. Considere los siguientes datos de la tabla 1.3.

Tabla 1.3: Resistencia a la tensión

Porcentaje del algodón	Resistencia a la tensión
15	7,7,9,8,10
20	19,20,21,20,22
25	21,21,17,19,20
30	8,7,8,9,10

Se fabrican cinco especímenes de tela para cada uno de los cuatro porcentajes de algodón. En este caso, tanto el modelo para el experimento como el tipo de análisis que se utiliza deberían tomar en cuenta el objetivo del experimento y los insumos importantes del científico textil. Algunas gráficas sencillas aclararían la distinción entre las muestras. Véase la figura 1.5; las medias de las muestras y la variabilidad se describen bien en la gráfica de los datos. Un posible objetivo de este experimento es simplemente la determinación de cuáles porcentajes de algodón son en realidad distintos de los otros. En otras palabras, como en el caso de los datos con nitrógeno/sin nitrógeno, ¿para cuáles porcentajes de algodón hay distinciones claras entre las poblaciones o, de forma más específica, entre las medias de las poblaciones? En este caso, quizás un modelo razonable sea que cada muestra viene de una distribución normal. Aquí el objetivo es muy semejante al de los datos con nitrógeno/sin nitrógeno, excepto en que se incluyen más muestras. El formalismo del análisis implica nociones de prueba de hipótesis que se examinan en el capítulo 10. A propósito, tal vez este formalismo no sea necesario a la luz de la gráfica de diagnóstico. Pero, ¿describe el objetivo real del experimento y por consiguiente el enfoque adecuado para el análisis de datos? Es probable que el científico anticipe la existencia de una *resistencia a la tensión máxima de la media de la población*, en el rango de concentración de algodón en el experimento. Aquí el análisis de los datos debería girar alrededor de un tipo diferente de modelo, es decir, uno que postule un tipo de estructura que

relacione la resistencia a la tensión de la media de la población con la concentración de algodón. En otras palabras, un modelo se escribe como

$$\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2,$$

donde $\mu_{t,c}$ es la resistencia a la tensión de la media de la población, que varía con la cantidad de algodón en el producto C. La implicación de este modelo es que para un nivel fijo de algodón, hay una población de mediciones de resistencia a la tensión y la media de la población es $\mu_{t,c}$. Este tipo de modelo, que se denomina **modelo de regresión**, se estudia en los capítulos 11 y 12. La forma funcional la elige el científico. A veces el análisis de datos puede sugerir que se cambie el modelo. Entonces, el analista de datos "considera" un modelo que es posible alterar después de que se haga algún análisis. El uso de un modelo empírico se acompaña por la **teoría de estimación**, donde β_0, β_1 y β_2 se estiman de los datos. Además, se utiliza la inferencia estadística para determinar lo adecuado del modelo.

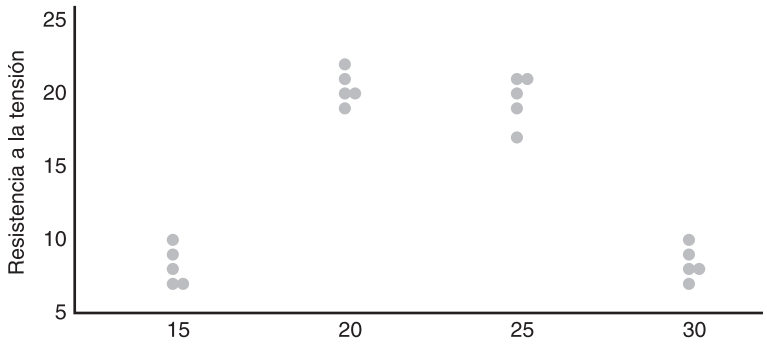


Figura 1.5: Gráfica de resistencia a la tensión y porcentajes de algodón.

Aquí se hacen evidentes dos puntos de las dos ilustraciones de datos: **1.** el tipo de modelo que se emplea para describir los datos a menudo depende del objetivo del experimento, y **2.** la estructura del modelo debería aprovecharse del insumo científico no estadístico. La selección de un modelo representa una **suposición fundamental** sobre la que se basa la inferencia estadística resultante. Se hará evidente a lo largo del libro qué tan importantes llegan a ser las gráficas. A menudo, las gráficas ilustran información que permite que los resultados de la inferencia estadística formal se comuniquen mejor al científico o al ingeniero. A veces, las gráficas o el **análisis exploratorio de los datos** pueden enseñar al analista algo que no se obtiene del análisis formal. Casi cualquier análisis formal requiere suposiciones que se desarrollan a partir del modelo de datos. Las gráficas pueden resaltar bien la **violación de suposiciones** que, de otra forma, no se notarían. A lo largo del libro, las gráficas se utilizan de manera extensa para complementar el análisis formal de los datos. En las siguientes secciones se presentan algunas herramientas gráficas útiles que sirven para el análisis exploratorio o descriptivo de los datos.

1.8 Métodos gráficos y descripción de datos

Evidentemente, el usuario de los métodos estadísticos no puede generar información o datos experimentales suficientes como para caracterizar totalmente a la población.

Sin embargo, a menudo, se emplean conjuntos de datos para aprender acerca de ciertas propiedades de la población. Los científicos y los ingenieros están acostumbrados a trabajar con conjuntos de datos. La importancia de caracterizar o *resumir* la naturaleza de agrupaciones de datos debería ser clara. Con frecuencia un resumen de un conjunto de datos que utilice gráficas daría una visión sobre el sistema a partir del cual se tomaron los datos.

En esta sección se estudian con detalle el papel del muestreo y de la presentación de los datos para reafirmar la **inferencia estadística** respecto de sistemas científicos. Examinaremos sólo alguna visualización sencilla pero a menudo eficaz que complemente el análisis de las poblaciones estadísticas. Los datos estadísticos obtenidos de poblaciones grandes podrían ser muy útiles para estudiar el comportamiento de la distribución, si se presentan junto con recursos tabulares y gráficos conocidos como **diagramas de tallo y hojas**.

Para ejemplificar la elaboración de un diagrama de tallo y hojas, considere los datos de la tabla 1.4, que especifican la "vida" de 40 baterías para automóvil similares, registradas al décimo de año más cercano. Las baterías se garantizan por tres años. Primero, divida cada observación en dos partes: una para el tallo y otra para las hojas, de manera que el tallo represente el dígito entero que antecede al decimal, y la hoja corresponda a la parte decimal del número. En otras palabras, para el número 3.7 el dígito 3 designa al tallo; y el 7, a la hoja. Para nuestros datos los cuatro tallos 1, 2, 3 y 4 se listan verticalmente del lado izquierdo de la tabla 1.5; en tanto que las hojas se registran en el lado derecho correspondiente del valor del tallo adecuado. Entonces, la hoja 6 del número 1.6 se registra enfrente del tallo 1; la hoja 5 del número 2.5 enfrente del tallo 2; y así sucesivamente. El número de hojas registrado junto a cada uno de los tallos se anota debajo de la columna de frecuencia.

Tabla 1.4: Vida de las baterías para automóvil

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Tabla 1.5: Diagrama de tallo y hojas de la vida de las baterías

Tallo	**Hoja**	**Frecuencia**
1	69	2
2	25669	5
3	00111122233344455677889 9	25
4	11234577	8

El diagrama de tallo y hojas de la tabla 1.5 contiene tan sólo cuatro tallos y, por lo tanto, no ofrece una representación adecuada de la distribución. Para solucionar ese inconveniente, es necesario aumentar el número de tallos en nuestro diagrama. Una manera sencilla de hacerlo consiste en escribir dos veces cada valor del tallo y después registrar las hojas 0, 1, 2, 3 y 4 enfrente del valor del tallo adecuado, donde aparezca por primera vez; y las hojas 5, 6, 7, 8 y 9 enfrente de este mismo valor del tallo, donde aparece la segunda vez. El diagrama doble de tallo y hojas modificado

se ilustra en la tabla 1.6, donde a los tallos que corresponden a las hojas 0 a 4 se les anotó un símbolo $\star$, y al tallo correspondiente a las hojas 5 a 9, el símbolo $\cdot$.

En cualquier problema específico, debemos decidir cuáles son los valores del tallo adecuados. Se trata de una decisión que se toma algo arbitrariamente, aunque nos guiamos por el tamaño de nuestra muestra. Por lo general, elegimos entre 5 y 20 tallos. Cuanto menor sea el número de datos disponibles, menor será nuestra elección respecto del número de tallos. Por ejemplo, si los datos consisten en números del 1 al 21, los cuales representan el número de personas en la fila de una cafetería en 40 días laborables elegidos aleatoriamente y elegimos un diagrama doble de tallo y hojas, los tallos serían $0\star$, $0\cdot$, $1\star$, $1\cdot$ y $2\star$, de manera que la observación de 1 más pequeña tiene tallo $0\star$ y hoja 1, el número 18 tiene tallo $1\cdot$ y hoja 8, y la observación de 21 más grande tiene tallo $2\star$ y hoja 1. Por otro lado, si los datos consisten en números de \$18,800 a \$19,600 que representan las mejores ventas posibles de 100 automóviles nuevos, obtenidos de cierto concesionario, y elegimos un diagrama sencillo de tallo y hojas, los tallos serían 188, 189, 190, . . . , y 196, y las hojas contendrían ahora dos dígitos cada una. Un automóvil que se vende en \$19.385 tendría un valor de tallo de 193 y 85 en los dos dígitos de la hoja. En el diagrama de tallo y hojas las hojas de dígitos múltiples que pertenecen al mismo tallo, por lo general, están separadas por comas. En los datos generalmente se ignoran los puntos decimales cuando todos los números a la derecha del punto decimal representan hojas, como en el caso de las tablas 1.5 y 1.6. Sin embargo, si los datos consisten en números que van de 21.8 a 74.9, podríamos elegir los dígitos 2, 3, 4, 5, 6 y 7 como nuestros tallos, de manera que un número como, por ejemplo, 48.3 tendría un valor de tallo de 4, y un valor de hoja de 8.3.

Tabla 1.6: Diagrama doble de tallo y hojas para la vida de las baterías

Tallo	**Hoja**	**Frecuencia**
$1\cdot$	69	2
$2\star$	2	1
$2\cdot$	5669	4
$3\star$	001111222333444	15
$3\cdot$	5567778899	10
$4\star$	11234	5
$4\cdot$	577	3

El diagrama de tallo y hojas representa una manera eficaz de resumir los datos. Otra forma consiste en usar la **distribución de frecuencias**, donde los datos, agrupados en diferentes clases o intervalos, se pueden construir contando las hojas que pertenecen a cada tallo y considerando que cada tallo define un intervalo de clase. En la tabla 1.5 el tallo 1 con 2 hojas define el intervalo 1.0-1.9 que contiene 2 observaciones; el tallo 2 con 5 hojas define el intervalo 2.0-2.9 que contiene 5 observaciones; el tallo 3 con 25 hojas define el intervalo 3.0-3.9 con 25 observaciones; y el tallo 4 con 8 hojas define el intervalo 4.0-4.9 que contiene 8 observaciones. Para el diagrama doble de tallo y hojas de la tabla 1.6 los tallos definen los siete intervalos de clase 1.5-1.9, 2.0-2.4, 2.5-2.9, 3.0-3.4, 3.5-3.9, 4.0-4.4 y 4.5-4.9 con frecuencias 2, 1, 4, 15, 10, 5 y 3, respectivamente. Al dividir cada frecuencia de clase entre el número total de observaciones, obtenemos la proporción del conjunto de observaciones en cada una de las clases. Una tabla que lista las frecuencias relativas se denomina

distribución de frecuencias relativas. La distribución de frecuencias relativas para los datos de la tabla 1.4, que muestra los puntos medios de cada intervalo de clase, se presenta en la tabla 1.7.

Tabla 1.7: Distribución de frecuencias relativas de la vida de las baterías

Intervalo de clase	**Punto medio de la clase**	**Frecuencia, f**	**Frecuencia relativa**
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

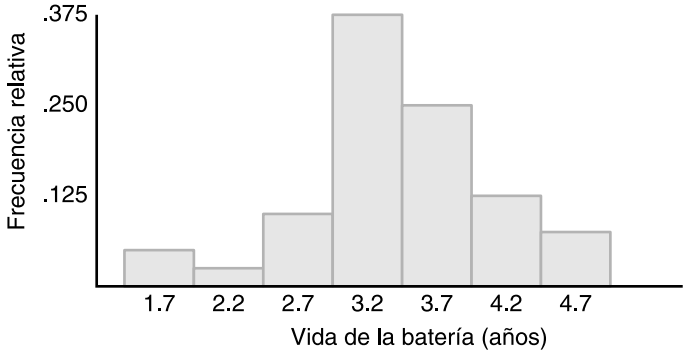


Figura 1.6: Histograma de frecuencias relativas.

La información que brinda una distribución de frecuencias relativas en forma tabular es más fácil de entender si se presenta en forma gráfica. Utilizando los puntos medios de cada intervalo y las frecuencias relativas correspondientes, construimos un **histograma de frecuencias relativas** (figura 1.6).

Muchas distribuciones de frecuencias continuas se representan gráficamente mediante la curva en forma de campana característica de la figura 1.7. Herramientas gráficas como las de las figuras 1.6 y 1.7 ayudan a comprender la naturaleza de la población. En los capítulos 5 y 6 examinaremos una propiedad de la población que se conoce como su **distribución**. Mientras que una definición más precisa de una distribución o de **distribución de probabilidad** se examinará más adelante en este texto, ahora podemos visualizarla como lo que habría sido el límite de la figura 1.7, conforme el tamaño de la muestra se vuelve más grande.

Se dice que una distribución en **simétrica** si se puede doblar a lo largo de un eje vertical, de manera que ambos lados coincidan. Una distribución que carece de simetría respecto de un eje vertical es **asimétrica** o sesgada. Entonces, la distribución que se ilustra en la figura 1.8a está sesgada porque tiene una cola derecha larga y una cola izquierda mucho más corta. En la figura 1.8b observamos que la distribución es simétrica; mientras que en la figura 1.8c está sesgada a la izquierda.

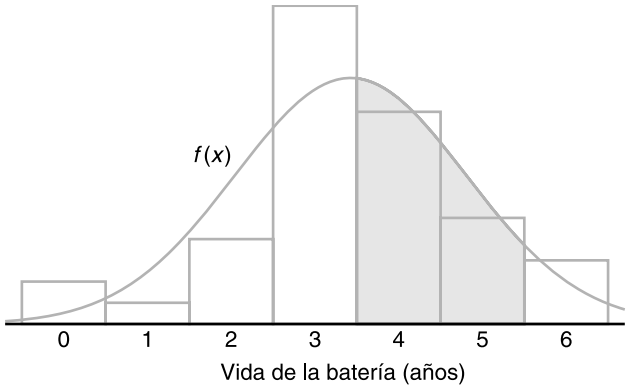


Figura 1.7: Estimación de la distribución de frecuencias.

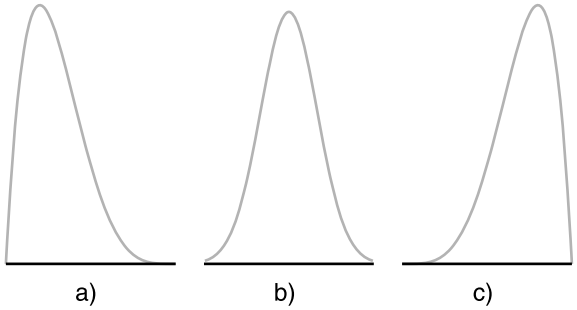


Figura 1.8: Asimetría de los datos.

Al girar un diagrama de tallo y hojas en dirección contraria a la de las manecillas del reloj en un ángulo de 90°, vemos que las columnas de hojas que resultan forman una imagen parecida a un histograma. Por lo tanto, si al observar los datos nuestro objetivo principal es determinar la forma general de la distribución, rara vez será necesario construir un histograma de frecuencias relativas. Se utilizan otros tipos diferentes de recursos y herramientas gráficas, los cuales se estudiarán en el capítulo 8, cuando presentemos detalles teóricos adicionales.

Otras características distintivas de una muestra

Hay características de la distribución o de la muestra a parte de las medidas del centro de ubicación y variabilidad que van más allá al definir su naturaleza. Por ejemplo, en tanto que la mediana divide los datos (o su distribución) en dos partes, existen otras medidas que dividen partes o piezas de la distribución que podrían resultar muy útiles. Una separación en cuatro partes se hace en *cuartiles*, donde el tercer cuartil separa el cuarto superior del resto de los datos, el segundo cuartil es la mediana y el primer cuartil separa el cuartil inferior del resto de los datos. Incluso la distribución puede dividirse más detalladamente calculando los percentiles de la distribución. Tales cantidades dan al analista una noción de las denominadas *colas* de la distribución (es decir, los valores que son relativamente extremos, ya sean pequeños o grandes). Por ejemplo, el 95º. percentil separa el 5% superior del 95% inferior. Definiciones similares prevalecen para los extremos en el lado inferior o

cola inferior de la distribución. El 1er percentil separa el 1% inferior del resto de la distribución. El concepto de percentiles tendrá un papel significativo en buena parte de lo que estudiaremos en los siguientes capítulos.

1.9 Tipos generales de estudios estadísticos: Diseño experimental, estudio observacional y estudio retrospectivo

En las siguientes secciones destacaremos la noción de muestreo de una población y el uso de los métodos estadísticos para aprender o quizá para reafirmar la información relevante acerca de una población. La información que se busca y que se obtiene mediante el uso de tales métodos estadísticos a menudo llega a influir en la toma de decisiones, así como en la resolución de problemas en diversas áreas importantes de ingeniería y científicas. Como ilustración, el ejemplo 1.3 describe un experimento sencillo, en el cual los resultados brindan ayuda para determinar los tipos de condiciones bajo las cuales se recomienda utilizar una aleación de aluminio específica, para prevenir la vulnerabilidad riesgosa ante la corrosión. Los resultados serían útiles no sólo para quienes fabrican la aleación, sino también para los clientes que consideren adquirirla. Este caso, y muchos otros que se incluyen en los capítulos 13 a 15, resaltan el concepto de condiciones experimentales diseñadas o controladas (combinaciones de condiciones de recubrimiento y humedad), que son de interés para aprender sobre algunas características o mediciones (nivel de corrosión) que surgen de tales condiciones. En el estudio de la corrosión se emplean métodos estadísticos que utilizan tanto medidas de tendencia central como de variabilidad. Como usted verá más adelante en este texto, tales métodos con frecuencia nos guían hacia un modelo estadístico como el que se examinó en la sección 1.7. En este caso, el modelo puede usarse para estimar (o predecir) las medidas de la corrosión como una función de la humedad y el tipo de recubrimiento utilizado. De nuevo, para desarrollar este tipo de modelos su vuelve muy útil emplear la estadística descriptiva que destaca las medidas de tendencia central y de variabilidad.

La información que se ofrece en el ejemplo 1.3 ilustra significativamente los tipos de preguntas de ingeniería que se plantean y se responden usando los métodos estadísticos que son útiles para el diseño experimental y que se presentan en este texto. Tales preguntas son las siguientes:

i. ¿Cuál es la naturaleza de la influencia de la humedad relativa sobre la corrosión de la aleación de aluminio dentro del rango de humedad relativa en este experimento?

ii. ¿El recubrimiento químico contra la corrosión reduce los niveles de corrosión y el efecto puede cuantificarse de alguna manera?

iii. ¿Hay **interacción** entre el tipo de recubrimiento y la humedad relativa que influya en la corrosión de la aleación? Si es así, ¿cual sería su interpretación?

¿Qué es interacción?

La importancia de las preguntas **i.** y **ii.** debería ser clara para el lector, en la medida en que tienen que ver con aspectos importantes tanto para los productores como para los usuarios de la aleación. ¿Y qué sucede con la pregunta **iii.**? El concepto de *interacción* se estudiará con detalle en los capítulos 14 y 15. Considere la gráfica de la figura 1.3. Se trata de un caso de detección de la interacción entre dos **factores** en un diseño experimental simple. Note que las líneas que conectan las medias

de la muestra no son paralelas. El **paralelismo** habría indicado que el efecto (visto como un resultado de la pendiente de las líneas) de la humedad relativa de la humedad relativa es el mismo, es decir, un efecto negativo, tanto para una condición sin recubrimiento como para otra con recubrimiento químico contra la corrosión. Recuerde que la pendiente "negativa" implica que la corrosión se vuelve más significativa conforme se incrementa la humedad. La ausencia de paralelismo implica una interacción entre el tipo de recubrimiento y la humedad relativa. A diferencia de la pendiente más pronunciada para la condición sin recubrimiento, la línea casi "horizontal" para el recubrimiento contra la corrosión sugiere que *no sólo el recubrimiento químico contra la corrosión es benéfico (note el desplazamiento entre las líneas), sino que la presencia del recubrimiento ilustra el efecto de la humedad despreciable.* Claramente, todas estas cuestiones son muy importantes para el efecto de los dos factores individuales y para la interpretación de la interacción, si está presente.

Los modelos estadísticos son bastante útiles para responder preguntas como las numeradas i, ii y iii anteriormente, donde los datos se obtienen de un diseño experimental. Sin embargo, uno no siempre cuenta con el tiempo o los recursos que permiten el uso de un diseño experimental. Por ejemplo, hay muchos casos en que las condiciones de interés para el científico o el ingeniero simplemente no pueden implementarse *debido a la imposibilidad de controlar los factores importantes.* En el ejemplo 1.3 la humedad relativa y el tipo de recubrimiento (o la ausencia de éste) son bastante fáciles de controlar. Desde luego, se trata del rasgo distintivo de un diseño experimental. En muchos campos, los factores que deben estudiarse no pueden ser controlados por cualesquiera diversas razones. Un control riguroso como el del ejemplo 1.3 permite al analista tener la confianza de que las diferencias encontradas (como en los niveles de corrosión) se deben a los factores que se controlan. Considere el ejercicio 1.6 de la página 14 como otro ejemplo. En este caso suponga que se elige 24 especímenes de caucho de silicón y 12 se asignan a cada uno de los niveles de temperatura de vulcanizado. Las temperaturas se controlan cuidadosamente, de manera que se trata de un ejemplo de diseño experimental con **solo factor**, que es la temperatura de vulcanizado. Se supondría que las diferencias encontradas en la resistencia a la tensión de la media son atribuibles a las diferentes temperaturas de vulcanizado.

¿Qué sucede si no se controlan los factores?

Suponga que los factores no se controlan y que *no hay asignación aleatoria* a los tratamientos específicos para las unidades experimentales, y que se busca deducir información a partir de un conjunto de datos. Como ejemplo considere el estudio realizado donde el interés se centra en la relación entre los niveles de colesterol sanguíneo y la cantidad de sodio medida en la sangre. Durante cierto periodo se monitoreó a un grupo de individuos, así como su colesterol sanguíneo y su sodio. En efecto, es posible obtener alguna información útil de tal conjunto de datos. No obstante, debería quedar claro que aquí ciertamente no hay control estricto de los niveles de sodio. De manera ideal, los sujetos deberían dividirse aleatoriamente en dos grupos, donde uno fuera el asignado a un "nivel alto" específico de sodio en la sangre, y el otro a un "nivel bajo" específico de sodio en la sangre. En efecto, esto no es posible. Evidentemente los cambios en los niveles de colesterol se deben a cambios en uno o diversos factores que no se controlaron. Este tipo de estudio, sin control de factores, se denomina **estudio observacional** (o por observación), el cual la mayoría de las veces implica una situación en que los sujetos se observan a través del tiempo.

Los estudios biológicos y biomédicos a menudo son necesariamente de este tipo. Sin embargo, los estudios observacionales no se restringen a dichas áreas. Por ejemplo, considere un estudio diseñado para determinar la influencia de la temperatura ambiental sobre la energía eléctrica que consumen las instalaciones de una planta química. Indudablemente los niveles de la temperatura ambiental no pueden controlarse y, por lo tanto, la estructura de los datos tan sólo se monitorea a partir de los datos de la planta a través del tiempo.

Debería notarse que una diferencia básica entre un experimento bien diseñado y un estudio observacional es la dificultad para determinar los verdaderos causa y efecto en este último. Asimismo, las diferencias encontradas en la reacción fundamental (por ejemplo, niveles de corrosión, colesterol sanguíneo, consumo de energía eléctrica en una planta) podría deberse a otros factores subyacentes que no se controlaron. De manera ideal, en un diseño experimental, los *factores perturbadores* estarían compensados gracias al proceso de aleatoriedad. De hecho, los cambios en los niveles de colesterol sanguíneo podrían deberse a la ingestión de grasa, a la realización de actividad física, etcétera. El consumo de energía eléctrica podría estar afectado por la cantidad de bienes producidos o incluso por la calidad de éstos.

En los estudios observacionales otra desventaja que a menudo se ignora cuando se comparan con los experimentos cuidadosamente diseñados es que, a diferencia de éstos, los primeros están a merced de circunstancias naturales, ambientales u otras no controladas que influyen en los niveles de los factores de interés. Por ejemplo, en el estudio biomédico respecto de la influencia de los niveles de sodio en la sangre sobre el colesterol sanguíneo, es posible que, de hecho, haya una influencia significativa, pero que el conjunto de datos específico que se usa no implique una variación observada suficiente en los niveles de sodio a causa de la naturaleza del sujeto elegido. Evidentemente, en un diseño experimental, el analista elige y controla los niveles de los factores.

Un tercer tipo de estudio estadístico que podría ser muy útil, pero que tiene notables desventajas cuando se le compara con un experimento bien diseñado, es un **estudio retrospectivo**. Esta clase de estudio emplea estrictamente **datos históricos**, que se obtienen durante un periodo específico. Una ventaja evidente con los datos retrospectivos es que prácticamente no hay costo por recabar los datos. Sin embargo, como podría esperarse, también tiene desventajas claras:

- **i.** A menudo es cuestionable la validez y la confiabilidad de los datos históricos.
- **ii.** Si el tiempo es un aspecto relevante en la estructura de los datos podría haber datos faltantes.
- **iii.** Existirían errores en la recopilación de los datos que no se conocen.
- **iv.** De nuevo, como en el caso de los datos observacionales, no hay control en los niveles de las variables que se miden (es decir, en los factores que se estudian). De hecho, las variaciones que se encuentran en los datos históricos a menudo no son significativas para estudios actuales.

Estudios que no determinan relaciones entre variables

En la sección 1.7 se le dio cierto énfasis al modelado de las relaciones entre variables. Presentamos la noción de análisis de regresión, el cual se estudia en los capítulos 11 y 12, y se considera una forma del análisis de datos para los diseños experimentales que se examinarán en los capítulos 14 y 15. En la sección 1.7, un modelo que relaciona la resistencia a la tensión de la media de la población (la tela) con los porcentajes de algodón, se utilizó para ilustrar los 20 especímenes que representaban las unidades experimentales. En este caso, los datos provienen de un diseño experimental simple, en el que los porcentajes de algodón individuales fueron seleccionados por científicos.

Con frecuencia tanto los datos observacionales como los retrospectivos se utilizan con la finalidad de observar relaciones entre variables a través de procedimientos de construcción que se estudian en los capítulos 11 y 12. Mientras que, de hecho, las ventajas de los diseños experimentales se aplican cuando la finalidad es la construcción del modelo estadístico, hay muchas áreas en que no es posible diseñar experimentos, de manera que *habrá que utilizar los datos históricos u observacionales*. Aquí nos referimos al conjunto de datos históricos que se incluye en el ejercicio 12.9 de la página 454. El objetivo es construir un modelo que resulte en una ecuación o relación que vincule el consumo mensual de energía eléctrica con la temperatura ambiental promedio x_1, el número de días en el mes x_2, la pureza promedio del producto x_3 y las toneladas de bienes producidos x_4. Se trata de los datos históricos del año anterior.

Ejercicios

1.13 Un fabricante de componentes electrónicos se interesa en determinar el tiempo de vida de cierto tipo de batería. La que sigue es una muestra, en horas de vida:

123, 116, 122, 110, 175, 126, 125, 111, 118, 117.

a) Encuentre la media y la mediana de la muestra.

b) ¿Qué característica en este conjunto de datos es la responsable de la diferencia sustancial entre ambas?

1.14 Un fabricante de neumáticos quiere determinar el diámetro interior de un neumático de cierto grado de calidad. Idealmente el diámetro sería de 570 mm. Los datos son los siguientes:

572, 572, 573, 568, 569, 575, 565, 570.

a) Encuentre la media y la mediana de la muestra.

b) Encuentre la varianza, la desviación estándar y el rango de la muestra.

c) Usando los estadísticos calculados en los incisos *a*) y *b*) ¿qué comentaría acerca de la calidad de los neumáticos?

1.15 Cinco lanzamientos independientes de una moneda tienen como resultado *cinco caras*. Resulta que si la moneda es legal, la probabilidad de este resultado es $(1/2)^5 = 0.03125$. ¿Produce esto evidencia sólida de que la moneda no sea legal? Comente y utilice el concepto de valor-P que se discutió en la sección 1.2.

1.16 Muestre que las n piezas de información en $\sum_{i=1}^{n} (x_1 - \bar{x}_2)^2$ no son independientes; es decir, muestre que

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

1.17 Se realiza un estudio acerca de los efectos del tabaquismo sobre los patrones de sueño. La medición que se observa es el tiempo, en minutos, que toma quedar dormido. Se obtienen estos datos:

Fumadores:	69.3	56.0	22.1	47.6
	53.2	48.1	52.7	34.4
	60.2	43.8	23.2	13.8
No fumadores:	28.6	25.1	26.4	34.9
	29.8	28.4	38.5	30.2
	30.6	31.8	41.6	21.1
	36.0	37.9	13.9	

a) Encuentre la media de la muestra para cada grupo.

b) Encuentre la desviación estándar de la muestra para cada grupo.

c) Usando una gráfica de puntos grafique los conjuntos de datos A y B en la misma línea.

d) Comente qué clase de impacto parece tener el hecho de fumar sobre el tiempo que se requiere para quedarse dormido.

1.18 Las siguientes puntuaciones representan la calificación en el examen final para un curso de estadística elemental:

23	60	79	32	57	74	52	70	82
36	80	77	81	95	41	65	92	85
55	76	52	10	64	75	78	25	80
98	81	67	41	71	83	54	64	72
88	62	74	43	60	78	89	76	84
48	84	90	15	79	34	67	17	82
69	74	63	80	85	61			

a) Elabore un diagrama de tallo y hojas para las calificaciones del examen, donde los tallos sean 1, 2, $3, \ldots, 9$.

b) Determine una distribución de frecuencias relativas.

c) Elabore un histograma de frecuencias relativas, trace un estimado de la gráfica de la distribución y discuta la asimetría de la distribución.

d) Calcule la media, la mediana y la desviación estándar de la muestra.

1.19 Los siguientes datos representan la duración de vida, en años, medida al décimo más cercano, de 30 bombas de combustible similares.

2.0	3.0	0.3	3.3	1.3	0.4
0.2	6.0	5.5	6.5	0.2	2.3
1.5	4.0	5.9	1.8	4.7	0.7
4.5	0.3	1.5	0.5	2.5	5.0
1.0	6.0	5.6	6.0	1.2	0.2

a) Construya un diagrama de tallo y hojas para la vida, en años, de las bombas de combustible, utilizando el dígito a la izquierda del punto decimal como el tallo para cada observación.

b) Determine una distribución de frecuencias relativas.

c) Calcule la media, el rango y la desviación estándar de la muestra.

1.20 Los siguientes datos representan la duración de la vida, en segundos, de 50 moscas frutales que se someten a un nuevo aerosol en un experimento de laboratorio controlado.

17	20	10	9	23	13	12	19	18	24
12	14	6	9	13	6	7	10	13	7
16	18	8	13	3	32	9	7	10	11
13	7	18	7	10	4	27	19	16	8
7	10	5	14	15	10	9	6	7	15

a) Elabore un diagrama doble de tallo y hojas para el periodo de vida de las moscas, usando los tallos $0\star$, $0\cdot$, $1\star$, $1\cdot$, $2\star$, $2\cdot$ y $3\star$ de manera que los tallos codificados con los símbolos $\star$ y $\cdot$ se asocien, respectivamente, con las hojas 0 a 4 y 5 a 9.

b) Determine una distribución de frecuencias relativas.

c) Construya un histograma de frecuencias relativas.

d) Calcule la mediana.

1.21 El contenido de nicotina, en miligramos, en 40 cigarrillos de cierta marca se registraron como sigue:

1.09	1.92	2.31	1.79	2.28
1.74	1.47	1.97	0.85	1.24
1.58	2.03	1.70	2.17	2.55
2.11	1.86	1.90	1.68	1.51
1.64	0.72	1.69	1.85	1.82
1.79	2.46	1.88	2.08	1.67
1.37	1.93	1.40	1.64	2.09
1.75	1.63	2.37	1.75	1.69

a) Encuentre la media y la mediana de la muestra.

b) Calcule la desviación estándar de la muestra.

1.22 Los siguientes datos constituyen mediciones del diámetro de 36 cabezas de remache en centésimos de una pulgada.

6.72	6.77	6.82	6.70	6.78	6.70	6.62	6.75
6.66	6.66	6.64	6.76	6.73	6.80	6.72	6.76
6.76	6.68	6.66	6.62	6.72	6.76	6.70	6.78
6.76	6.67	6.70	6.72	6.74	6.81	6.79	6.78
6.66	6.76	6.76	6.72				

a) Calcule la media y la desviación estándar de la muestra.

b) Construya un histograma de frecuencias relativas para los datos.

c) Comente sobre si habría una indicación clara o no de que la muestra proviene de una población que describe una distribución en forma de campana.

1.23 En 20 automóviles elegidos aleatoriamente, se tomaron las emisiones de hidrocarburos en velocidad en vacío, en partes por millón (ppm), para modelos de 1980 y 1990.

odelos 1980:

141	359	247	940	882	494	306	210	105	880
200	223	188	940	241	190	300	435	241	380

odelos 1990:

140	160	20	20	223	60	20	95	360	70
220	400	217	58	235	380	200	175	85	65

a) Construya una gráfica de puntos como la de la figura 1.1.

b) Calcule la media de la muestra para los dos años y sobreponga las dos medias en las gráficas.

c) Comente sobre lo que indica la gráfica de puntos, respecto de si cambiaron o no las emisiones de la población de 1980 a 1990. Utilice el concepto de variabilidad en su respuesta a este inciso.

1.24 Los siguientes son datos históricos de los sueldos del personal (dólares por alumno en 30 escuelas seleccionadas de la región este de Estados Unidos a principios de la década de 1970).

3.79	2.99	2.77	2.91	3.10	1.84	2.52	3.22
2.45	2.14	2.67	2.52	2.71	2.75	3.57	3.85
3.36	2.05	2.89	2.83	3.13	2.44	2.10	3.71
3.14	3.54	2.37	2.68	3.51	3.37		

a) Calcule la media y la desviación estándar de la muestra.

b) Con los datos elabore un histograma de frecuencias relativas.

c) Construya un diagrama de tallo y hojas con los datos.

1.25 El siguiente conjunto de datos se relaciona con el ejercicio anterior y representa el porcentaje de las familias que se ubican en el nivel superior de ingresos en las mismas escuelas individuales y con el mismo orden del ejercicio 1.24.

72.2	31.9	26.5	29.1	27.3	8.6	22.3	26.5
20.4	12.8	25.1	19.2	24.1	58.2	68.1	89.2
55.1	9.4	14.5	13.9	20.7	17.9	8.5	55.4
38.1	54.2	21.5	26.2	59.1	43.3		

a) Calcule la media de la muestra.

b) Calcule la mediana de la muestra.

c) Construya un histograma de frecuencias relativas con los datos.

d) Determine la media recortada 10%. Compárela con los resultados de los incisos *a*) y *b*) y exprese su comentario.

1.26 Suponga que le interesa emplear los conjuntos de datos de los ejercicios 1.24 y 1.25 para derivar un

modelo que prediga los salarios del personal como una función del porcentaje de familias en un nivel alto de ingresos para los sistemas escolares actuales. Comente sobre cualquier desventaja de llevar a cabo este tipo de análisis.

1.27 Se realizó un estudio para determinar la influencia del desgaste, y, de un cojinete como una función de la carga, x, sobre el cojinete. Se utiliza un diseño experimental para este estudio. Se emplearon tres niveles de carga: 700 lb, 1000 lb y 1300 lb. Se utilizaron cuatro especímenes en cada nivel y las medias muestrales fueron, respectivamente, 210, 325 y 375.

a) Grafique el promedio de desgaste contra la carga.

b) A partir de la gráfica del inciso anterior, ¿parece que haya una relación entre desgaste y carga?

c) Suponga que tenemos los siguientes valores individuales de desgaste para cada uno de los cuatro especímenes en los respectivos niveles de carga.

	x		
	700	**1000**	**1300**
y_1	145	250	150
y_2	105	195	180
y_3	260	375	420
y_4	330	480	750
	$\bar{y}_1 = 210$	$\bar{y}_2 = 325$	$\bar{y}_3 = 375$

Grafique los resultados para todos los especímenes contra los tres valores de carga.

d) A partir de la gráfica del inciso anterior, ¿parece que haya una relación clara? Si su respuesta es diferente de la del inciso *b*), explique por qué.

1.28 En Estados Unidos y otros países muchas compañías de manufactura utilizan piezas moldeadas como componentes de un proceso. La contracción (encogimiento) a menudo es un problema importante, de manera que un dado de metal moldeado para una pieza se construye más grande que el tamaño nominal para considerar su contracción. En un estudio de moldeado por inyección se descubrió que la contracción está influida por múltiples factores, entre los cuales están la velocidad de la inyección en pies/segundo y la temperatura de moldeado en °C. Los dos conjuntos de datos siguientes muestran los resultados de un experimento diseñado, donde la velocidad de inyección se mantuvo a dos niveles ("bajo" y "alto") y la temperatura de moldeado se mantuvo constantemente en un nivel "bajo". La contracción de midió en cm $\times 10^4$.

Los valores de contracción a una velocidad de inyección baja fueron:

72.68	72.62	72.58	72.48	73.07
72.55	72.42	72.84	72.58	72.92

Los valores de contracción a una velocidad de inyección alta fueron:

71.62	71.68	71.74	71.48	71.55
71.52	71.71	71.56	71.70	71.50

a) Construya una gráfica de puntos para ambos conjuntos de datos en la misma gráfica. Sobre ésta indique ambas medias de la contracción, tanto para la velocidad de inyección baja como para la velocidad de inyección alta.

b) Con base en los resultados de la gráfica del inciso anterior, y considerando la ubicación de las dos medias y su sentido de variabilidad, ¿cuál es su conclusión respecto del efecto de la velocidad inyección sobre la contracción a una temperatura de moldeado "baja"?

1.29 Considere la situación del ejercicio 1.28; pero ahora utilice el siguiente conjunto de datos, en el cual la contracción se mide de nuevo a una velocidad de inyección baja y a una velocidad de inyección alta. Sin embargo, esta vez la temperatura de moldeado se aumenta a un nivel "alto" y se mantiene constante.

Los valores de contracción a una velocidad de inyección baja fueron:

76.20	76.09	75.98	76.15	76.17
75.94	76.12	76.18	76.25	75.82

Los valores de contracción a una velocidad de inyección alta fueron:

93.25	93.19	92.87	93.29	93.37
92.98	93.47	93.75	93.89	91.62

a) Como en el ejercicio 1.28, elabore una gráfica de puntos con ambos conjuntos de datos en la misma gráfica e identifique las dos medias (es decir, la contracción media para la velocidad de inyección baja y para la velocidad de inyección alta).

b) Como en el ejercicio 1.28, comente la influencia de la velocidad de inyección en la contracción para la temperatura de moldeado alta. Tome en cuenta la posición de las dos medias y la variabilidad de cada media.

c) Compare su conclusión en el inciso *b*) actual con la del inciso *b*) del ejercicio 1.28, en el cual la temperatura de moldeado se mantuvo a un nivel bajo. ¿Diría que hay interacción entre la velocidad de inyección y la temperatura de moldeado? Explique.

1.30 Utilice los resultados de los ejercicios 1.28 y 1.29 para crear una gráfica que ilustre la interacción evidente entre los datos. Use como guía la gráfica de la figura 1.3 del ejemplo 1.3. ¿El tipo de información encontrada en los ejercicios 1.28, 1.29 y 1.30 podría encontrarse en el caso de un estudio observacional, donde el analista no tiene control sobre la velocidad de inyección ni sobre la temperatura de moldeado? Explique.

Capítulo 2

Probabilidad

2.1 Espacio muestral

En el estudio de la estadística tratamos básicamente con la presentación e interpretación de **resultados fortuitos** que ocurren en un estudio planeado o en una investigación científica. Por ejemplo, al registrar el número de accidentes que ocurren mensualmente en la intersección de Driftwood Lane y Royal Oak Drive, con la finalidad de justificar la instalación de un semáforo; o al clasificar los artículos que salen de una línea de ensamble como "defectuosos" o "no defectuosos"; o al revisar el volumen de gas que se libera en una reacción química cuando se varía la concentración de un ácido. Por ello, el estadístico a menudo trata con datos experimentales, conteos o mediciones representativos, o quizá con **datos categóricos** que se podrían clasificar de acuerdo con algún criterio.

Nos referiremos a cualquier registro de información, ya sea numérico o categórico, como una **observación**. Así, los números 2, 0, 1 y 2, que representan el número de accidentes que ocurrieron cada mes, de enero a abril, durante el año pasado en la intersección de Driftwood Lane y Royal Oak Drive, constituyen un conjunto de observaciones. Asimismo, los datos categóricos N, D, N, N y D, que representan los artículos defectuosos o no defectuosos cuando se inspeccionan cinco artículos, se registran como observaciones.

Los estadísticos utilizan la palabra **experimento** para describir cualquier proceso que genere un conjunto de datos. Un ejemplo simple de experimento estadístico es el lanzamiento de una moneda al aire. En tal experimento sólo hay dos resultados posibles: cara o cruz. Otro experimento sería el lanzamiento de un misil y la observación de su velocidad en tiempos específicos. Las opiniones de los votantes respecto de un nuevo impuesto sobre ventas también se pueden considerar como observaciones de un experimento. Estamos particularmente interesados en las observaciones que se obtienen por la repetición del experimento varias veces. En la mayoría de los casos los resultados dependerán del azar y, por lo tanto, no se predicen con certeza. Si un químico realiza un análisis varias veces con las mismas condiciones, obtendrá diferentes medidas, que indican un elemento de probabilidad en el procedimiento experimental. Incluso cuando se lanza una moneda al aire de forma repetida, no podemos tener la certeza de que un lanzamiento dado tendrá cara como resultado. Sin embargo, conocemos el conjunto completo de posibilidades para cada lanzamiento.

Considerando el análisis realizado en la sección 1.9, deberíamos considerar el alcance del término *experimento*. Se revisaron tres tipos de estudios estadísticos y se dieron varios ejemplos de cada uno. En cada uno de los tres casos —*experimentos diseñados*, *estudios observacionales* y *estudios retrospectivos*—, el resultado final fue un conjunto de *datos* que, en efecto, está sujeto a la **incertidumbre**. Aunque sólo uno de ellos tiene la palabra *experimento* en su descripción, el proceso de generar los datos o el proceso de observarlos forman parte de un experimento. El estudio de la corrosión discutido en la sección 1.3 con seguridad implica un experimento cuyas mediciones de la corrosión representan los datos. El ejemplo de la sección 1.9, en el cual se observaron el colesterol y el sodio en la sangre de un conjunto de individuos, representó un estudio observacional (que es diferente de un experimento *diseñado*); e incluso el proceso de generación de datos y el resultado fueron inciertos. Así son los experimentos. Un tercer ejemplo de la sección 1.9 representó un estudio retrospectivo, en el cual se observaron datos históricos sobre el consumo mensual de energía eléctrica y el promedio mensual de la temperatura ambiental. Sin embargo, aun cuando los datos hayan estado archivados durante décadas, el proceso se considerará un experimento.

Definición 2.1: El conjunto de todos los resultados posibles de un experimento estadístico se llama **espacio muestral** y se representa con el símbolo S.

A cada resultado en un espacio muestral se le llama **elemento** o **miembro** del espacio muestral, o simplemente **punto muestral**. Si el espacio muestral tiene un número finito de elementos, podemos *listar* los miembros separados por comas y encerrarlos entre llaves. De esta forma, el espacio muestral S, de los resultados posibles cuando se lanza una moneda al aire, se escribe como

$$S = \{H, T\},$$

donde H y T corresponden a "caras" y "cruces", respectivamente.

Ejemplo 2.1: Considere el experimento de lanzar un dado. Si nos interesamos en el número que muestre en la cara superior, el espacio muestral sería

$$S_1 = \{1, 2, 3, 4, 5, 6\}.$$

Si nos interesamos sólo en si el número es par o impar, el espacio muestral es simplemente

$$S_2 = \{\text{par, impar}\}.$$

El ejemplo 2.1 ilustra el hecho de que se puede usar más de un espacio muestral para describir los resultados de un experimento. En este caso S_1 brinda más información que S_2. Si sabemos cuál elemento en S_1 ocurre, podremos indicar cuál resultado tiene lugar en S_2; no obstante, el conocimiento de lo que pasa en S_2 no ayuda mucho en la determinación de qué elemento ocurre en S_1. En general, se desea utilizar un espacio muestral que dé la mayor información acerca de los resultados del experimento. En algunos experimentos es útil listar los elementos del espacio muestral de forma sistemática utilizando un **diagrama de árbol**.

Ejemplo 2.2: Un experimento consiste en lanzar una moneda y después lanzarla una segunda vez si sale cara. Si sale cruz en el primer lanzamiento, entonces se lanza un dado una vez. Para listar los elementos del espacio muestral que proporcione la mayor información, construimos el diagrama de árbol de la figura 2.1. Las diversas trayectorias a lo largo de las ramas del árbol dan los distintos puntos muestrales. Al comenzar con la rama superior izquierda y movernos a la derecha a lo largo de la primera trayectoria, obtenemos el punto muestral HH, que indica la posibilidad de que ocurran caras en dos lanzamientos sucesivos de la moneda. Asimismo, el punto muestral $T3$ indica la posibilidad

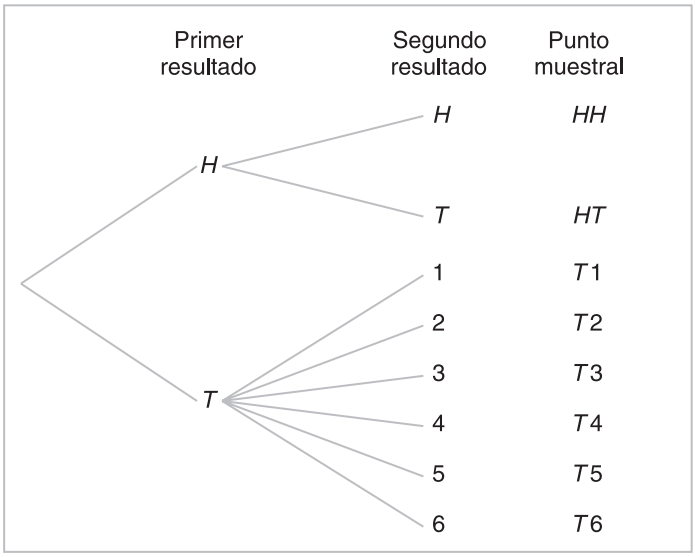


Figura 2.1: Diagrama de árbol para el ejemplo 2.2.

de que la moneda mostrará una cruz seguida por un *3* en el lanzamiento del dado. Al seguir a lo largo de todas las trayectorias, vemos que el espacio muestral es

$$S = \{HH, HT, T1, T2, T3, T4, T5, T6\}.$$

Muchos de los conceptos de este capítulo se ilustran mejor con ejemplos que tienen que ver con ilustraciones tales como el empleo de dados y cartas. Se trata de aplicaciones muy importantes para utilizar al principio del proceso de aprendizaje, lo cual nos permitirá el uso de los nuevos conceptos para avanzar con mayor facilidad el estudiar ejemplos de ciencia e ingeniería como el siguiente.

Ejemplo 2.3: Suponga que de un proceso de fabricación se seleccionan tres artículos de forma aleatoria. Cada artículo se inspecciona y clasifica como defectuoso, D, o sin defectos (no defectuoso), N. Para listar los elementos del espacio muestral que brinde la mayor información, construimos el diagrama de árbol de la figura 2.2, de manera que las diversas trayectorias a lo largo de las ramas del árbol dan los distintos puntos muestrales. Al comenzar con la primera trayectoria, obtenemos el punto muestral DDD, que indica la posibilidad de que los tres artículos inspeccionados estén defectuosos. Conforme continuamos a lo largo de las demás trayectorias, vemos que el espacio muestral es

$$S = \{DDD, DDN, DND, DNN, NDD, NDN, NND, NNN\}.$$

Los espacios muestrales con un número grande o infinito de puntos muestrales se describen mejor mediante un **enunciado** o **regla**. Por ejemplo, si los resultados posibles de un experimento son el conjunto de ciudades en el mundo con una población de más de un millón, nuestro espacio muestral se escribe como

$$S = \{x \mid x \text{ es una ciudad con una población de más de un millón}\},$$

que se lee "S es el conjunto de todas las x tales que x es una ciudad con una población de más de un millón". La barra vertical se lee "tal que". De manera similar, si S es el conjunto de todos los puntos (x, y) sobre la frontera o el interior de un círculo de radio 2 con centro en el origen, escribimos la **regla**

$$S = \{(x, y) \mid x^2 + y^2 \leq 4\}.$$

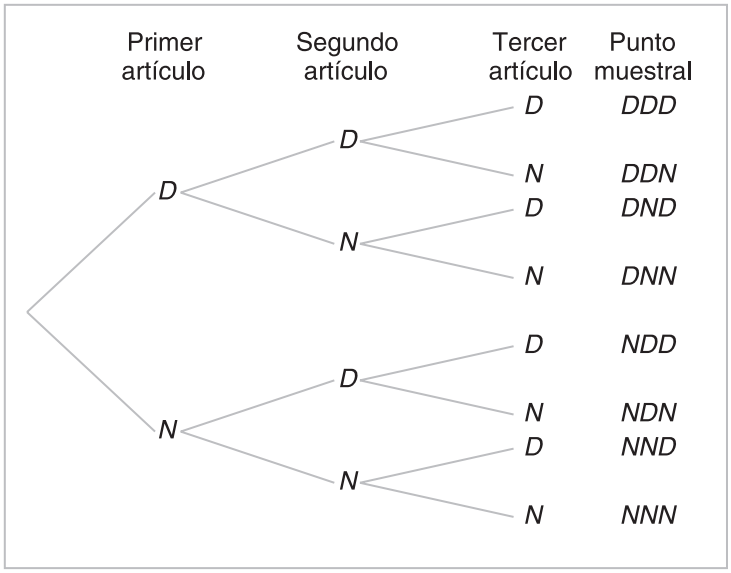


Figura 2.2: Diagrama de árbol para el ejemplo 2.3.

Si describimos el espacio muestral utilizando el método de la regla o listando los elementos dependerá del problema específico en cuestión. El método de la regla tiene ventajas prácticas, en especial para los diversos experimentos donde un listado se vuelve una tarea tediosa.

Considere la situación del ejemplo 2.3 donde los artículos que salen del proceso de fabricación están ya sea D, defectuoso, o N, sin defectos. Hay muchos procedimientos estadísticos importantes llamados planes de muestreo que determinan si un "lote" de artículos se considera satisfactorio o no. Un plan así implica tomar muestras hasta que se obtengan k artículos defectuosos. Suponga que el experimento consiste en tomar de forma aleatoria muestras de artículos hasta que salga uno defectuoso. En este caso, el espacio muestral sería

$$S = \{D,\, ND,\, NND,\, NNND, \ldots\}.$$

2.2 Eventos

Para cualquier experimento dado podemos enfocarnos en la ocurrencia de ciertos **eventos**, más que en el resultado de un elemento específico en el espacio muestral. Por ejemplo, quizás estemos interesados en el evento A, en el cual al lanzarse un dado el resultado es divisible entre 3. Éste ocurrirá si el resultado es un elemento del subconjunto $A = \{3, 6\}$ del espacio muestral S_1 del ejemplo 2.1. Como ilustración adicional, nos podemos interesar en el evento B de que el número de artículos defectuosos sea mayor que 1 en el ejemplo 2.3. Esto ocurrirá si el resultado es un elemento del subconjunto

$$B = \{DDN,\, DND,\, NDD,\, DDD\}$$

del espacio muestral S.

Para cada evento asignamos una colección de puntos muestrales, que constituye un subconjunto del espacio muestral. Ese subconjunto representa la totalidad de los elementos para los que el evento es cierto.

Definición 2.2: Un **evento** es un subconjunto de un espacio muestral.

Ejemplo 2.4: Dado el espacio muestral $S = \{t \mid t \geq 0\}$, donde t es la vida en años de cierto componente electrónico, entonces el evento A de que el componente falle antes de que finalice el quinto año es el subconjunto $A = \{t \mid 0 \leq t < 5\}$.

Es concebible que un evento sea un subconjunto que incluya todo el espacio muestral S, o un subconjunto de S que se denomina **conjunto vacío** y se denota con el símbolo ϕ, que no contiene elemento alguno. Por ejemplo, si hacemos que A sea el evento de detectar un organismo microscópico a simple vista en un experimento biológico, entonces $A = \phi$. También, si

$$B = \{x \mid x \text{ es un factor par de } 7\},$$

entonces B debe ser el conjunto vacío, pues los únicos factores posibles de 7 son los números nones 1 y 7.

Considere un experimento donde se registran los hábitos de fumar de los empleados de una compañía industrial. Un posible espacio muestral podría clasificar a un individuo como no fumador, fumador ligero, fumador moderado o fumador empedernido. Sea el subconjunto de los fumadores un evento. Entonces, la totalidad de los no fumadores corresponde a un evento diferente, también subconjunto de S, que se denomina complemento del conjunto de fumadores.

Definición 2.3: El **complemento** de un evento A respecto de S es el subconjunto de todos los elementos de S que no están en A. Denotamos el complemento de A mediante el símbolo A'.

Ejemplo 2.5: Sea R el evento de que se seleccione una carta roja de una baraja ordinaria de 52 cartas, y sea S toda la baraja. Entonces, R' es el evento de que la carta seleccionada de la baraja no sea una roja sino una negra.

Ejemplo 2.6: Considere el espacio muestral

$$S = \{\text{libro, catalizador, cigarrillo, precipitado, ingeniero, remache}\}.$$

Sea $A = \{\text{catalizador, remache, libro, cigarrillo}\}$. Entonces, el complemento de A es $A' = \{\text{precipitado, ingeniero}\}$.

Consideremos ahora ciertas operaciones con eventos que tendrán como resultado la formación de nuevos eventos. Tales eventos nuevos serán subconjuntos del mismo espacio muestral como los eventos dados. Suponga que A y B son dos eventos que se asocian con un experimento. En otras palabras, A y B son subconjuntos del mismo espacio muestral S. Por ejemplo, en el lanzamiento de un dado podemos hacer que A sea el evento de que ocurra un número par, y B el evento de que aparezca un número mayor que 3. Entonces, los subconjuntos $A = \{2, 4, 6\}$ y $B = \{4, 5, 6\}$ son subconjuntos del mismo espacio muestral

$$S = \{1, 2, 3, 4, 5, 6\}.$$

Note que *tanto* A como B ocurrirán en un lanzamiento dado, si el resultado es un elemento del subconjunto $\{4, 6\}$, el cual es precisamente la **intersección** de A y B.

Definición 2.4: La **intersección** de dos eventos A y B, que se denota con el símbolo $A \cap B$, es el evento que contiene todos los elementos que son comunes a A y a B.

Ejemplo 2.7: Sea C el evento de que una persona seleccionada al azar en un café Internet sea un estudiante universitario, y sea M el evento de que la persona sea hombre. Entonces $C \cap M$ es el evento de todos los estudiantes universitarios hombres en el café Internet.

Ejemplo 2.8: Sean $M = \{a, e, i, o, u\}$ y $N = \{r, s, t\}$; entonces, se sigue que $M \cap N = \phi$. Es decir, M y N no tienen elementos comunes y, por lo tanto, no pueden ocurrir ambos de forma simultánea.

Para ciertos experimentos estadísticos no es nada extraño definir dos eventos, A y B, que no pueden ocurrir de forma simultánea. Se dice entonces que los eventos A y B son **mutuamente excluyentes**. Expresado de manera más formal, tenemos la siguiente definición:

Definición 2.5: Dos eventos A y B son **mutuamente excluyentes** o **disjuntos** si $A \cap B = \phi$; es decir, si A y B no tienen elementos en común.

Ejemplo 2.9: Una compañía de televisión por cable ofrece programas en ocho diferentes canales, tres de los cuales están afiliados con ABC, dos con NBC y uno con CBS. Los otros dos son un canal educativo y el canal de deportes ESPN. Suponga que un individuo que se suscribe a este servicio enciende un televisor sin seleccionar de antemano el canal. Sea A el evento de que el programa pertenezca a la red NBC y B el evento de que pertenezca a la red CBS. Como un programa de televisión no puede pertenecer a más de una red, los eventos A y B no tienen programas en común. Por lo tanto, la intersección $A \cap B$ no contiene programa alguno y, en consecuencia, los eventos A y B son mutuamente excluyentes.

A menudo nos interesamos en la ocurrencia de al menos uno de dos eventos asociados con un experimento. Así, en el experimento del lanzamiento de un dado, si

$$A = \{2, 4, 6\} \text{ y } B = \{4, 5, 6\},$$

podemos interesarnos en que ocurran A o B, o en que ocurran tanto A como B. Tal evento, que se llama la unión de A y B, ocurrirá si el resultado es un elemento del subconjunto $\{2, 4, 5, 6\}$.

Definición 2.6: La **unión** de dos eventos A y B, que se denota con el símbolo $A \cup B$, es el evento que contiene todos los elementos que pertenecen a A o a B o a ambos.

Ejemplo 2.10: Sea $A = \{a, b, c\}$ y $B = \{b, c, d, e\}$, entonces, $A \cup B = \{a, b, c, d, e\}$.

Ejemplo 2.11: Sea P el evento de que un empleado seleccionado al azar de una compañía petrolera fume cigarrillos. Sea Q el evento de que el empleado seleccionado ingiera bebidas alcohólicas. Entonces, el evento $P \cup Q$ es el conjunto de todos los empleados que beben o fuman, o que hacen ambas cosas.

Ejemplo 2.12: Si $M = \{x \mid 3 < x < 9\}$ y $N = \{y \mid 5 < y < 12\}$, entonces,

$$M \cup N = \{z \mid 3 < z < 12\}.$$

La relación entre eventos y el correspondiente espacio muestral se puede ilustrar de forma gráfica utilizando **diagramas de Venn**. En un diagrama de Venn repre-

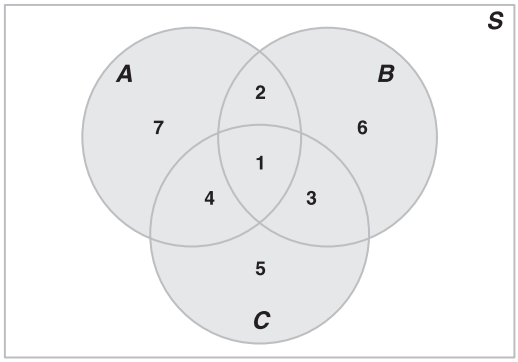


Figura 2.3: Eventos representados por varias regiones.

sentamos el espacio muestral como un rectángulo y los eventos con círculos trazados dentro del rectángulo. De esta forma, en la figura 2.3, vemos que

$$\begin{aligned} A \cap B &= \text{regiones 1 y 2,} \\ B \cap C &= \text{regiones 1 y 3,} \\ A \cup C &= \text{regiones 1, 2, 3, 4, 5 y 7,} \\ B' \cap A &= \text{regiones 4 y 7,} \\ A \cap B \cap C &= \text{región 1,} \\ (A \cup B) \cap C' &= \text{regiones 2, 6 y 7,} \end{aligned}$$

y así sucesivamente. En la figura 2.4 vemos que los eventos A, B y C son subconjuntos del espacio muestral S. También es claro que el evento B es un subconjunto del evento A; el evento $B \cap C$ no tiene elementos y, por ello, B y C son mutuamente excluyentes; el evento $A \cap C$ tiene al menos un elemento; y el evento $A \cup B = A$. La figura 2.4 puede, por lo tanto, representar una situación donde seleccionamos una carta al azar de una baraja ordinaria de 52 cartas y observamos si ocurren los siguientes eventos:

A: la carta es roja,

B: la carta es el jack, la reina o el rey de diamantes,

C: la carta es un as.

Claramente, el evento $A \cap C$ consiste sólo en los dos ases rojos.

Varios resultados que se derivan de las definiciones precedentes, y que se pueden verificar de forma sencilla empleando diagramas de Venn, son los que siguen:

1. $A \cap \phi = \phi$.

2. $A \cup \phi = A$.

3. $A \cap A' = \phi$.

4. $A \cup A' = S$.

5. $S' = \phi$.

6. $\phi' = S$.

7. $(A')' = A$.

8. $(A \cap B)' = A' \cup B'$.

9. $(A \cup B)' = A' \cap B'$.

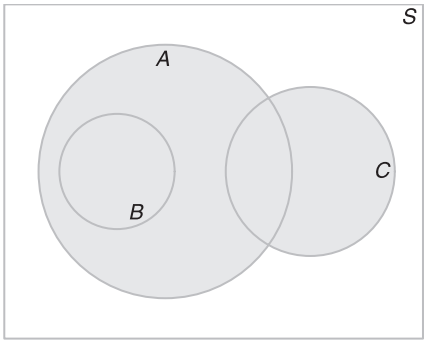


Figura 2.4: Eventos del espacio muestral S.

Ejercicios

2.1 Liste los elementos de cada uno de los siguientes espacios muestrales:

a) el conjunto de números enteros entre 1 y 50 que son divisibles entre 8;

b) el conjunto $S = \{x \mid x^2 + 4x - 5 = 0\}$;

c) el conjunto de resultados cuando se lanza una moneda al aire hasta que aparecen una cruz o tres caras;

d) el conjunto $S = (x \mid x$ es un continente);

e) el conjunto $S = \{x \mid 2x - 4 \geq 0$ y $x < 1\}$.

2.2 Utilice el método de la regla para describir el espacio muestral S, que consiste en todos los puntos del primer cuadrante dentro de un círculo de radio 3 con centro en el origen.

2.3 ¿Cuáles de los siguientes eventos son iguales?

a) $A = \{1, 3\}$;

b) $B = \{x \mid x$ es un número de un dado$\}$;

c) $C = \{x \mid x^2 - 4x + 3 = 0\}$;

d) $D = \{x \mid x$ es el número de caras cuando se lanzan seis monedas al aire$\}$.

2.4 Un experimento implica lanzar un par de dados, uno verde y uno rojo, y registrar los números que salen. Si x es igual al resultado en el dado verde y y es el resultado en el dado rojo, describa el espacio muestral S

a) mediante la lista de los elementos (x, y);

b) usando el método de la regla.

2.5 Un experimento consiste en lanzar un dado y después lanzar una moneda una vez, si el número en el dado es par. Si el número en el dado es impar, la moneda se lanza dos veces. Use la notación $4H$, por ejemplo, para denotar el resultado de que el dado muestre 4 y después la moneda salga cara, y $3HT$ para denotar el resultado de que el dado muestre 3 seguido por una cara y después una cruz en la moneda; construya un diagrama de árbol para mostrar los 18 elementos del espacio muestral S.

2.6 Se seleccionan dos jurados de cuatro suplentes para servir en un juicio por homicidio. Usando la notación A_1A_3, por ejemplo, para denotar el evento simple de que se seleccionen los suplentes 1 y 3, liste los 6 elementos del espacio muestral S.

2.7 Se seleccionan al azar cuatro estudiantes de una clase de química y se clasifican como masculino o femenino. Liste los elementos del espacio muestral S_1 usando de la letra M para "masculino", y F para "femenino". Defina un segundo espacio muestral S_2 donde los elementos representen el número de mujeres seleccionadas.

2.8 Para el espacio muestral del ejercicio 2.4:

a) liste los elementos que corresponden al evento A de que la suma sea mayor que 8;

b) liste los elementos que corresponden al evento B de que ocurra un 2 en cualquiera de los dos dados;

c) liste los elementos que corresponden al evento C de que salga un número mayor que 4 en el dado verde;

d) liste los elementos que corresponden al evento $A \cap C$,

e) liste los elementos que corresponden al evento $A \cap B$,

f) liste los elementos que corresponden al evento $B \cap C$,

g) construya un diagrama de Venn para ilustrar la intersecciones y uniones de los eventos A, B y C.

2.9 Para el espacio muestral del ejercicio 2.5:

a) liste los elementos que corresponden al evento A de que en el dado salga un número menor que 3;

b) liste los elementos que corresponden al evento B de que ocurran 2 cruces;

c) liste los elementos que corresponden al evento A',

d) liste los elementos que corresponden al evento $A' \cap B$,

e) liste los elementos que corresponden al evento $A \cup B$.

2.10 Se contrata a una firma de ingenieros para que determine si ciertas vías fluviales en Virginia son seguras para la pesca. Se toman muestras de tres ríos.

a) Liste los elementos de un espacio muestral S, y utilice las letras F para "seguro para la pesca", y N para "inseguro para la pesca".

b) Liste los elementos de S que correspondan al evento E de que al menos dos de los ríos son seguros para la pesca.

c) Defina un evento que tenga como sus elementos los puntos

$$\{FFF, NFF, FFN, NFN\}.$$

2.11 Los currícula de dos aspirantes masculinos para el puesto de profesor de química en una facultad se colocan en el mismo archivo que los currícula de dos aspirantes mujeres. Hay dos puestos disponibles y el primero, con el rango de profesor asistente, se cubre mediante la selección al azar de 1 de los 4 aspirantes. El segundo puesto, con el rango de profesor titular, se cubre después mediante la selección aleatoria de uno de los 3 aspirantes restantes. Utilizando la notación M_2F_1, por ejemplo, para denotar el evento simple de que el primer puesto se cubra con el segundo aspirante hombre y el segundo puesto se cubra después con la primera aspirante mujer:

a) liste los elementos de un espacio muestral S;

b) liste los elementos de S que corresponden al evento A de que el puesto de profesor asistente se cubra con un aspirante hombre;

c) liste los elementos de S que corresponden al evento B de que exactamente 1 de los 2 puestos se cubra con un aspirante hombre;

d) liste los elementos de S que corresponden al evento C de que ningún puesto se cubra con un aspirante hombre;

e) liste los elementos de S que corresponden al evento $A \cap B$,

f) liste los elementos de S que corresponden al evento $A \cup C$,

g) construya un diagrama de Venn para ilustrar las intersecciones y las uniones de los eventos A, B y C.

2.12 Se estudian el ejercicio y la dieta como posibles sustitutos de la medicación para bajar la presión sanguínea. Se utilizarán tres grupos de individuos para estudiar el efecto del ejercicio. El grupo uno es sedentario, mientras que el grupo dos camina, y el grupo tres nada una hora al día. La mitad de cada uno de los tres grupos de ejercicio tendrá una dieta sin sal. Un grupo adicional de individuos no hará ejercicio ni restringirá su consumo de sal, pero tomará la medicación estándar. Use Z para sedentario, W para caminante, S para nadador, Y para sal, N para sin sal, M para medicación, y F para sin medicamentos.

a) Muestre todos los elementos del espacio muestral S.

b) Dado que A es el conjunto de individuos sin medicamento y B es el conjunto de caminantes, liste los elementos de $A \cup B$.

c) Liste los elementos de $A \cap B$.

2.13 Construya un diagrama de Venn para ilustrar las posibles intersecciones y uniones para los siguientes eventos relativos al espacio muestral que consiste en todos los automóviles fabricados en Estados Unidos.

F: cuatro puertas, S: techo corredizo,
P: dirección hidráulica.

2.14 Si $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ y $A = \{0, 2, 4, 6, 8\}$, $B = \{1, 3, 5, 7, 9\}$, $C = \{2, 3, 4, 5\}$ y $D = \{1, 6, 7\}$, liste los elementos de los conjuntos que corresponden a los siguientes eventos:

a) $A \cup C$;

b) $A \cap B$;

c) C';

d) $(C' \cap D) \cup B$;

e) $(S \cap C)'$;

f) $A \cap C \cap D'$.

2.15 Considere el espacio muestral $S =$ {cobre, sodio, nitrógeno, potasio, uranio, oxígeno, cinc} y los eventos

$$A = \{\text{cobre, sodio, cinc}\},$$
$$B = \{\text{sodio, nitrógeno, potasio}\},$$
$$C = \{\text{oxígeno}\}.$$

Liste los elementos de los conjuntos que corresponden a los siguientes eventos:

a) A';

b) $A \cup C$;

c) $(A \cap B') \cup C'$;

d) $B' \cap C'$;

e) $A \cap B \cap C$;

f) $(A' \cup B') \cap (A' \cap C)$.

2.16 Si $S = \{x \mid 0 < x < 12\}$, $M = \{x \mid 1 < x < 9\}$, y $N = \{x \mid 0 < x < 5\}$, encuentre

a) $M \cup N$;

b) $M \cap N$;

c) $M' \cap N'$.

2.17 Sean A, B y C eventos relativos al espacio muestral S. Con el uso de diagramas de Venn, sombree las áreas que representan los siguientes eventos:

a) $(A \cap B)'$;

b) $(A \cup B)'$;

c) $(A \cap C) \cup B$.

2.18 ¿Cuál de los siguientes pares de eventos son mutuamente excluyentes?

a) Un golfista que se clasifica en último lugar en la vuelta del hoyo 18 en un torneo de 72 hoyos y pierde el torneo.

b) Un jugador de póquer que tiene flor (todas las cartas del mismo palo) y 3 de un tipo en la misma mano de 5 cartas.

c) Una madre que da a luz a una niña y un par de gemelas el mismo día.

d) Un jugador de ajedrez que pierde el último juego y gana el torneo.

2.19 Suponga que una familia sale de vacaciones de verano en su casa rodante y que M es el evento de que sufrirán fallas mecánicas, T es el evento de que recibirán una boleta de infracción por cometer una falta de tránsito y V es el evento de que llegarán a un lugar para acampar que esté lleno. Refiérase al diagrama de Venn de la figura 2.5, exprese con palabras los eventos representados por las siguientes regiones:

a) región 5;

b) región 3;

c) regiones 1 y 2 juntas;

d) regiones 4 y 7 juntas;

e) regiones 3, 6, 7 y 8 juntas.

2.20 Refiérase al ejercicio 2.19 y al diagrama de Venn de la figura 2.5, liste los números de las regiones que representan los siguientes eventos:

a) La familia no experimentará fallas mecánicas y no cometerá infracciones de tránsito, pero encontrará que el lugar para acampar estará lleno.

b) La familia experimentará tanto fallas mecánicas como problemas para localizar un lugar disponible para acampar, pero no recibirá una multa por infracción de tránsito.

c) La familia experimentará fallas mecánicas o encontrará un lugar para acampar lleno, pero no recibirá una multa por cometer una infracción de tránsito.

d) La familia no llegará a un lugar para acampar lleno.

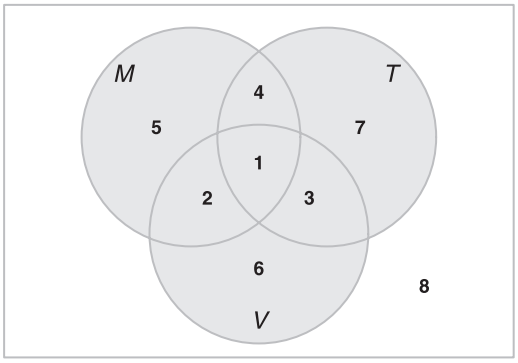


Figura 2.5: Diagrama de Venn para los ejercicios 2.19 y 2.20.

2.3 Conteo de puntos muestrales

Uno de los problemas que el estadístico debe considerar e intentar evaluar es el elemento de posibilidad asociado con la ocurrencia de ciertos eventos cuando se realiza un experimento. Estos problemas pertenecen al campo de la probabilidad, un tema que se estudiará en la sección 2.4. En muchos casos debemos ser capaces de resolver un problema de probabilidad mediante el conteo del número de puntos en el espacio muestral, sin listar realmente cada elemento. El principio fundamental del conteo, a menudo denominado **regla de multiplicación**, se establece como sigue:

Teorema 2.1: Si una operación se puede llevar a cabo en n_1 formas, y si para cada una de éstas se puede realizar una segunda operación en n_2 formas, entonces las dos operaciones se pueden ejecutar juntas de $n_1 n_2$ formas.

Ejemplo 2.13: ¿Cuántos puntos muestrales hay en el espacio muestral cuando un par de dados se lanza una vez?

Solución: El primer dado puede caer en cualquiera de $n_1 = 6$ maneras. Para cada una de esas 6 maneras el segundo dado también puede caer en $n_2 = 6$ formas. Por lo tanto, el par de dados puede caer en

$$n_1 n_2 = (6)(6) = 36 \text{ formas posibles.}$$

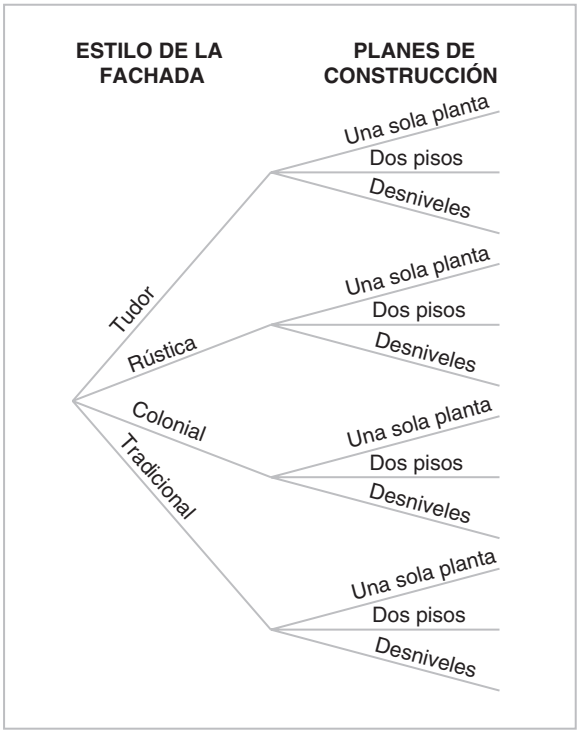


Figura 2.6: Diagrama de árbol para el ejemplo 2.14.

Ejemplo 2.14: Un urbanista de una nueva subdivisión ofrece a los futuros compradores de una casa la elección del estilo de la fachada entre Tudor, rústica, colonial y tradicional en una planta, dos pisos y desniveles. ¿En cuántas formas diferentes un comprador puede ordenar una de estas casas?

Solución: Como $n_1 = 4$ y $n_2 = 3$, un comprador debe elegir entre

$$n_1 n_2 = (4)(3) = 12 \text{ casas posibles.}$$

Las respuestas a los dos ejemplos anteriores se comprueba con la construcción de diagramas de árbol y el conteo de las diversas trayectorias a lo largo de las ramas.

Así, en el ejemplo 2.14 habrá $n_1 = 4$ ramas que corresponden a los diferentes estilos de la fachada, y después habrá $n_2 = 3$ ramas que se extienden de cada una de estas 4 ramas para representar los diferentes planes de plantas. Este diagrama de árbol da las $n_1n_2 = 12$ elecciones de casas dadas por las trayectorias a lo largo de las ramas, como se ilustra en la figura 2.6.

La regla de multiplicación del teorema 2.1 se puede extender para cubrir cualquier número de operaciones. Por ejemplo, suponga, que un cliente desea instalar un teléfono de AT&T^{TM} y puede elegir entre $n_1 = 10$ colores decorativos, que supondremos disponibles en cualquiera de $n_2 = 3$ longitudes opcionales de cordón, con $n_3 = 2$ tipos de marcado, es decir, de disco o por tonos. Estas tres clasificaciones dan como resultado $n_1n_2n_3 = (10)(3)(2) = 60$ diferentes formas para que un cliente ordene uno de estos teléfonos. La **regla de multiplicación generalizada** que cubre k operaciones se formula en el siguiente teorema.

Teorema 2.2: Si una operación se puede ejecutar en n_1 formas, y si para cada una de éstas se puede llevar a cabo una segunda operación en n_2 formas, y para cada una de las primeras dos se puede realizar una tercera operación en n_3 formas, y así sucesivamente, entonces la serie de k operaciones se puede realizar en $n_1n_2 \cdots n_k$ formas.

Ejemplo 2.15: Sam va a armar una computadora por sí mismo. Tiene la opción de comprar los chips entre dos marcas, un disco duro de cuatro marcas, la memoria de tres marcas y un conjunto de accesorios en cinco tiendas locales. ¿De cuántas formas diferentes puede Sam comprar las partes?

Solución: Como $n_1 = 2$, $n_2 = 4$, $n_3 = 3$ y $n_4 = 5$, hay

$$n_1 \times n_2 \times n_3 \times n_4 = 2 \times 4 \times 3 \times 5 = 120$$

formas diferentes de comprar las partes.

Ejemplo 2.16: ¿Cuántos números pares de cuatro dígitos se pueden formar con los dígitos 0, 1, 2 5, 6 y 9, si cada dígito se puede usar sólo una vez?

Solución: Como el número debe ser par, tenemos sólo $n_1 = 3$ elecciones para la posición de las unidades. Sin embargo, para un número de cuatro dígitos la posición de los millares no puede ser 0. Por lo tanto, consideramos la posición de las unidades en dos partes: 0 o diferente de 0. Si la posición de las unidades es 0 (es decir, $n_1 = 1$), tenemos $n_2 = 5$ elecciones para la posición de los millares, $n_3 = 4$ para la posición de las centenas y $n_4 = 3$ para la posición de las decenas. Por lo tanto, formamos un total de

$$n_1n_2n_3n_4 = (1)(5)(4)(3) = 60$$

números pares de cuatro dígitos. Por otro lado, si la posición de las unidades no es 0 (es decir, $n_1 = 2$), tenemos $n_2 = 4$ elecciones para la posición de los millares, $n_3 = 4$ para la posición de las centenas y $n_4 = 3$ para la posición de las decenas. En esta situación tenemos un total de

$$n_1n_2n_3n_4 = (2)(4)(4)(3) = 96$$

números pares de cuatro dígitos.

Puesto que los dos casos anteriores son mutuamente excluyentes entre sí, el número total de números pares de cuatro dígitos se calcula usando $60 + 96 = 156$.

Con frecuencia nos interesamos en un espacio muestral que contiene como elementos a todas las posibles ordenaciones o arreglos de un grupo de objetos. Por ejemplo, cuando queremos saber cuántos arreglos diferentes son posibles para sentar a seis personas alrededor de una mesa, o cuando nos preguntamos cuántas ordenaciones diferentes son posibles para sacar dos billetes de lotería de un total de 20. Los diferentes arreglos se llaman **permutaciones**.

Definición 2.7: Una permutación es un arreglo de todo o parte de un conjunto de objetos.

Considere las tres letras a, b y c. Las permutaciones posibles son abc, acb, bac, bca, cab y cba. De esta forma vemos que hay 6 arreglos distintos. Con el uso del teorema 2.2 podemos llegar a la respuesta 6 sin realmente listar las diferentes ordenaciones. Hay $n_1 = 3$ elecciones para la primera posición, después $n_2 = 2$ para la segunda, lo que deja sólo $n_3 = 1$ elección para la última posición, lo que da un total de

$$n_1 n_2 n_3 = (3)(2)(1) = 6 \text{ permutaciones.}$$

En general, n objetos distintos se pueden arreglar en

$$n(n-1)(n-2)\cdots(3)(2)(1) \text{ formas.}$$

Representamos este producto mediante el símbolo $n!$, que se lee "n factorial". Tres objetos se pueden arreglar en $3! = (3)(2)(1) = 6$ maneras. Por definición, $1! = 1$. También definimos $0! = 1$.

Teorema 2.3: El número de permutaciones de n objetos distintos es $n!$.

El número de permutaciones de las cuatro letras a, b, c y d será $4! = 24$. Consideremos ahora el número de permutaciones que son posibles al tomar, de las cuatro letras, dos a la vez. Éstas serían ab, ac, ad, ba, bc, bd, ca, cb, cd, da, db y dc. De nuevo, usando el teorema 2.1, tenemos dos posiciones para llenar con $n_1 = 4$ elecciones para la primera y después $n_2 = 3$ elecciones para la segunda, para un total de

$$n_1 n_2 = (4)(3) = 12$$

permutaciones. En general, n objetos distintos tomados de r a la vez se pueden arreglar en

$$n(n-1)(n-2)\cdots(n-r+1)$$

formas. Representamos este producto mediante el símbolo

$$_nP_r = \frac{n!}{(n-r)!}.$$

Como resultado tenemos el teorema que sigue.

Teorema 2.4: El número de permutaciones de n objetos distintos tomados de r a la vez es

$$_nP_r = \frac{n!}{(n-r)!}.$$

Ejemplo 2.17: En un año se otorgarán tres premios (a la investigación, la enseñanza y el servicio) en un grupo de 25 estudiantes de posgrado del departamento de estadística. Si cada estudiante puede recibir un premio como máximo, ¿cuántas selecciones posibles habría?

Solución: Como los premios son distinguibles, se trata de un problema de permutación. El número total de puntos muestrales es

$$_{25}P_3 = \frac{25!}{(25-3)!} = \frac{25!}{22!} = (25)(24)(23) = 13,800.$$

Ejemplo 2.18: Se van a elegir a un presidente y a un tesorero de un club estudiantil compuesto por 50 personas. ¿Cuántas opciones diferentes de funcionarios son posibles si

a) no hay restricciones;

b) A participará sólo si él es el presidente;

c) B y C participarán juntos o no lo harán;

d) D y E no participarán juntos?

Solución: *a*) El número total de elecciones de los funcionarios, si no hay restricciones, es

$$_{50}P_2 = \frac{50!}{48!} = (50)(49) = 2450.$$

b) Como A participaría sólo si es el presidente, tenemos dos situaciones: **i.** A se elige como presidente, lo cual produce 49 resultados posibles; o **ii.** los funcionarios se eligen de entre las 49 personas restantes cuyo número de opciones es $_{49}P_2 = (49)(48) = 2352$. Por lo tanto, el número total de elecciones es $49 + 2352 = 2401$.

c) El número de selecciones cuando B y C participan juntos es 2. El número de selecciones cuando ni B ni C se eligen es $_{48}P_2 = 2256$. Por lo tanto, el número total de opciones en esta situación es $2 + 2256 = 2258$.

d) El número de selecciones cuando D participa como funcionario pero sin E es $(2)(48) = 96$, donde 2 es el número de posiciones que D puede tomar y 48 es el número selecciones de los otros funcionarios de las personas restantes en el club, excepto E. El número de selecciones cuando E participa como funcionario pero sin D también es $(2)(48) = 96$. El número de selecciones cuando tanto D como E no son elegidos es $_{48}P_2 = 2256$. Por lo tanto, el número total de opciones es $(2)(96) + 2256 = 2448$. Este problema también tiene otra solución corta: como D y E sólo pueden participar juntos de dos maneras, la respuesta es $2450 - 2 = 2448$.

Las permutaciones que ocurren al arreglar objetos en un círculo se llaman **permutaciones circulares**. Dos permutaciones circulares no se consideran diferentes a menos que los objetos correspondientes en los dos arreglos estén precedidos o seguidos por un objeto diferente, conforme avancemos en la dirección de las manecillas del reloj. Por ejemplo, si cuatro personas juegan *bridge*, no tenemos una permutación nueva si se mueven una posición en la dirección de las manecillas del reloj. Al considerar a una persona en una posición fija y arreglar a las otras tres de 3! formas, encontramos que hay seis arreglos distintos para el juego de *bridge*.

Teorema 2.5: El número de permutaciones de n objetos distintos arreglados en un círculo es $(n-1)!$.

Hasta aquí consideramos permutaciones de objetos distintos. Es decir, todos los objetos fueron por completo diferentes o distinguibles. Evidentemente, si las letras b y c son ambas iguales a x, entonces las 6 permutaciones de las letras a, b y c se convierten en axx, axx, xax, xax, xxa y xxa, de las cuales sólo 3 son diferentes. Por lo tanto, con 3 letras, en las que 2 son la misma, tenemos $3!/2! = 3$ permutaciones distintas. Con 4 letras diferentes a, b, c y d tenemos 24 permutaciones distintas. Si hacemos $a = b = x$ y $c = d = y$, podemos listar sólo las siguientes permutaciones distintas: $xxyy$, $xyxy$, $yxxy$, $yyxx$, $xyyx$ y $yxyx$. De esta forma tenemos $4!/(2!\,2!) = 6$ permutaciones distintas.

Teorema 2.6: El número de permutaciones distintas de n objetos de los que n_1 son de una clase, n_2 de una segunda clase, $\ldots$, n_k de una k-ésima clase es

$$\frac{n!}{n_1!n_2!\quad n_k!}.$$

Ejemplo 2.19: Durante un entrenamiento del equipo de fútbol americano de la universidad, el coordinador defensivo necesita tener a 10 jugadores parados en una fila. Entre estos 10 jugadores, hay 1 de primer año, 2 de segundo año, 4 de tercer año y 3 de cuarto año, respectivamente. ¿De cuántas formas diferentes se pueden arreglar en una fila, si sólo se distingue su nivel de clase?

Solución: Usando directamente el teorema 2.6, el número total de arreglos es

$$\frac{10!}{1!\,2!\,4!\,3!} = 12,600.$$

Con frecuencia nos interesa el número de formas de dividir un conjunto de n objetos en r subconjuntos denominados **celdas**. Se consigue una partición si la intersección de todo par posible de los r subconjuntos es el conjunto vacío ϕ, y si la unión de todos los subconjuntos da el conjunto original. El orden de los elementos dentro de una celda no tiene importancia. Considere el conjunto $\{a, e, i, o, u\}$. Las particiones posibles en dos celdas en las que la primera celda contenga 4 elementos y la segunda 1 elemento son

$$\{(a,e,i,o),(\ )\},\{(a,i,o,\ ),(e)\},\{(e,i,o,\ ),(a)\},\{(a,e,o,\ ),(i)\},\{(a,e,i,\ ),(o)\}.$$

Vemos que hay 5 formas de partir un conjunto de 4 elementos en dos subconjuntos o celdas que contengan 4 elementos en la primera celda y 1 en la segunda.

El número de particiones para esta ilustración se denota con la expresión

$$\binom{5}{4,1} = \frac{5!}{4!\,1!} = 5,$$

donde el número superior representa el número total de elementos y los números inferiores representan el número de elementos que van en cada celda. Establecemos esto de forma más general en el siguiente teorema.

Teorema 2.7: El número de formas de partir un conjunto de n objetos en r celdas con n_1 elementos en la primera celda, n_2 elementos en la segunda, y así sucesivamente, es

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \quad n_r!},$$

donde $n_1 + n_2 + \cdots + n_r = n$.

Ejemplo 2.20: ¿En cuántas formas se pueden asignar siete estudiantes de posgrado a una habitación de hotel triple y a dos dobles, durante su asistencia a una conferencia?

Solución: El número total de particiones posibles sería

$$\binom{7}{3, 2, 2} = \frac{7!}{3!\ 2!\ 2!} = 210.$$

En muchos problemas nos interesamos en el número de formas de seleccionar r objetos de n sin importar el orden. Tales selecciones se llaman **combinaciones**. Una combinación es realmente una partición con dos celdas, donde una celda contiene los r objetos seleccionados y la otra contiene los $(n - r)$ objetos restantes. El número de tales combinaciones, denotado con

$$\binom{n}{r, n-r}, \text{ por lo general, se reduce a } \binom{n}{r},$$

debido a que el número de elementos en la segunda celda debe ser $n - r$.

Teorema 2.8: El número de combinaciones de n objetos distintos tomados de r a la vez es

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Ejemplo 2.21: Un niño le pide a su mamá que le lleve cinco cartuchos de Game-Boy$^{\text{TM}}$ de su colección de 10 juegos de arcada y 5 de deportes. ¿Cuántas maneras hay en que su mamá le llevará 3 juegos de arcada y 2 de deportes, respectivamente?

Solución: El número de formas de seleccionar 3 cartuchos de 10 es

$$\binom{10}{3} = \frac{10!}{3!\ (10-3)!} = 120.$$

El número de formas de seleccionar 2 cartuchos de 5 es

$$\binom{5}{2} = \frac{5!}{2!\ 3!} = 10.$$

Utilizando la regla de la multiplicación del teorema 2.1 con $n_1 = 120$ y $n_2 = 10$, hay $(120)(10) = 1200$ formas.

Ejemplo 2.22: ¿Cuántos arreglos diferentes de letras se pueden hacer con las letras de la palabra STATISTICS?

Solución: Utilizando el mismo argumento de la discusión del teorema 2.8, en este ejemplo aplicamos en realidad el teorema 2.7 para obtener

$$\binom{10}{3,3,2,1,1} = \frac{10!}{3!\ 3!\ 2!\ 1!\ 1!} = 50,400.$$

Aquí tenemos 10 letras en total, donde 2 letras (S, T) aparecen tres veces cada una, la letra I aparece dos veces, y las letras A y C aparecen una vez cada una.

Ejercicios

2.21 A los participantes de una convención se les ofrecen seis recorridos a sitios de interés cada uno de los tres días. ¿De cuántas maneras se puede acomodar una persona para ir a uno de los recorridos planeados por la convención?

2.22 En un estudio médico los pacientes se clasifican en 8 formas de acuerdo con su tipo sanguíneo: AB^+, AB^-, A^+, A^-, B^+, B^-, O^- u O^-; y también de acuerdo con su presión sanguínea: baja, normal o alta. Encuentre el número de formas en las que se puede clasificar a un paciente.

2.23 Si un experimento consiste en lanzar un dado y después extraer una letra al azar del alfabeto inglés, ¿cuántos puntos habrá en el espacio muestral?

2.24 Los estudiantes de una universidad privada de humanidades se clasifican como estudiantes de primer año, de segundo año, de penúltimo año o de último año, y también de acuerdo con su género (hombres o mujeres). Encuentre el número total de clasificaciones posibles para los estudiantes de esa universidad.

2.25 Cierto calzado se recibe en 5 diferentes estilos y cada estilo está disponible en 4 colores distintos. Si la tienda desea mostrar pares de estos zapatos que muestren la totalidad de los diversos estilos y colores, ¿cuántos diferentes pares tendría que mostrar?

2.26 Un estudio en California concluyó que al seguir siete sencillas reglas para la salud, la vida de un hombre se puede prolongar 11 años en promedio y la vida de una mujer 7 años. Estas 7 reglas son: no fumar, hacer ejercicio, uso moderado del alcohol, dormir siete u ocho horas, mantener el peso adecuado, desayunar y no ingerir alimentos entre comidas. De cuántas formas puede una persona adoptar cinco de esas reglas a seguir:

a) ¿Si la persona actualmente infringe las siete reglas?

b) ¿Si la persona nunca bebe y siempre desayuna?

2.27 Un urbanista de un nuevo fraccionamiento ofrece a un futuro comprador de una casa la elección de 4 diseños, 3 diferentes sistemas de calefacción, un garaje o cobertizo, y un patio o un porche cubierto. ¿De cuántos planes diferentes dispone el comprador?

2.28 Un medicamento contra el asma se puede adquirir de 5 diferentes laboratorios en forma de líquido, comprimidos o cápsulas, todas en concentración normal o alta. ¿De cuántas formas diferentes un doctor puede recetar la medicina a un paciente que sufre de asma?

2.29 En un estudio económico de combustibles, cada uno de 3 autos de carreras se prueba con 5 marcas diferentes de gasolina en 7 lugares de prueba que se localizan en diferentes regiones del país. Si se utilizan 2 pilotos en el estudio y las pruebas se realizan una vez bajo cada uno de los distintos grupos de condiciones, ¿cuántas pruebas se necesitan?

2.30 ¿De cuántas formas distintas se puede responder una prueba de falso-verdadero que consta de 9 preguntas?

2.31 Si una prueba de opción múltiple consiste en 5 preguntas, cada una con 4 respuestas posibles de las cuales sólo 1 es correcta,

a) ¿de cuántas formas diferentes un estudiante puede elegir una respuesta a cada pregunta?

b) ¿de cuántas maneras un estudiante puede elegir una respuesta a cada pregunta y tener incorrectas todas las respuestas?

2.32 *a*) ¿Cuántas permutaciones distintas se pueden hacer con las letras de la palabra *columna*?

b) ¿Cuántas de estas permutaciones comienzan con la letra m?

2.33 Un testigo de un accidente de tránsito, en el cual huyó el culpable, dice a la policía que el número de la matrícula contenía las letras RLH seguidas de 3 dígitos, cuyo primer número es un 5. Si el testigo no puede recordar los últimos 2 dígitos, pero tiene la certeza de que los 3 eran diferentes, encuentre el número máximo de matrículas de automóvil que la policía tiene que verificar.

2.34 *a*) ¿De cuántas maneras se pueden formar 6 personas para abordar un autobús?

b) Si 3 personas específicas, de las 6, insisten en estar una después de la otra, ¿cuántas maneras son posibles?

c) Si 2 personas específicas, de las 6, rehúsan seguir una a la otra, ¿cuántas maneras son posibles?

2.35 Un contratista desea construir 9 casas, cada una con diferente diseño. ¿De cuántas formas puede colocar estas casas en una calle si hay 6 lotes en un lado de la calle y 3 en el lado opuesto?

2.36 *a*) ¿Cuántos números de tres dígitos se pueden formar con los dígitos 0, 1, 2, 3, 4, 5 y 6, si cada dígito se puede usar sólo una vez?

b) ¿Cuántos de estos números son impares?

c) ¿Cuántos son mayores que 330?

2.37 ¿De cuántas maneras se pueden sentar 4 niños y 5 niñas en una fila, si unos y otras se deben alternar?

2.38 Cuatro matrimonios compran 8 lugares en la misma fila para un concierto. ¿De cuántas maneras diferentes se pueden sentar

a) sin restricciones?

b) si cada pareja se sienta junta?

c) si todos los hombres se sientan juntos a la derecha de todas las mujeres?

2.39 En un concurso regional de ortografía, los 8 finalistas son 3 niños y 5 niñas. Encuentre el número de puntos muestrales en el espacio muestral S para el número de ordenamientos posibles al final del concurso para

a) los 8 finalistas;

b) las primeras 3 posiciones.

2.40 ¿De cuántas formas de pueden llenar las cinco posiciones iniciales en un equipo de baloncesto con 8 jugadores que pueden jugar cualquiera de las posiciones?

2.41 Encuentre el número de formas en que 6 profesores se pueden asignar a 4 secciones de un curso introductorio de psicología, si ningún profesor se asigna a más de una sección.

2.42 Se sacan 3 billetes de lotería para el primer, segundo y tercer premios de un grupo de 40 boletos. Encuentre el número de puntos muestrales en S para dar los 3 premios, si cada concursante sólo tiene un billete.

2.43 ¿De cuántas maneras se pueden plantar 5 árboles diferentes en un círculo?

2.44 ¿De cuántas formas se puede acomodar en círculo una caravana de ocho carretas que proviene de Arizona?

2.45 ¿Cuántas permutaciones distintas se pueden hacer con las letras de la palabra *infinito*?

2.46 ¿De cuántas maneras se pueden colocar 3 robles, 4 pinos y 2 arces a lo largo de la línea divisoria de una propiedad, si no se distingue entre árboles del mismo tipo?

2.47 Una universidad participa en 12 juegos de fútbol durante una temporada. ¿De cuántas formas puede el equipo terminar la temporada con 7 ganados, 3 perdidos y 2 empates?

2.48 Nueve personas se dirigen a esquiar en tres automóviles que llevan 2, 4 y 5 pasajeros, respectivamente. ¿De cuántas maneras es posible transportar a las 9 personas hasta el albergue en todos los autos?

2.49 ¿Cuántas formas hay para seleccionar a 3 candidatos de 8 recién graduados igualmente calificados para las vacantes de una empresa contable?

2.50 ¿Cuántas formas hay en que dos estudiantes no tengan la misma fecha de cumpleaños en un grupo de 60?

2.4 Probabilidad de un evento

Quizá fue la insaciable sed del juego lo que condujo al desarrollo temprano de la teoría de la probabilidad. En un esfuerzo por aumentar sus ganancias, algunos pidieron a los matemáticos que les proporcionaran las estrategias óptimas para los diversos juegos de azar. Algunos de los matemáticos que brindaron tales estrategias fueron Pascal, Leibniz, Fermat y James Bernoulli. Como resultado de este desarrollo inicial de la teoría de la probabilidad, la inferencia estadística, con todas sus predicciones y generalizaciones, se extiende más allá de los juegos de azar para abarcar muchos

otros campos asociados con los eventos aleatorios, como la política, los negocios, la predicción del clima y la investigación científica. Para que estas predicciones y generalizaciones sean razonablemente precisas, resulta esencial una comprensión de la teoría básica de la probabilidad.

¿Qué queremos decir cuando hacemos afirmaciones como "Juan probablemente ganará el torneo de tenis", o "tengo una oportunidad de cincuenta por ciento de obtener un número par cuando se lanza un dado", o "no tengo posibilidad de ganar en la lotería esta noche", o "la mayoría de nuestros graduados probablemente estará casados dentro de tres años"? En cada caso expresamos un resultado del cual no estamos seguros; pero debido a la información del pasado o a partir de una comprensión de la estructura del experimento, tenemos algún grado de confianza en la validez de la afirmación.

En el resto de este capítulo consideraremos sólo aquellos experimentos para los cuales el espacio muestral contiene un número finito de elementos. La probabilidad de la ocurrencia de un evento que resulta de tal experimento estadístico se evalúa utilizando un conjunto de números reales denominados **pesos** o **probabilidades**, que van de 0 a 1. Para todo punto en el espacio muestral asignamos una probabilidad tal que la suma de todas las probabilidades es 1. Si tenemos razón para creer que es bastante probable que ocurra cierto punto muestral cuando se lleva a cabo el experimento, la probabilidad que se le asigne debería ser cercana a 1. Por otro lado, una probabilidad cercana a cero se asigna a un punto muestral que no es probable que ocurra. En muchos experimentos, como lanzar una moneda o un dado, todos los puntos muestrales tienen la misma oportunidad de ocurrencia y se les asignan probabilidades iguales. Para puntos fuera del espacio muestral, es decir, para eventos simples que no es posible que ocurran, asignamos una probabilidad de cero.

Para encontrar la probabilidad de un evento A, sumamos todas las probabilidades que se asignan a los puntos muestrales en A. Esta suma se denomina probabilidad de A y se denota con $P(A)$.

Definición 2.8: La probabilidad de un evento A es la suma de los pesos de todos los puntos muestrales en A. Por lo tanto,

$$0 \leq P(A) \leq 1, \quad P(\phi) = 0, \quad \text{y} \quad P(S) = 1.$$

Además, si $A_1, A_2, A_3, \ldots$ es una serie de eventos mutuamente excluyentes, entonces

$$P_1(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots.$$

Ejemplo 2.23: Se lanza dos veces una moneda. ¿Cuál es la probabilidad de que ocurra al menos una cara?

Solución: El espacio muestral para este experimento es

$$S = \{HH,\ HT,\ TH,\ TT\}.$$

Si la moneda está balanceada, cada uno de estos resultados tendrá la misma probabilidad de ocurrencia. Por lo tanto, asignamos una probabilidad de ω a cada uno de los puntos muestrales. Entonces, $4\omega = 1$ o $\omega = 1/4$. Si A representa el evento de que ocurra al menos una cara, entonces

$$A = \{HH, HT, TH\} \quad \text{y} \quad P(A) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}.$$

Ejemplo 2.24: Se carga un dado de forma que sea dos veces más probable que salga un número par que uno non. Si E es el evento de que ocurra un número menor que 4 en un solo lanzamiento del dado, encuentre $P(E)$.

Solución: El espacio muestral es $S = \{1, 2, 3, 4, 5, 6\}$. Asignamos una probabilidad de w a cada número non y una probabilidad de $2w$ a cada número par. Como la suma de las probabilidades debe ser 1, tenemos $9w = 1$ o $w = 1/9$. Por ello se asignan probabilidades de $1/9$ y $2/9$ a cada número non y par, respectivamente. Por lo tanto,

$$E = \{1, 2, 3\} \text{ y } P(E) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}.$$

Ejemplo 2.25: En el ejemplo 2.24, sea A el evento de que salga un número par y sea B el evento de que salga un número divisible entre 3. Encuentre $P(A \cup B)$ y $P(A \cap B)$.

Solución: Para los eventos $A = \{2, 4, 6\}$ y $B = \{3, 6\}$, tenemos

$$A \cup B = \{2, 3, 4, 6\} \text{ y } A \cap B = \{6\}.$$

Al asignar una probabilidad de $1/9$ a cada número non y de $2/9$ a cada número par, tenemos

$$P(A \cup B) = \frac{2}{9} + \frac{1}{9} + \frac{2}{9} + \frac{2}{9} = \frac{7}{9} \quad \text{y} \quad P(A \cap B) = \frac{2}{9}.$$

Si el espacio muestral para un experimento contiene N elementos, todos los cuales tienen la misma probabilidad de ocurrencia, asignamos una probabilidad igual a $1/N$ a cada uno de los N puntos. La probabilidad de cualquier evento A que contenga n de estos N puntos muestrales es entonces la razón del número de elementos en A al número de elementos en S.

Teorema 2.9: Si un experimento puede tener como resultado cualquiera de N diferentes resultados igualmente probables, y si exactamente n de estos resultados corresponden al evento A, entonces la probabilidad del evento A es

$$P(A) = \frac{n}{N}.$$

Ejemplo 2.26: Una clase de estadística para ingenieros consta de 25 estudiantes de ingeniería industrial, 10 de mecánica, 10 de eléctrica y 8 de civil. Si el profesor elige a una persona al azar para que conteste una pregunta, encuentre la probabilidad de que el estudiante elegido sea *a*) un estudiante de ingeniería industrial, *b*) uno que de ingeniería civil o eléctrica.

Solución: Se denotan con I, M, E y C las especialidades de los estudiantes en ingenierías industrial, mecánica, eléctrica y civil, respectivamente. El número total de estudiantes en la clase es 53, todos los cuales tienen la misma probabilidad de ser seleccionados.

a) Como 25 de los 53 estudiantes tienen la especialidad en ingeniería industrial, la probabilidad del evento I, elegir al azar a alguien de ingeniería industrial, es

$$P(I) = \frac{25}{53}.$$

b) Como 18 de los 53 estudiantes son de las especialidades de ingeniería civil o eléctrica, se sigue que

$$P(C \cup E) = \frac{18}{53}.$$

Ejemplo 2.27: En una mano de póquer que consiste en 5 cartas, encuentre la probabilidad de tener 2 ases y 3 jacks.

Solución: El número de formas de tener 2 ases de 4 es

$$\binom{4}{2} = \frac{4!}{2!\ 2!} = 6,$$

y el número de formas de tener 3 jacks de 4 es

$$\binom{4}{3} = \frac{4!}{3!\ 1!} = 4.$$

Mediante la regla de multiplicación del teorema 2.1, hay $n = (6)(4) = 24$ manos con 2 ases y 3 jacks. El número total de manos de póquer de 5 cartas, las cuales son igualmente probables, es

$$N = \binom{52}{5} = \frac{52!}{5!\ 47!} = 2,598,960.$$

Por lo tanto, la probabilidad del evento C de obtener 2 ases y 3 jacks en una mano de póquer de 5 cartas es

$$P(C) = \frac{24}{2,598,960} = 0.9 \times 10^{-5}.$$

Si los resultados de un experimento no tienen igual probabilidad de ocurrencia, las probabilidades se deben asignar sobre la base de un conocimiento previo o de evidencia experimental. Por ejemplo, si una moneda no está balanceada, podemos estimar las probabilidades de caras y cruces al lanzar la moneda un número grande de veces, y registrar los resultados. De acuerdo con la definición de **frecuencia relativa** de la probabilidad, las probabilidades verdaderas serían las fracciones de caras y cruces que ocurren a largo plazo.

Para encontrar un valor numérico que represente de forma adecuada la probabilidad de ganar en el tenis, debemos depender de nuestro desempeño previo en el juego, así como también del de nuestro oponente y, hasta cierto punto, en nuestra creencia de ser capaces de ganar. De manera similar, para encontrar la probabilidad de que un caballo gane una carrera, debemos llegar a una probabilidad que se base en las marcas anteriores de todos los caballos que participan en la carrera, así como de las marcas de los jockeys que montan los caballos. La intuición, sin duda, también juega una parte en la determinación del monto de la apuesta que estemos dispuestos a arriesgar. El uso de la intuición, las creencias personales y otra información indirecta para llegar a probabilidades se denomina la definición **subjetiva** de la probabilidad.

En la mayoría de las aplicaciones de probabilidad de este libro la interpretación de frecuencia relativa de probabilidad es la que opera. Su fundamento es el experimento estadístico en vez de la subjetividad. Se le considera más bien como **frecuencia**

relativa limitante. Como resultado, muchas aplicaciones de probabilidad en ciencia e ingeniería se deben basar en experimentos que se puedan repetir. Nociones menos objetivas de probabilidad se encuentran cuando asignamos probabilidades que se basan en información y opiniones previas. Como ejemplo, "hay una buena oportunidad de que los Leones pierdan el Super Bowl". Cuando las opiniones y la información previa difieren de un individuo a otro, la probabilidad subjetiva se vuelve el recurso pertinente.

2.5 Reglas aditivas

A menudo resulta más sencillo calcular la probabilidad de algún evento a partir del conocimiento de las probabilidades de otros eventos. Esto puede ser cierto si el evento en cuestión se puede representar como la unión de otros dos eventos o como el complemento de algún evento. A continuación se presentan varias leyes importantes que con frecuencia simplifican el cálculo de las probabilidades. La primera, que se denomina **regla aditiva**, se aplica a uniones de eventos.

Teorema 2.10: Si A y B son dos eventos, entonces

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Prueba: Considere el diagrama de Venn de la figura 2.7. $P(A \cup B)$ es la suma de las probabilidades de los puntos muestrales en $A \cup B$. Así, $P(A) + P(B)$ es la suma de todas las probabilidades en A más la suma de todas las probabilidades en B. Por lo tanto, sumamos dos veces las probabilidades en $(A \cap B)$. Como estas probabilidades se suman a $P(A \cap B)$, debemos restar esta probabilidad una vez para obtener la suma de las probabilidades en $A \cup B$.

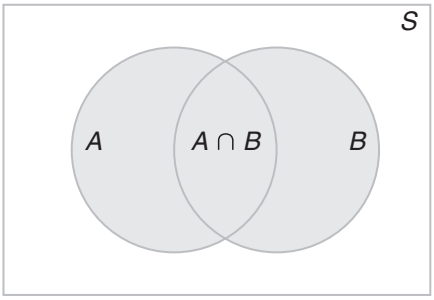


Figura 2.7: Regla aditiva de probabilidad.

Corolario 2.1: Si A y B son mutuamente excluyentes, entonces

$$P(A \cup B) = P(A) + P(B).$$

El corolario 2.1 es un resultado inmediato del teorema 2.10, pues si A y B son mutuamente excluyentes, $A \cap B = 0$ y entonces $P(A \cap B) = P(\phi) = 0$. En general, escribimos:

Corolario 2.2: Si $A_1, A_2, \ldots, A_n$ son mutuamente excluyentes, entonces

$$P(A_1 \cup A_2 \cup \quad \cup A_n) = P(A_1) + P(A_2) + \quad + P(A_n).$$

Una colección de eventos $\{A_1, A_2, \ldots A_n\}$ de un espacio muestral S se denomina una **partición** de S si $A_1, A_2, \ldots, A_n$ son mutuamente excluyentes y $A_1 \cup A_2 \cup \cdots \cup A_n = S$. Por lo tanto, tenemos

Corolario 2.3: Si $A_1, A_2, \ldots, A_n$ es una partición de un espacio muestral S, entonces

$$P(A_1 \cup A_2 \cup \quad \cup A_n) = P(A_1) + P(A_2) + \quad + P(A_n) = P(S) = 1.$$

Como se esperaría, el teorema 2.10 se extiende de forma análoga.

Teorema 2.11: Para tres eventos A, B y C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Ejemplo 2.28: Al final del semestre, Juan se va a graduar en la facultad de ingeniería industrial en una universidad. Después de tener entrevistas en dos compañías donde quiere trabajar, él evalúa la probabilidad que tiene de lograr una oferta de empleo en la compañía A como 0.8, y la probabilidad de obtenerla de la compañía B como 0.6. Si, por otro lado, considera que la probabilidad de que reciba ofertas de ambas compañías es 0.5, ¿cuál es la probabilidad de que obtendrá al menos una oferta de esas dos compañías?

Solución: Con la regla aditiva tenemos

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.8 + 0.6 - 0.5 = 0.9.$$

Ejemplo 2.29: ¿Cuál es la probabilidad de obtener un total de 7 u 11 cuando se lanza un par de dados?

Solución: Sea A el evento de que ocurra 7 y B el evento de que salga 11. Entonces, un total de 7 ocurre para 6 de los 36 puntos muestrales y un total de 11 ocurre sólo para 2. Como todos los puntos muestrales son igualmente probables, tenemos $P(A) = 1/6$ y $P(B) = 1/18$. Los eventos A y B son mutuamente excluyentes, pues un total de 7 y 11 no pueden ocurrir en el mismo lanzamiento. Por lo tanto,

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9}.$$

Este resultado también se podría obtener al contar el número total de puntos para el evento $A \cup B$, es decir 8, y escribir

$$P(A \cup B) = \frac{n}{N} = \frac{8}{36} = \frac{2}{9}.$$

El teorema 2.10 y sus tres corolarios deberían ayudar al lector a lograr una mejor comprensión de la probabilidad y de su interpretación. Los corolarios 1 y 2 sugieren el resultado muy intuitivo que trata con la probabilidad de ocurrencia de, al menos, uno de entre un número de eventos, sin que puedan ocurrir dos de ellos simultáneamente. La probabilidad de que al menos ocurra uno es la suma de las probabilidades de ocurrencia de los eventos individuales. El tercer corolario simplemente establece que el valor mayor de una probabilidad (uno) se asigna a todo el espacio muestral S.

Ejemplo 2.30: Si las probabilidades de que un individuo que compra un automóvil nuevo elija color verde, blanco, rojo o azul son, respectivamente, 0.09, 0.15, 0.21 y 0.23, ¿cuál es la probabilidad de que un comprador dado adquiera un automóvil nuevo que tenga uno de esos colores?

Solución: Sean G, W, R y B los eventos de que un comprador seleccione, respectivamente, un automóvil verde, blanco, rojo o azul. Como estos cuatro eventos son mutuamente excluyentes, la probabilidad es

$$\begin{aligned} P(G \cup W \cup R \cup B) &= P(G) + P(W) + P(R) + P(B) \\ &= 0.09 + 0.15 + 0.21 + 0.23 = 0.68. \end{aligned}$$

A menudo es más difícil calcular la probabilidad de que ocurra un evento que calcular la probabilidad de que el evento no ocurra. Si éste es el caso para algún evento A, simplemente encontramos primero $P(A')$ y, después, con el teorema 2.10, encontramos $P(A)$ por sustracción.

Teorema 2.12: Si A y A' son eventos complementarios, entonces

$$P(A) + P(A') = 1.$$

Prueba: Como $A \cup A' = S$ y los conjuntos A y A' son disjuntos, entonces

$$1 = P(S) = P(A \cup A') = P(A) + P(A').$$

Ejemplo 2.31: Si las probabilidades de que un mecánico automotriz dé servicio a 3, 4, 5, 6, 7, 8 o más vehículos en un día de trabajo dado son 0.12, 0.19, 0.28, 0.24, 0.10 y 0.07, respectivamente, ¿cuál es la probabilidad de que dé servicio al menos a 5 vehículos el siguiente día de trabajo?

Solución: Sea E el evento de que al menos 5 automóviles reciban servicio. Así, $P(E) = 1 - P(E')$, donde E' es el evento de que menos de 5 automóviles reciban servicio. Como

$$P(E') = 0.12 + 0.19 = 0.31,$$

del teorema 2.12 se sigue que

$$P(E) = 1 - 0.31 = 0.69.$$

Ejemplo 2.32: Suponga que las especificaciones del fabricante para la longitud del cable de cierto tipo de computadora son 2000 ± 10 milímetros. En esta industria, se sabe que el cable pequeño tiene la misma probabilidad de salir defectuoso (no cumplir con las especificaciones) que el cable grande. Es decir, la probabilidad de que aleatoriamente

se produzca un cable con una longitud mayor que 2010 milímetros es igual a la probabilidad de producirlo con una longitud menor que 1990 milímetros. Se sabe que la probabilidad de que el procedimiento de producción cumpla con las especificaciones es de 0.99.

a) ¿Cuál es la probabilidad de que un cable elegido aleatoriamente sea muy largo?

b) ¿Cuál es la probabilidad de que un cable elegido aleatoriamente sea mayor que 1990 milímetros?

Solución: Sea B el evento de que un cable cumpla con las especificaciones. Sean S y L los eventos de que el cable sea muy pequeño o muy grande, respectivamente. Entonces,

a) $P(M) = 0.99$ y $P(S) = P(L) = \frac{1-0.99}{2} = 0.005$.

b) Si se denota con X la longitud de un cable seleccionado aleatoriamente, tenemos

$$P(1990 \leq X \leq 2010) = P(M) = 0.99.$$

Como $P(X \geq 2010) = P(L) = 0.005$, entonces

$$P(X \geq 1990) = P(M) + P(L) = 0.995.$$

Lo cual también se resuelve utilizando el teorema 2.12:

$$P(X \geq 1990) + P(X < 1990) = 1.$$

Así, $P(X \geq 1990) = 1 - P(S) = 1 - 0.005 = 0.995$.

Ejercicios

2.51 Encuentre los errores en cada una de las siguientes aseveraciones:

a) Las probabilidades de que un vendedor de automóviles venda 0, 1, 2 o 3 unidades en un día dado de febrero son 0.19, 0.38, 0.29 y 0.15, respectivamente.

b) La probabilidad de que llueva mañana es 0.40 y la probabilidad de que no llueva es 0.52.

c) Las probabilidades de que una impresora cometa 0, 1, 2, 3 o 4 o más errores al imprimir un documento son 0.19, 0.34, −0.25, 0.43 y 0.29, respectivamente.

d) Al sacar una carta de una baraja en un solo intento la probabilidad de seleccionar corazones es 1/4, la probabilidad de seleccionar una carta negra es 1/2, y la probabilidad de seleccionar una carta negra de corazones es 1/8.

2.52 Suponga que todos los elementos de S en el ejercicio 2.8 de la página 38 tienen la misma probabilidad de ocurrencia y encuentre

a) la probabilidad del evento A;

b) la probabilidad del evento C;

c) la probabilidad del evento $A \cap C$.

2.53 Una caja contiene 500 sobres, de los cuales 75 contienen \$100 en efectivo, 150 contienen \$25 y 275 contienen \$10. Se puede comprar un sobre en \$25. ¿Cuál es el espacio muestral para las diferentes cantidades de dinero? Asigne probabilidades a los puntos muestrales y después encuentre la probabilidad de que el primer sobre que se compre contenga menos de \$100.

2.54 Suponga que en un grupo de último año de facultad de 500 estudiantes se encuentra que 210 fuman, 258 consumen bebidas alcohólicas, 216 comen entre comidas, 122 fuman y consumen bebidas alcohólicas, 83 comen entre comidas y consumen bebidas alcohólicas, 97 fuman y comen entre comidas, y 52 tienen esos tres hábitos nocivos para la salud. Si se selecciona al azar a un miembro de este grupo, encuentre la probabilidad de que el estudiante

a) fume pero no consuma bebidas alcohólicas;

b) coma entre comidas y consuma bebidas alcohólicas pero no fume;

c) ni fume ni coma entre comidas.

2.55 La probabilidad de que una industria estadounidense se ubique en Shanghai, China, es 0.7, la

probabilidad de que se ubique en Beijin, China, es 0.4 y la probabilidad de que se ubique en Shanghai o Beijin o en ambas es 0.8. ¿Cuál es la probabilidad de que la industria se ubique

a) en ambas ciudades?

b) en ninguna de esas ciudades?

2.56 De experiencias pasadas un agente bursátil considera que con las condiciones económicas actuales un cliente invertirá en bonos libres de impuestos con una probabilidad de 0.6, que invertirá en fondos mutualistas con una probabilidad de 0.3 y que invertirá en ambos con una probabilidad de 0.15. Ahora, encuentre la probabilidad de que un cliente invierta

a) en bonos libres de impuestos o en fondos mutualistas;

b) en ninguno de esos instrumentos.

2.57 Si se elige al azar una letra del alfabeto inglés, encuentre la probabilidad de que la letra

a) sea una vocal excepto *y*;

b) esté listada en algún lugar antes de la letra *j*;

c) esté listada en algún lugar después de la letra *g*.

2.58 Un fabricante de automóviles está preocupado por el posible retiro de su sedán de cuatro puertas con mayor venta. Si hubiera un retiro, existe una probabilidad de 0.25 de que haya un defecto en el sistema de frenos, de 0.18 en la transmisión, de 0.17 en el sistema de combustible y de 0.40 en alguna otra área.

a) ¿Cuál es la probabilidad de que el defecto esté en los frenos o en el sistema de combustible, si la probabilidad de defectos simultáneos en ambos sistemas es 0.15?

b) ¿Cuál es la probabilidad de que no haya defecto en los frenos o en el sistema de combustible?

2.59 Si cada artículo codificado en un catálogo empieza con 3 letras distintas seguidas por 4 dígitos distintos de cero, encuentre la probabilidad de seleccionar aleatoriamente uno de estos artículos codificados que tenga como primera letra una vocal y el último dígito sea par.

2.60 Se lanza un par de dados. Encuentre la probabilidad de obtener

a) un total de 8;

b) a lo más un total de 5.

2.61 Se sacan dos cartas sucesivamente de una baraja sin remplazo. ¿Cuál es la probabilidad de que ambas cartas sean mayores que 2 y menores que 8?

2.62 Si se toman 3 libros al azar de un librero que contiene 5 novelas, 3 libros de poemas y 1 diccionario, ¿cuál es la probabilidad de que

a) se seleccione el diccionario?

b) se seleccionen 2 novelas y 1 libro de poemas?

2.63 En una mano de póquer que consiste en 5 cartas, encuentre la probabilidad de tener

a) 3 ases;

b) 4 cartas de corazones y 1 de tréboles.

2.64 En un juego de Yahtzee, donde se lanzan 5 dados de forma simultánea, encuentre la probabilidad de obtener 4 del mismo tipo.

2.65 En una clase de 100 estudiantes graduados de preparatoria, 54 estudiaron matemáticas; 69, historia, y 35 cursaron matemáticas e historia. Si se selecciona al azar uno de estos estudiantes, encuentre la probabilidad de que

a) el estudiante haya cursado matemáticas o historia;

b) el estudiante no haya llevado ninguna de estas materias;

c) el estudiante haya cursado historia pero no matemáticas.

2.66 La empresa Dom's Pizza utiliza pruebas de sabor y el análisis estadístico de los datos antes de comercializar cualquier producto nuevo. Considere un estudio que incluye tres tipos de pastas (delgada, delgada con ajo y orégano, y delgada con trozos de queso). Dom's también estudia tres salsas (estándar, una nueva salsa con más ajo y una nueva salsa con albahaca fresca).

a) ¿Cuántas combinaciones de pasta y salsa se incluyen?

b) ¿Cuál es la probabilidad de que un juez tenga una pasta delgada sencilla con salsa estándar en su primera prueba de sabor?

2.67 De acuerdo con *Consumer Digest* (julio/agosto de 1996), la ubicación probable de las PC en una casa son:

Dormitorio de adultos:	0.03
Dormitorio de niños:	0.15
Otro dormitorio:	0.14
Oficina o estudio:	0.40
Otra habitación	0.28

a) ¿Cuál es la probabilidad de que una PC esté en un dormitorio?

b) ¿Cuál es la probabilidad de que no esté en un dormitorio?

c) Suponga que se selecciona una familia al azar entre las familias con una PC; ¿en qué habitación esperaría encontrar la PC?

2.68 El interés se enfoca en la vida de un componente electrónico. Suponga que se sabe que la probabilidad de que el componente funcione más de 6000 horas es 0.42. Suponga, además, que la probabilidad de que el componente *no dure más de* 4000 horas es 0.04.

a) ¿Cuál es la probabilidad de que la vida del componente sea menor o igual a 6000 horas?

b) ¿Cuál es la probabilidad de que la vida sea mayor que 4000 horas?

2.69 Considere la situación del ejercicio 2.68. Sea A el evento de que el componente falle en una prueba específica y B el evento de que el componente se deforme pero en realidad no falle. El evento A ocurre con una probabilidad de 0.20 y el evento B ocurre con una probabilidad de 0.35.

a) ¿Cuál es la probabilidad de que el componente no falle en la prueba?

b) ¿Cuál es la probabilidad de que el componente funcione perfectamente bien (es decir, que ni se deforme ni que falle en la prueba)?

c) ¿Cuál es la probabilidad de que el componente falle o se deforme en la prueba?

2.70 En las fábricas a los trabajadores constantemente se les motiva para que practiquen la tolerancia cero para prevenir los accidentes en el lugar de trabajo. Los accidentes pueden ocurrir porque el ambiente o las condiciones laborales son inseguros en sí mismos. Por otro lado, los accidentes pueden ocurrir por negligencia o simplemente por fallas humanas. Además, los horarios de trabajo de 7:00 A.M. a 3:00 P.M. (turno matutino), de 3:00 P.M. a 11:00 P.M. (turno vespertino) y de 11:00 P.M. a 7:00 A.M. (turno nocturno) pueden ser un factor. El año pasado ocurrieron 300 accidentes. Los porcentajes de los accidentes por la combinación de condiciones son como sigue:

Turno	Condiciones inseguras	Fallas humanas
Matutino	5%	32%
Vespertino	6%	25%
Nocturno	2%	30%

Si se elige aleatoriamente un reporte de accidente de entre los 300 reportes,

a) ¿Cuál es la probabilidad de que el accidente haya ocurrido en el turno nocturno?

b) ¿Cuál es la probabilidad de que el accidente haya ocurrido debido a una falla humana?

c) ¿Cuál es la probabilidad de que el accidente haya ocurrido debido a las condiciones inseguras?

d) ¿Cuál es la probabilidad de que el accidente haya ocurrido durante los turnos vespertino o nocturno?

2.71 Considere la situación del ejemplo 2.31 de la página 54.

a) ¿Cuál es la probabilidad de que no más de 4 automóviles recibirán servicio del mecánico?

b) ¿Cuál es la probabilidad de que el mecánico dará servicio a menos de 8 automóviles?

c) ¿Cuál es la probabilidad de que el mecánico dará servicio a 3 o 4 automóviles?

2.72 El interés se enfoca en la naturaleza de un horno que se compra en una tienda por departamentos específica. Puede ser de gas o eléctrico. Considere la decisión tomada por seis clientes distintos.

a) Suponga que la probabilidad de que, a lo más, dos de esos individuos compren un horno eléctrico es 0.40. ¿Cuál será la probabilidad de que al menos tres compren un horno eléctrico?

b) Suponga que se sabe que la probabilidad de que los seis compren el horno eléctrico es 0.007, mientras que 0.104 es la probabilidad de que los seis compren el horno de gas. ¿Cuál es la probabilidad de que al menos se compre un horno de cada tipo?

2.73 En muchas industrias es común que se utilicen máquinas para llenar los envases de un producto. Esto ocurre tanto en la industria alimentaria como en otras áreas cuyos productos son de uso doméstico, como los detergentes. Dichas máquinas no son perfectas y, de hecho, podrían A cumplir las especificaciones de llenado, B quedar por debajo del llenado establecido y C llenar de más. Por lo general, se busca evitar la práctica de llenado insuficiente. Sea $P(B) = 0.001$, mientras que $P(A) = 0.990$.

a) Determine $P(C)$.

b) ¿Cuál es la probabilidad de que la máquina no dé llenado insuficiente?

c) ¿Cuál es la probabilidad de que la máquina llene de más o de menos?

2.74 Considere la situación del ejercicio 2.73. Suponga que se producen 50,000 bolsas de detergente por semana y también que las bolsas con llenado insuficiente se "devuelven" con la petición de rembolsar al cliente el precio de compra. Suponga que se sabe que el "costo" de producción es de $4.00 por bolsa, en tanto que el precio de compra es de $4.50 por bolsa.

a) ¿Cuál es la utilidad semanal cuando no se tienen bolsas defectuosas?

b) ¿Cuál es la perdida en utilidades esperada debido al llenado insuficiente?

2.75 Como sugeriría la situación del ejercicio 2.73, a menudo los procedimientos estadísticos se utilizan para control de calidad (es decir, control de calidad industrial). A veces, el *peso* de un producto es una variable importante que hay que controlar. Se dan especificaciones de peso para ciertos productos empacados, y si un paquete está muy ligero o muy pesado se rechaza. Los datos históricos sugieren que 0.95 es la probabilidad de que el producto cumpla con las especificaciones de peso; mientras que 0.002 es la probabilidad de que el producto esté muy ligero. Por cada uno de los productos empacados el fabricante invierte $20.00 en producción y el precio de compra para el consumidor son $25.00.

a) ¿Cuál es la probabilidad de que un paquete elegido aleatoriamente de la línea de producción esté muy pesado?

b) Por cada 10,000 paquetes que se venden, que utilidad recibirá el fabricante si todos los paquetes cumplen con las especificaciones de peso?

c) Considerando que los paquetes "defectuosos" se rechazan y pierden todo su valor, ¿en cuánto se reduce la utilidad por cada 10,000 paquetes debido a que no cumplen con las especificaciones?

2.76 Demuestre que

$$P(A' \cap B') = 1 + P(A \cap B) - P(A) - P(B).$$

2.6 Probabilidad condicional

La probabilidad de que un evento B ocurra cuando se sabe que ya ocurrió algún evento A se llama **probabilidad condicional** y se denota con $P(B|A)$. El símbolo $P(B|A)$, por lo general, se lee "la probabilidad de que ocurra B dado que ocurrió A" o simplemente "la probabilidad de B, dado A".

Considere el evento B de obtener un cuadrado perfecto cuando se lanza un dado. El dado se construye de modo que los números pares tengan el doble de probabilidad de ocurrencia que los números nones. Con base en el espacio muestral $S = \{1, 2, 3, 4, 5, 6\}$, con probabilidades asignadas de 1/9 y 2/9, respectivamente, a los números impares y a los pares, la probabilidad de que ocurra B es 1/3. Suponga ahora que se sabe que el lanzamiento del dado tiene como resultado un número mayor que 3. Tenemos ahora un espacio muestral reducido $A = \{4, 5, 6\}$, que es un subconjunto de S. Para encontrar la probabilidad de que ocurra B, en relación con el espacio A, debemos asignar primero nuevas probabilidades a los elementos de A proporcionales a sus probabilidades originales de modo que su suma sea 1. Al asignar una probabilidad de w al número non en A y una probabilidad de $2w$ a los dos números pares, tenemos $5w = 1$ o $w = 1/5$. En relación con el espacio A, encontramos que B contiene sólo el elemento 4. Si denotamos este evento con el símbolo $B|A$, escribimos $B|A = (4)$, y de aquí

$$P(B|A) = \frac{2}{5}.$$

Este ejemplo ilustra que los eventos pueden tener probabilidades diferentes cuando se consideran en relación con diferentes espacios muestrales.

También podemos escribir

$$P(B|A) = \frac{2}{5} = \frac{2/9}{5/9} = \frac{P(A \cap B)}{P(A)},$$

donde $P(A \cap B)$ y $P(A)$ se encuentran a partir del espacio muestral original S. En otras palabras, una probabilidad condicional relativa a un subespacio A de S se puede calcular de forma directa de las probabilidades que se asignan a los elementos del espacio muestral original S.

Definición 2.9: La probabilidad condicional de B, dado A, que se denota con $P(B|A)$, se define como

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{si} \quad P(A) > 0.$$

Como ilustración adicional, suponga que nuestro espacio muestral S es la población de adultos en una pequeña ciudad que cumplen con los requisitos para obtener

un título universitario. Debemos clasificarlos de acuerdo con su sexo y situación laboral:

Tabla 2.1: Clasificación de los adultos en una ciudad pequeña

	Empleado	**Desempleado**	**Total**
Hombre	460	40	500
Mujer	140	260	400
Total	600	300	900

Uno de estos individuos se seleccionará al azar para que realice un viaje a través del país para promover las ventajas de establecer industrias nuevas en la ciudad. Nos interesaremos en los eventos siguientes:

M: se elige a un hombre,

E: el elegido tiene empleo.

Al utilizar el espacio muestral reducido E, encontramos que

$$P(M|E) = \frac{460}{600} = \frac{23}{30}.$$

Sea $n(A)$ el número de elementos en cualquier conjunto A. Con el uso de esta notación, podemos escribir

$$P(M|E) = \frac{n(E \cap M)}{n(E)} = \frac{n(E \cap M)/n(S)}{n(E)/n(S)} = \frac{P(E \cap M)}{P(E)},$$

donde $P(E \cap M)$ y $P(E)$ se encuentran a partir del espacio muestral original S. Para verificar este resultado, note que

$$P(E) = \frac{600}{900} = \frac{2}{3} \qquad \text{y} \qquad P(E \cap M) = \frac{460}{900} = \frac{23}{45}.$$

Por lo tanto,

$$P(M|E) = \frac{23/45}{2/3} = \frac{23}{30},$$

como antes.

Ejemplo 2.33: La probabilidad de que un vuelo programado normalmente salga a tiempo es $P(D) = 0.83$; la probabilidad de que llegue a tiempo es $P(A) = 0.82$; y la probabilidad de que salga y llegue a tiempo es $P(D \cap A) = 0.78$. Encuentre la probabilidad de que un avión *a*) llegue a tiempo, dado que salió a tiempo; y *b*) salió a tiempo, dado que llegó a tiempo.

Solución: *a*) La probabilidad de que un avión llegue a tiempo, dado que salió a tiempo es

$$P(A|D) = \frac{P(D \cap A)}{P(D)} = \frac{0.78}{0.83} = 0.94.$$

b) La probabilidad de que un avión haya salido a tiempo, dado que llegó a tiempo es

$$P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{0.78}{0.82} = 0.95.$$

En el experimento del lanzamiento de un dado que se discutió en la página 58 notamos que $P(B|A) = 2/5$; mientras que $P(B) = 1/3$. Es decir, $P(B|A) \neq P(B)$, lo cual indica que B depende de A. Consideremos ahora un experimento donde se sacan 2 cartas, una después de la otra, de una baraja ordinaria, con reemplazo. Los eventos se definen como

A: la primera carta es un as,

B: la segunda carta es una espada.

Como la primera carta se reemplaza, nuestro espacio muestral para la primera y segunda cartas consiste en 52 cartas, que contienen 4 ases y 13 espadas. Entonces,

$$P(B|A) = \frac{13}{52} = \frac{1}{4} \quad \text{y} \quad P(B) = \frac{13}{52} = \frac{1}{4}.$$

Es decir, $P(B|A) = P(B)$. Cuando esto es cierto, se dice que los eventos A y B son **independientes**.

La noción de probabilidad condicional brinda la capacidad de reevaluar la idea de probabilidad de un evento a la luz de la información adicional; es decir, cuando se sabe que ocurrió otro evento. La probabilidad $P(A|B)$ es una "actualización" de $P(A)$ basada en el conocimiento de que ocurrió el evento B. En el ejemplo 2.33 es importante conocer la probabilidad de que el vuelo llegue a tiempo. Se nos da la información de que el vuelo no salió a tiempo. Con esta información adicional, la probabilidad más pertinente es $P(A|D')$, esto es, la probabilidad de que llegue a tiempo, dado que no salió a tiempo. En muchas situaciones las conclusiones que se obtienen de observar la probabilidad condicional más importante cambian drásticamente la situación. En este ejemplo el cálculo de $P(A|D')$ es

$$P(A|D') = \frac{P(A \cap D')}{P(D')} = \frac{0.82 - 0.78}{0.17} = 0.24.$$

Entonces, la probabilidad de una llegada a tiempo disminuye significativamente ante la presencia de la información adicional.

Ejemplo 2.34: El concepto de probabilidad condicional tiene innumerables aplicaciones industriales y biomédicas. Considere un proceso industrial en el ramo textil, donde se producen franjas (tiras) para una clase de ropa específica. Las franjas pueden estar defectuosas de dos maneras: en longitud y en textura. En cuanto a esta última el proceso de identificación es muy complicado. A partir de información histórica del proceso se sabe que 10% de las franjas no pasan la prueba de longitud, que 5% no pasan la prueba de textura y que sólo 0.8% no pasan ambas pruebas. Si en el proceso se elige aleatoriamente una franja y una medición rápida identifica que no pasa la prueba de longitud, ¿cuál es la probabilidad de que esté defectuosa en textura?

Solución: Considere los eventos

L : defecto en longitud, T : defecto en textura

Así, como la franja está defectuosa en longitud, la probabilidad de que esta franja esté defectuosa en textura está dada por

$$P(T|L) = \frac{P(T \cap L)}{P(L)} = \frac{0.008}{0.1} = 0.08.$$

Entonces, el conocimiento que da la probabilidad condicional aporta información considerablemente mayor que tan sólo saber $P(T)$.

Eventos independientes

Aunque la probabilidad condicional tiene en cuenta la alteración de la probabilidad de un evento a la luz de material adicional, también nos permite entender mejor el muy importante concepto de **independencia** o, en el contexto actual, de eventos independientes. En el ejemplo 2.33, del aeropuerto, $P(A|D)$ difiere de $P(A)$. Esto sugiere que la ocurrencia de D influye en A y esto realmente se espera en este caso. Sin embargo, considere la situación donde tenemos los eventos A y B y

$$P(A|B) = P(A).$$

En otras palabras, la ocurrencia de B no influye en las probabilidades de ocurrencia de A. Aquí la ocurrencia de A es independiente de la ocurrencia de B. La importancia del concepto de independencia no se debe enfatizar en exceso. Juega un papel vital en el material de casi todos los capítulos de este libro y en todas las áreas de la estadística aplicada.

Definición 2.10: Dos eventos A y B son **independientes** si y sólo si

$$P(B|A) = P(B) \quad \text{o} \quad P(A|B) = P(A),$$

dada la existencia de probabilidad condicional. De otra forma, A y B son **dependientes**.

La condición $P(B|A) = P(B)$ implica que $P(A|B) = P(A)$, y viceversa. Para los experimentos de extracción de una carta, donde mostramos que $P(B|A) = P(B) = 1/4$, también podemos ver que $P(A|B) = P(A) = 1/13$.

2.7 Reglas multiplicativas

Al multiplicar la fórmula de la definición 2.9 por $P(A)$, obtenemos la siguiente **regla multiplicativa** importante, que nos permite calcular la probabilidad de que ocurran dos eventos.

Teorema 2.13: Si en un experimento pueden ocurrir los eventos A y B, entonces

$$P(A \cap B) = P(A)P(B|\mathrm{A}), \quad \text{dado que } P(A) > 0.$$

Así la probabilidad de que ocurran A y B es igual a la probabilidad de que ocurra A multiplicada por la probabilidad condicional de que ocurra B, dado que ocurre A.

Como los eventos $A \cap B$ y $B \cap A$ son equivalentes, del teorema 2.13 se sigue que también podemos escribir

$$P(A \cap B) = P(B \cap A) = P(B)P(A|B).$$

En otras palabras, no importa qué evento se considere como A y cuál como B.

Ejemplo 2.35: Suponga que tenemos una caja de fusibles que contiene 20 unidades, de las cuales 5 están defectuosas. Si se seleccionan 2 fusibles al azar y se retiran de la caja, uno después del otro, sin reemplazar el primero, ¿cuál es la probabilidad de que ambos fusibles estén defectuosos?

Solución: Sean A el evento de que el primer fusible esté defectuoso y B el evento de que el segundo esté defectuoso; entonces, interpretamos $A \cap B$ como el evento de que ocurra A, y entonces B ocurre después de que haya ocurrido A. La probabilidad de separar primero un fusible defectuoso es $1/4$; entonces, la probabilidad de separar un segundo fusible defectuoso de los restantes 4 es $4/19$. Por lo tanto,

$$P(A \cap B) = \left(\frac{1}{4}\right)\left(\frac{4}{19}\right) = \frac{1}{19}.$$

Ejemplo 2.36: Una bolsa contiene 4 bolas blancas y 3 negras, y una segunda bolsa contiene 3 blancas y 5 negras. Se saca una bola de la primera bolsa y se coloca sin verla en la segunda bolsa. ¿Cuál es la probabilidad de que ahora se saque una bola negra de la segunda bolsa?

Solución: Sean B_1, B_2 y W_1, respectivamente, la extracción de una bola negra de la bolsa 1, una bola negra de la bolsa 2 y una bola blanca de la bolsa 1. Nos interesa la unión de los eventos mutuamente excluyentes $B_1 \cap B_2$ y $W_1 \cap B_2$. Las diversas posibilidades y sus probabilidades se ilustran en la figura 2.8. Entonces

$$\begin{aligned} P[(B_1 \cap B_2) \text{ o } (W_1 \cap B_2)] &= P(B_1 \cap B_2) + P(W_1 \cap B_2) \\ &= P(B_1)P(B_2|B_1) + P(W_1)P(B_2|W_1) \\ &= \left(\frac{3}{7}\right)\left(\frac{6}{9}\right) + \left(\frac{4}{7}\right)\left(\frac{5}{9}\right) = \frac{38}{63}. \end{aligned}$$

Si, en el ejemplo 2.35, el primer fusible se reemplaza y los fusibles se reacomodan por completo antes de que se extraiga el segundo, entonces la probabilidad de un fusible defectuoso en la segunda selección aún es $1/4$; es decir, $P(B \mid A) = P(B)$, y los eventos A y B son independientes. Cuando esto es cierto, podemos sustituir $P(B \mid A)$ por $P(B)$ en el teorema 2.13 para obtener la siguiente regla multiplicativa especial.

Teorema 2.14: Dos eventos A y B son independientes si y sólo si

$$P(A \cap B) = P(A)P(B).$$

Por lo tanto, para obtener la probabilidad de que ocurran dos eventos independientes, simplemente calculamos el producto de sus probabilidades individuales.

Ejemplo 2.37: Una pequeña ciudad tiene un carro de bomberos y una ambulancia disponibles para emergencias. La probabilidad de que el carro de bomberos esté disponible cuando

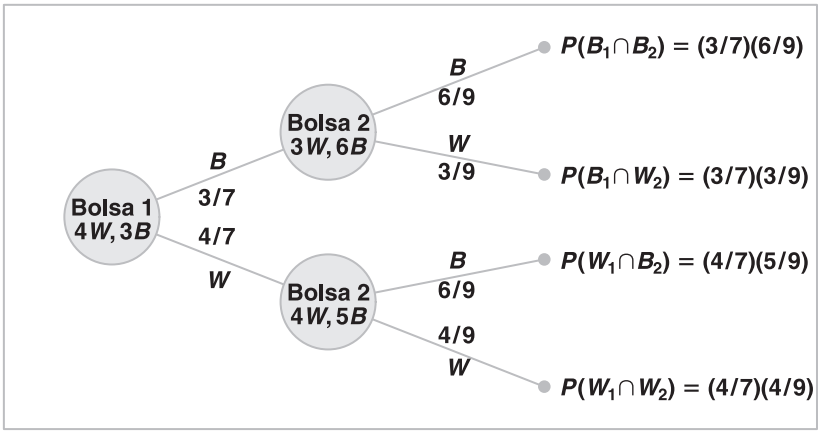


Figura 2.8: Diagrama de árbol para el ejemplo 2.36.

se necesite es 0.98 y la probabilidad de que la ambulancia esté disponible cuando se le requiera es 0.92. En el caso de que resulte un herido de un edificio en llamas, encuentre la probabilidad de que tanto la ambulancia como el carro de bomberos estén disponibles.

Solución: Sean A y B los respectivos eventos de que estén disponibles el carro de bomberos y la ambulancia. Entonces,

$$P(A \cap B) = P(A)P(B) = (0.98)(0.92) = 0.9016.$$

Ejemplo 2.38: Un sistema eléctrico consiste en cuatro componentes como se ilustra en la figura 2.9. El sistema funciona si los componentes A y B funcionan, y ya sea que funcionen los componentes C o D. La confiabilidad (probabilidad de que funcionen) de cada uno de los componentes también se muestra en la figura 2.9. Encuentre la probabilidad de que *a*) el sistema completo funcione, y *b*) que el componente C no funcione, dado que el sistema completo funciona. Suponga que los cuatro componentes funcionan de manera independiente.

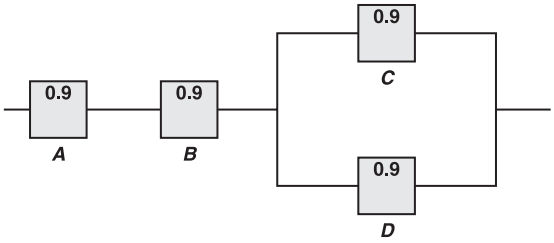


Figura 2.9: Un sistema eléctrico para el ejemplo 2.38.

Solución: En esta configuración del sistema, A, B, y el subsistema C y D constituyen un sistema de circuitos en serie; mientras que el mismo subsistema C y D es un sistema de circuitos en paralelo.

a) En efecto, la probabilidad de que el sistema completo funcione se calcula de la siguiente manera:

$$\begin{aligned} P(A \cap B \cap (C \cup D)) &= P(A)P(B)P(C \cup D) \\ &= P(A)P(B)[1 - P(C' \cap D')] \\ &= P(A)P(B)[1 - P(C')P(D')] \\ &= (0.9)(0.9)[1 - (1 - 0.8)(1 - 0.8)] \\ &= 0.7776. \end{aligned}$$

Las igualdades anteriores son válidas por la independencia entre los cuatro componentes.

b) En este caso para calcular la probabilidad condicional, note que

$$\begin{aligned} P &= \frac{P(\text{El sistema funciona pero } C \text{ no funciona})}{P(\text{el sistema funciona})} \\ &= \frac{P(A \cap B \cap C' \cap D)}{P(\text{el sistema funciona})} = \frac{(0.9)(0.9)(1 - 0.8)(0.8)}{0.7776} = 0.1667. \end{aligned}$$

Teorema 2.15: Si, en un experimento, pueden ocurrir los eventos A_1, $A_2, \ldots, A_k$, entonces

$$\begin{aligned} & P(A_1 \cap A_2 \cap \cdots \cap A_k) \\ = \; & P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k|A_1 \cap A_2 \cap \cdots \cap A_{k-1}). \end{aligned}$$

Si los eventos A_1, $A_2, \ldots,$ A_k son independientes, entonces

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1)P(A_2) \cdots P(A_k).$$

Ejemplo 2.39: Se sacan tres cartas una tras otra, sin reemplazo, de una baraja ordinaria. Encuentre la probabilidad de que ocurra el evento $A_1 \cap A_2 \cap A_3$, donde A_1 es el evento de que la primera carta sea un as rojo, A_2 el evento de que la segunda carta sea un 10 o un *jack*, y A_3 el evento de que la tercera carta sea mayor que 3 pero menor que 7.

Solución: Primero definimos los eventos:

A_1: la primera carta es un as rojo,

A_2: la segunda carta es un 10 o un jack,

A_3: la tercera carta es mayor que 3 pero menor que 7.

Entonces,

$$P(A_1) = \frac{2}{52}, \quad P(A_2|A_1) = \frac{8}{51}, \quad P(A_3|A_1 \cap A_2) = \frac{12}{50},$$

y de aquí, por el teorema 2.15,

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A2) \\ &= \left(\frac{2}{52}\right)\left(\frac{8}{51}\right)\left(\frac{12}{50}\right) = \frac{8}{5525}. \end{aligned}$$

Ejemplo 2.40: Se carga una moneda de manera que la cara tenga una probabilidad de ocurrir dos veces mayor que la cruz. Si se lanza tres veces la moneda, ¿cuál es la probabilidad de obtener dos cruces y una cara?

Solución: El espacio muestral para el experimento consiste en los 8 elementos,

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

Sin embargo, con una moneda cargada ya no es posible asignar probabilidades iguales a cada punto muestral. Es fácil observar que $P(H) = 2/3$ y $P(T) = 1/3$ para un lanzamiento, ya que una cara tiene una probabilidad de ocurrir mayor que una cruz. Sea ahora A el evento de obtener dos cruces y una cara en los tres lanzamientos de la moneda. Entonces,

$$A = \{TTH, THT, HTT\},$$

y como los resultados en cada uno de los 3 lanzamientos son independientes, del teorema 2.15 se sigue que

$$P(TTH) = P(T)P(T)P(H) = \left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right) = \frac{2}{27}.$$

De manera similar, $P(THT) = P(HTT) = 2/27$ y, por ello, $P(A) = 2/27 + 2/27 + 2/27 = 2/9$.

Ejercicios

2.77 Si R es el evento de que un convicto cometiera un robo a mano armada y D es el evento de que el convicto promoviera el consumo de drogas, exprese en palabras lo que en probabilidades se indica como

a) $P(R|D)$;
b) $P(D'|R)$;
c) $P(R'|D')$.

2.78 Una clase de física avanzada se compone de 10 estudiantes de primer año, 30 del último año y 10 graduados. Las calificaciones finales muestran que 3 estudiantes de primer año, 10 del último año y 5 de los graduados obtuvieron A en el curso. Si se elige un estudiante al azar de esta clase y se encuentra que es uno de los que obtuvieron A, ¿cuál es la probabilidad de que él o ella sea un estudiante de último año?

2.79 Una muestra aleatoria de 200 adultos se clasifica a continuación por sexo y nivel de educación.

Educación	**Hombre**	**Mujer**
Primaria	38	45
Secundaria	28	50
Universidad	22	17

Si se elige una persona al azar de este grupo, encuentre la probabilidad de que

a) la persona sea hombre, dado que la persona tiene educación secundaria;
b) la persona no tiene un grado universitario, dado que la persona es mujer.

2.80 En un experimento para estudiar la relación de la hipertensión arterial con los hábitos de fumar, se reúnen los siguientes datos para 180 individuos:

	No fumadores	**Fumadores moderados**	**Fumadores empedernidos**
H	21	36	30
NH	48	26	19

donde H y NH en la tabla representan *Hipertensión* y *Sin hipertensión*, respectivamente. Si se selecciona uno de estos individuos al azar, encuentre la probabilidad de que la persona

a) sufra hipertensión, dado que la persona es un fumador empedernido;
b) sea un no fumador, dado que la persona no sufre de hipertensión.

2.81 En el último año de una clase de bachillerato con 100 estudiantes, 42 cursaron matemáticas; 68, psicología; 54, historia; 22, matemáticas e historia; 25, matemáticas y psicología, 7 historia pero ni matemáticas ni psicología; 10, las tres materias; y 8 no tomaron ninguna de las tres. Si se selecciona un estudiante al azar, encuentre la probabilidad de que

a) una persona inscrita en psicología curse las tres materias;

b) una persona que no se inscribió en psicología curse historia y matemáticas.

2.82 Un fabricante de una vacuna para la gripe se interesa en la calidad de su suero. Tres departamentos diferentes procesan los lotes de suero y tienen tasas de rechazo de 0.10, 0.08 y 0.12, respectivamente. Las inspecciones de los tres departamentos son secuenciales e independientes.

a) ¿Cuál es la probabilidad de que un lote de suero sobreviva a la primera inspección departamental, pero sea rechazado por el segundo departamento?

b) ¿Cuál es la probabilidad de que un lote de suero sea rechazado por el tercer departamento?

2.83 En *USA Today* (5 de septiembre de 1996) se listaron como sigue los resultados de una encuesta sobre el uso de ropa para dormir mientras se viaja:

	Hombre	**Mujer**	**Total**
Ropa interior	0.220	0.024	0.244
Camisón	0.002	0.180	0.182
Nada	0.160	0.018	0.178
Pijama	0.102	0.073	0.175
Camiseta	0.046	0.088	0.134
Otros	0.084	0.003	0.087

a) ¿Cuál es la probabilidad de que un viajero sea una mujer que duerme desnuda?

b) ¿Cuál es la probabilidad de que un viajero sea hombre?

c) Suponiendo que el viajero sea hombre, ¿cuál es la probabilidad de que duerma en pijama?

d) ¿Cuál es la probabilidad de que un viajero sea hombre si duerme en pijama o en camiseta?

2.84 La probabilidad de que un automóvil al que se llena el tanque de gasolina también necesite un cambio de aceite es 0.25, la probabilidad de que necesite un nuevo filtro de aceite es 0.40, y la probabilidad de que necesite cambio de aceite y filtro es 0.14.

a) Si se tiene que cambiar el aceite, ¿cuál es la probabilidad de que se necesite un nuevo filtro?

b) Si necesita un nuevo filtro de aceite, ¿cuál es la probabilidad de que se tenga que cambiar el aceite?

2.85 La probabilidad de que un hombre casado vea cierto programa de televisión es 0.4 y la probabilidad de que una mujer casada vea el programa es 0.5. La probabilidad de que un hombre vea el programa, dado que su esposa lo hace, es 0.7. Encuentre la probabilidad de que

a) un matrimonio vea el programa;

b) una esposa vea el programa dado que su esposo lo ve;

c) al menos 1 persona de un matrimonio vea el programa.

2.86 Para matrimonios que viven en cierto suburbio, la probabilidad de que el esposo vote en un referéndum es 0.21, la probabilidad de que su esposa vote es 0.28 y la probabilidad de que ambos voten es 0.15. ¿Cuál es la probabilidad de que

a) al menos un miembro de un matrimonio vote?

b) una esposa vote, dado que su esposo votará?

c) un esposo vote, dado que su esposa no vota?

2.87 La probabilidad de que un vehículo que entra a las Cavernas Luray tenga matrícula de Canadá es 0.12, la probabilidad de que sea una casa rodante es 0.28, y la probabilidad de que sea una casa rodante con matrícula de Canadá es 0.09. ¿Cuál es la probabilidad de que

a) una casa rodante que entra a las Cavernas Luray tenga matrícula de Canadá?

b) un vehículo con matrícula de Canadá que entra a las Cavernas Luray sea una casa rodante?

c) un vehículo que entra a las Cavernas Luray no tenga matrícula de Canadá o que no sea una casa rodante?

2.88 La probabilidad de que el jefe de familia esté en casa cuando llame un representante de marketing es 0.4. Dado que el jefe de familia está en casa, la probabilidad de que se compren bienes de la compañía es 0.3. Encuentre la probabilidad de que el jefe de familia esté en casa y se compren bienes de la compañía.

2.89 La probabilidad de que un doctor diagnostique de manera correcta una enfermedad específica es 0.7. Dado que el doctor hace un diagnóstico incorrecto, la probabilidad de que el paciente entable una demanda legal es 0.9. ¿Cuál es la probabilidad de que el doctor haga un diagnóstico incorrecto y el paciente lo demande?

2.90 En 1970, 11% de los estadounidenses completaron cuatro años de universidad, de los cuales 43% eran mujeres. En 1990, 22% de los estadounidenses completaron cuatro años de universidad, de los cuales 53% fueron mujeres. (*Time*, 19 de enero de 1996.)

a) Dado que una persona completó cuatro años de universidad en 1970, ¿cuál es la probabilidad de que la persona sea mujer?

b) ¿Cuál es la probabilidad de que una mujer terminara cuatro años de universidad en 1990?

c) ¿Cuál es la probabilidad de que en 1990 un hombre no haya terminado la universidad?

2.91 Un agente de bienes raíces tiene ocho llaves maestras para abrir varias casas nuevas. Sólo 1 llave maestra abrirá cualesquiera de las casas. Si 40% de estas casas por lo general se dejan abiertas, ¿cuál es la probabilidad de que el agente de bienes raíces pueda entrar en una casa específica, si selecciona 3 llaves maestras al azar antes de salir de la oficina?

2.92 Antes de la distribución de cierto software estadístico se prueba la precisión de cada cuarto disco compacto (CD). El proceso de prueba consiste en correr cuatro programas independientes y verificar los resultados. La tasa de falla para los 4 programas de prueba son 0.01, 0.03, 0.02 y 0.01, respectivamente.

a) ¿Cuál es la probabilidad de que un CD que se pruebe falle cualquier prueba?

b) Dado que se prueba un CD, ¿cuál es la probabilidad de que falle el programa 2 o 3?

c) En una muestra de 100, ¿cuántos CD esperaría que se rechazaran?

d) Dado que un CD está defectuoso, ¿cuál es la probabilidad de que se pruebe?

2.93 Una ciudad tiene dos carros de bomberos que operan de forma independiente. La probabilidad de que un carro específico esté disponible cuando se le necesite es 0.96.

a) ¿Cuál es la probabilidad de que ninguno esté disponible cuando se les necesite?

b) ¿Cuál es la probabilidad de que un carro de bomberos esté disponible cuando se le necesite?

2.94 La probabilidad de que Tom viva 20 años más es 0.7, y la probabilidad de que Nancy viva 20 años más es 0.9. Si suponemos independencia para ambos, ¿cuál es la probabilidad de que ninguno viva 20 años más?

2.95 Un neceser contiene 2 frascos de aspirina y 3 frascos de comprimidos para la tiroides. Un segundo bolso grande contiene 3 frascos de aspirinas, 2 frascos de comprimidos para la tiroides y 1 frasco de pastillas laxantes. Si se saca 1 frasco al azar de cada equipaje, encuentre la probabilidad de que

a) ambos frascos contengan comprimidos para la tiroides;

b) ningún frasco contenga comprimidos para la tiroides;

c) los 2 frascos contengan cosas diferentes.

2.96 La probabilidad de que una persona que visita a su dentista necesite rayos X es 0.6, la probabilidad de que una persona que necesite una placa de rayos X también tenga una amalgama es 0.3, y la probabilidad de que una persona que tenga una placa de rayos X y una amalgama también tenga una extracción dental es 0.1. ¿Cuál es la probabilidad de que una persona que visita a su dentista tenga una placa de rayos X, una amalgama y una extracción dental?

2.97 Encuentre la posibilidad de seleccionar aleatoriamente 4 litros de leche en buenas condiciones sucesivamente de un refrigerador que contiene 20 litros, de los cuales 5 están echados a perder, utilizando

a) La primera fórmula del teorema 2.15 de la página 64.

b) Las fórmulas de los teoremas 2.8 y 2.9 de las páginas 46 y 50, respectivamente.

2.98 Suponga que el diagrama de un sistema eléctrico se muestra en la figura 2.10. ¿Cuál es la probabilidad de que el sistema funcione? Suponga que los componentes fallan de forma independiente.

2.99 Un sistema de circuitos se muestra en la figura 2.11. Suponga que los componentes fallan de manera independiente.

a) ¿Cuál es la probabilidad de que el sistema completo funcione?

b) Dado que el sistema funciona, ¿cuál es la probabilidad de que el componente A no funcione?

2.100 En la situación del ejercicio 2.99, se sabe que el sistema no funciona. ¿Cuál es la probabilidad de que el componente A tampoco funcione?

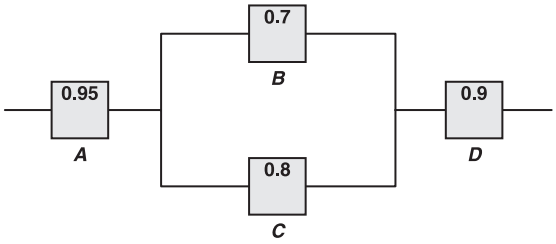


Figura 2.10: Diagrama para el ejercicio 2.98.

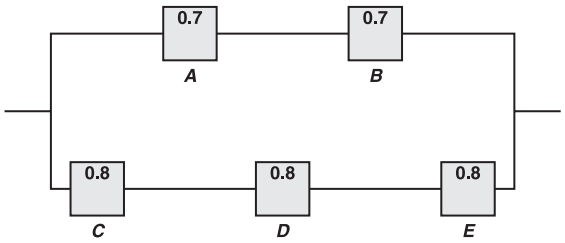


Figura 2.11: Diagrama para el ejercicio 2.99.

2.8 Regla de Bayes

Regresemos a la ilustración de la sección 2.6, donde un individuo se selecciona al azar de entre los adultos de una pequeña ciudad, para viajar por el país y promover las ventajas de establecer industrias nuevas en la ciudad. Suponga que ahora se nos da la información adicional de que 36 de los empleados y 12 de los desempleados son miembros del Club Rotario. Deseamos encontrar la probabilidad del evento A de que el individuo seleccionado sea miembro del Club Rotario. Con referencia a la figura 2.12, podemos escribir A como la unión de los dos eventos mutuamente excluyentes $E \cap A$ y $E' \cap A$. De aquí $A = (E \cap A) \cup (E' \cap A)$ y por el corolario 2.1 del teorema 2.10, y además el teorema 2.13, podemos escribir

$$\begin{aligned} P(A) &= P[(E \cap A) \cup (E' \cap A)] = P(E \cap A) + P(E' \cap A) \\ &= P(E)P(A|E) + P(E')P(A|E'). \end{aligned}$$

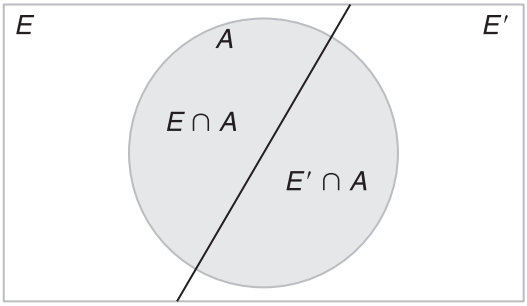


Figura 2.12: Diagrama de Venn para los eventos A, E y E'.

Los datos de la sección 2.6, junto con los datos adicionales dados arriba para el conjunto A, nos permiten calcular

$$P(E) = \frac{600}{900} = \frac{2}{3}, \quad P(A|E) = \frac{36}{600} = \frac{3}{50},$$

y

$$P(E') = \frac{1}{3}, \quad P(A|E') = \frac{12}{300} = \frac{1}{25}.$$

Si mostramos estas probabilidades con el diagrama de árbol de la figura 2.13, donde la primera rama da la probabilidad $P(E)P(A\,|\,E)$ y la segunda rama da la probabilidad $P(E')P(A\,|\,E')$, se sigue que

$$P(A) = \left(\frac{2}{3}\right)\left(\frac{3}{50}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{25}\right) = \frac{4}{75}.$$

Una generalización de la ilustración precedente al caso donde el espacio muestral se parte en k subconjuntos la cubre el siguiente teorema, que algunas veces se denomina **teorema de probabilidad total** o **regla de eliminación**.

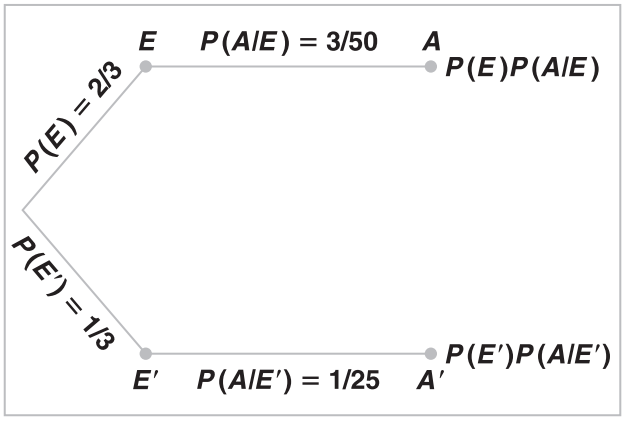


Figura 2.13: Diagrama de árbol para los datos de la página 59 con información adicional de la página 68.

Teorema 2.16: Si los eventos $B_1, B_2, \ldots, B_k$ constituyen una partición del espacio muestral S tal que $P(B_i) \neq 0$ para $i = 1, 2, \ldots, k$, entonces, para cualquier evento A de S,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i).$$

Prueba: Considere el diagrama de Venn de la figura 2.14. Se observa que el evento A es la unión de los eventos mutuamente excluyentes

$$B_1 \cap A,\ B_2 \cap A,\ \ldots,\ B_k \cap A;$$

es decir,

$$A = (B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_k \cap A).$$

Usando el corolario 2.2 del teorema 2.10 y además el teorema 2.13, tenemos

$$\begin{aligned} P(A) &= P[(B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_k \cap A)] \\ &= P(B_1 \cap A) + P(B_2 \cap A) + \cdots + P(B_k \cap A) \\ &= \sum_{i=1}^{k} P(B_i \cap A) \\ &= \sum_{i=1}^{k} P(B_i)P(A|B_i). \end{aligned}$$

Ejemplo 2.41: En cierta planta de ensamble, tres máquinas, B_1, B_2 y B_3, montan 30, 45 y 25% de los productos, respectivamente. Por la experiencia pasada se sabe que 2, 3 y 2% de los productos ensamblados por cada máquina, respectivamente, tienen defectos.

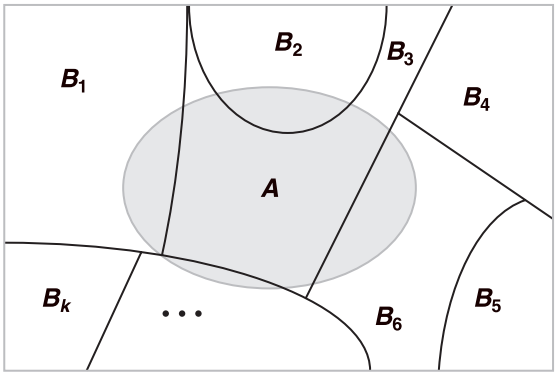


Figura 2.14: Partición del espacio muestral S.

Ahora, suponga que se selecciona de forma aleatoria un producto terminado. ¿Cuál es la probabilidad de que esté defectuoso?

Solución: Considere los siguientes eventos:

A: el producto está defectuoso,

B_1: el producto está ensamblado con la máquina B_1,

B_2: el producto está ensamblado con la máquina B_2,

B_3: el producto está ensamblado con la máquina B_3.

Al aplicar la regla de eliminación, podemos escribir

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3).$$

Con referencia al diagrama de árbol de la figura 2.15, encontramos que las tres ramas dan las probabilidades

$$\begin{aligned} P(B_1)P(A|B_1) &= (0.3)(0.02) = 0.006, \\ P(B_2)P(A|B_2) &= (0.45)(0.03) = 0.0135, \\ P(B_3)P(A|B_3) &= (0.25)(0.02) = 0.005, \end{aligned}$$

y de aquí

$$P(A) = 0.006 + 0.0135 + 0.005 = 0.0245.$$

En vez de preguntar por $P(A)$, por la regla de eliminación, suponga que consideramos ahora el problema de encontrar la probabilidad condicional $P(B_i|\,A)$ en el ejemplo 2.41. En otras palabras, suponga que se seleccionó un producto de forma aleatoria y está defectuoso. ¿Cuál es la probabilidad de que este producto fuera ensamblado con la máquina B_i? Preguntas de este tipo se pueden contestar usando el siguiente teorema, la **regla de Bayes**:

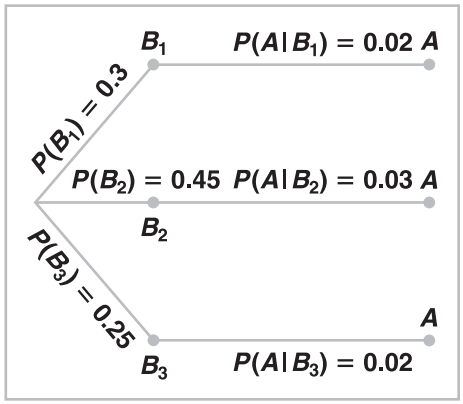


Figura 2.15: Diagrama de árbol para el ejemplo 2.41.

Teorema 2.17: (**Regla de Bayes**) Si los eventos B_1, $B_2, \ldots$, B_k constituyen una partición del espacio muestral S, donde $P(B_i) \neq 0$ para i / 1, $2, \ldots, k$, entonces, para cualquier evento A en S tal que $P(A) \neq 0$,

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} \text{ para } r = 1, 2, \ldots, k.$$

Prueba: Por la definición de probabilidad condicional,

$$P(B_r|A) = \frac{P(B_r \cap A)}{P(A)},$$

y con el teorema 2.16 en el denominador, tenemos

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)},$$

que completa la demostración.

Ejemplo 2.42: Con referencia al ejemplo 2.41, si se elige al azar un producto y se encuentra que está defectuoso, ¿cuál es la probabilidad de que esté ensamblado con la máquina B_3?

Solución: Utilizando la regla de Bayes para escribir

$$P(B_3|A) = \frac{P(B_3)P(A|B_3)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)},$$

y sustituyendo después las probabilidades calculadas en el ejemplo 2.41, tenemos

$$P(B_3|A) = \frac{0.005}{0.006 + 0.0135 + 0.005} = \frac{0.005}{0.0245} = \frac{10}{49}.$$

En vista del hecho de que se seleccionó un producto defectuoso, este resultado sugiere que probablemente no fue ensamblado con la máquina B_3.

Ejemplo 2.43: Una empresa de manufactura emplea tres planes analíticos para el diseño y desarrollo de un producto específico. Por razones de costos, los tres se utilizan en momentos diferentes. De hecho, los planes 1, 2 y 3 se utilizan, respectivamente, para 30, 20 y 50% de los productos. La "tasa de defectuosos" es diferente para los tres procedimientos, es decir,

$$P(D|P_1) = 0.01, \qquad P(D|P_2) = 0.03, \qquad P(D|P_3) = 0.02,$$

donde $P(D|P_j)$ es la probabilidad de un producto defectuoso, dado el plan j. Si se observa un producto al azar y se encuentra que está defectuoso, ¿cuál fue el plan que se usó con mayor probabilidad y fue el responsable?

Solución: De la declaración del problema

$$P(P_1) = 0.30, \quad P(P_2) = 0.20, \quad \text{y} \quad P(P_3) = 0.50,$$

debemos encontrar $P(P_j|D)$ para $j = 1, 2, 3$. La regla de Bayes del teorema 2.17 muestra

$$\begin{aligned} P(P_1|D) &= \frac{P(P_1)P(D|P_1)}{P(P_1)P(D|P_1) + P(P_2)P(D|P_2) + P(P_3)P(D|P_3)} \\ &= \frac{(0.30)(0.01)}{(0.3)(0.01) + (0.20)(0.03) + (0.50)(0.02)} = \frac{0.003}{0.019} = 0.158. \end{aligned}$$

Asimismo,

$$P(P_2|D) = \frac{(0.03)(0.20)}{0.019} = 0.316 \quad \text{y} \quad P(P_3|D) = \frac{(0.02)(0.50)}{0.019} = 0.526.$$

La probabilidad condicional de un defecto dado el plan 3 es la mayor de las tres; de manera que el resultado de un defecto en un producto elegido al azar es más probable usando el plan 3.

Con la regla de Bayes, un método estadístico, llamado método bayesiano, tiene mucha utilidad para las aplicaciones. En el capítulo 18 estudiaremos una introducción al método bayesiano.

Ejercicios

2.101 En cierta región del país se sabe por experiencia que la probabilidad de seleccionar un adulto mayor de 40 años de edad con cáncer es 0.05. Si la probabilidad de que un doctor diagnostique de forma correcta que una persona con cáncer tiene la enfermedad es 0.78, y la probabilidad de que diagnostique de forma incorrecta que una persona sin cáncer tiene la enfermedad es 0.06, ¿cuál es la probabilidad de que a una persona se le diagnostique cáncer?

2.102 La policía planea hacer cumplir los límites de velocidad usando un sistema de radar en 4 diferentes puntos dentro de la ciudad. Las trampas de radar en cada uno de los sitios L_1, L_2, L_3 y L_4 operan 40, 30, 20 y 30% del tiempo, y si una persona maneja a gran velocidad cuando va a su trabajo tiene las probabilidades de 0.2, 0.1, 0.5 y 0.2, respectivamente, de pasar por esos lugares. ¿Cuál es la probabilidad de que reciba una multa por conducir con exceso de velocidad?

2.103 Refiérase al ejercicio 2.101. ¿Cuál es la probabilidad de que una persona a la que se le diagnostica cáncer realmente tenga la enfermedad?

2.104 Si en el ejercicio 2.102 la persona es multada por conducir con exceso de velocidad en su camino al

trabajo, ¿cuál es la probabilidad de que pase por el sistema de radar que se ubica en L_2?

2.105 Suponga que los cuatro inspectores de una fábrica de película colocan la fecha de caducidad en cada paquete de película al final de la línea de montaje. John, quien coloca la fecha de caducidad en 20% de los paquetes, no la pone una vez en cada 200 paquetes; Tom, quien la coloca en 60% de los paquetes, no la coloca una vez en cada 100 paquetes; Jeff, quien la coloca en 15% de los paquetes, no lo hace una vez en cada 90 paquetes; y Pat, que fecha 5% de los paquetes, falla una vez en cada 200 paquetes. Si un consumidor se queja de que su paquete de película no muestra la fecha de caducidad, ¿cuál es la probabilidad de que haya sido inspeccionado por John?

2.106 Una compañía telefónica regional opera tres estaciones de retransmisión idénticas en diferentes sitios. Durante un periodo de un año, el número de desperfectos reportados por cada estación y las causas se muestran a continuación.

Estaciones	**A**	**B**	**C**
Problemas con el suministro de electricidad	2	1	1
Desperfecto de la computadora	4	3	2
Fallas del equipo eléctrico	5	4	2
Fallas ocasionadas por otros errores humanos	7	7	5

Suponga que se reporta una falla y que se encuentra que fue ocasionada por otros errores humanos. ¿Cuál es la probabilidad de que provenga de la estación **C**?

2.107 La contaminación de los ríos en Estados Unidos es un problema desde hace varios años. Considere los siguientes eventos:

$A = \{$El río está contaminado.$\}$

$B = \{$Una prueba en una muestra de agua detecta contaminación.$\}$

$C = \{$Se permite la pesca.$\}$

Suponga $P(A) = 0.3$, $P(B|A) = 0.75$, $P(B|A') = 0.20$, $P(C|A \cap B) = 0.20$, $P(C|A' \cap B) = 0.15$, $P(C|A \cap B') = 0.80$ y $P(C|A' \cap B) = 0.90$.

a) Encuentre $P(A \cap B \cap C)$.

b) Encuentre $P(B' \cap C)$.

c) Encuentre $P(C)$.

d) Encuentre la probabilidad de que el río esté contaminado, dado que se permite la pesca y que la prueba de la muestra no detecta contaminación.

2.108 Una cadena de tiendas de pintura produce y vende pintura látex y semiesmaltada. Con base en las ventas de largo plazo, la probabilidad de que un cliente compre pintura látex es 0.75. De los que compran pintura de látex, 60% también compran rodillos. Pero 30% de los compradores de pintura semiesmaltada compran rodillos. Un comprador que se selecciona al azar compra un rodillo y una lata de pintura. ¿Cuál es la probabilidad de que sea pintura látex?

Ejercicios de repaso

2.109 Un suero de la verdad tiene la propiedad de que 90% de los sospechosos culpables se juzgan de forma adecuada; mientras que, por supuesto, 10% de los sospechosos culpables erróneamente se consideran inocentes. Por otro lado, a los sospechosos inocentes se les juzga de manera errónea 1% de las veces. Si el sospechoso se selecciona de un grupo de sospechosos, de los cuales sólo 5% alguna vez han cometido un delito, y el suero indica que es culpable, ¿cuál es la probabilidad de que sea inocente?

2.110 Una alergista afirma que 50% de los pacientes que examina son alérgicos a algún tipo de hierba. ¿Cuál es la probabilidad de que

a) exactamente tres de sus cuatro próximos pacientes sean alérgicos a hierbas?

b) ninguno de sus siguientes 4 pacientes sea alérgico a hierbas?

2.111 Mediante la comparación de las regiones apropiadas en un diagrama de Venn, verifique que

a) $(A \cap B) \cup (A \cap B') = A$;

b) $A' \cap (B' \cup C) = (A' \cap B') \cup (A' \cap C)$.

2.112 Las probabilidades de que una estación de servicio bombee gasolina en 0, 1, 2, 3, 4, 5 o más automóviles durante cierto periodo de 30 minutos son, respectivamente, 0.03, 0.18, 0.24, 0.28, 0.10 y 0.17. Encuentre la probabilidad de que en este periodo de 30 minutos

a) más de 2 automóviles reciban gasolina;

b) a lo más 4 automóviles reciban gasolina;

c) 4 o más automóviles reciban gasolina.

2.113 ¿Cuántas manos de *bridge* que contengan 4 espadas, 6 diamantes, 1 trébol y 2 corazones son posibles?

2.114 Si la probabilidad de que una persona cometa un error en su declaración de impuestos sobre la renta es 0.1, encuentre la probabilidad de que

a) cuatro personas no relacionadas cometan cada una un error;

b) el señor Jones y la señora Clark cometan un error, y el señor Roberts y la señora Williams no cometan errores.

2115 Una empresa industrial grande usa tres hoteles locales para ofrecer hospedaje nocturno a sus clientes. Por la experiencia pasada se sabe que a 20% de los clientes se les asignan habitaciones en el Ramada Inn, a 50% en el Sheraton y a 30% en el Lakeview Motor Lodge. Si hay una falla en la plomería en 5% de las habitaciones del Ramada Inn, en 4% de las habitaciones del Sheraton y en 8% de las habitaciones del Lakeview Motor Lodge, ¿cuál es la probabilidad de que

a) a un cliente se le asigne una habitación con fallas en la plomería?

b) a una persona con una habitación que tiene fallas de plomería se le haya asignado acomodo en el Lakeview Motor Lodge?

2.116 De un grupo de 4 hombres y 5 mujeres, ¿cuántos comités de 3 miembros son posibles

a) sin restricciones?

b) con 1 hombre y 2 mujeres?

c) con 2 hombres y 1 mujer, si cierto hombre debe estar en el comité?

2.117 La probabilidad de que un paciente se recupere de una operación de corazón delicada es 0.8. ¿Cuál es la probabilidad de que

a) exactamente 2 de los siguientes 3 pacientes que tienen esta operación sobrevivan?

b) los siguientes 3 pacientes que tengan esta operación sobrevivan?

2.118 En cierta prisión federal se sabe que 2/3 de los reclusos son menores de 25 años de edad. También se sabe que 3/5 de los reos son hombres y que 5/8 son mujeres de 25 años de edad o mayores. ¿Cuál es la probabilidad de que un prisionero seleccionado al azar de esta prisión sea mujer y de al menos 25 años de edad?

2.119 De 4 manzanas rojas, 5 verdes y 6 amarillas, ¿cuántas selecciones de 9 manzana son posibles si se deben seleccionar 3 de cada color?

2.120 De una caja que contiene 6 bolas negras y 4 verdes se extraen tres bolas sucesivamente, cada bola se reemplaza en la caja antes de que se extraiga la siguiente. ¿Cuál es la probabilidad de que

a) las 3 sean del mismo color?

b) cada color esté representado?

2.121 Un cargamento de 12 televisores contiene tres defectuosos. ¿De cuántas formas un hotel puede comprar 5 de estas unidades y recibir al menos 2 defectuosas?

2.122 Se examinaron los planes de estudio de ingeniería eléctrica, química, industrial y mecánica. Se encontró que algunos estudiantes no cursan estadística, algunos cursan un semestre y otros cursan dos semestres. Considere los siguientes eventos:

A: Se cursa algo de estadística.

B: Ingenieros eléctricos e industriales.

C: Ingenieros químicos.

Utilice diagramas de Venn y sombree las áreas que representan los siguientes eventos:

a) $(A \cap B)'$;

b) $(A \cup B)'$;

c) $(A \cap C) \cup B$.

2.123 Cierta dependencia federal emplea a tres empresas consultoras (A, B y C) con probabilidades de 0.40, 0.35 y 0.25, respectivamente. De la experiencia pasada se sabe que las probabilidades de excesos en costos de las empresas son 0.05, 0.03 y 0.15, respectivamente. Suponga que la agencia experimenta un exceso en los costos.

a) ¿Cuál es la probabilidad de que la empresa consultora implicada sea la compañía C?

b) ¿Cuál es la probabilidad de que sea la compañía A?

2.124 Un fabricante estudia los efectos de la temperatura de cocción, tiempo de cocción y tipo de aceite para la cocción al elaborar papas fritas. Se utilizan 3 diferentes temperaturas, 4 diferentes tiempos de cocción y 3 diferentes aceites.

a) ¿Cuál es el número total de combinaciones a estudiar?

b) ¿Cuántas combinaciones se utilizarán para cada tipo de aceite?

c) Discuta por qué las permutaciones no son un problema en este ejercicio.

2.125 Considere la situación del ejercicio 2.124 y suponga que el fabricante puede probar sólo dos combinaciones en un día.

a) ¿Cuál es la probabilidad de que se elija cualquier conjunto dado de 2 corridas?

b) ¿Cuál es la probabilidad de que se utilice la temperatura más alta en cualquiera de estas 2 combinaciones?

2.126 Se sabe que en las mujeres de más de 60 años se desarrolla cierta forma de cáncer con una probabilidad de 0.07. Se dispone de una prueba de sangre para la detección de tal padecimiento, aunque no es infalible. De hecho, se sabe que 10% de las veces la prueba da negativo falso (es decir, incorrectamente la prueba da un resultado negativo) y 5% de las veces la prueba da positivo falso (es decir, incorrectamente la prueba da un resultado positivo). Si una mujer de más de 60 años que se sometió a la prueba y recibió un resultado favorable (negativo), ¿cuál es la probabilidad de que ella tenga la enfermedad?

2.127 Un fabricante de cierto tipo de componente electrónico abastece a los proveedores en lotes de 20. Suponga que 60% de todos los lotes no contienen componentes defectuosos, que 30% contienen un componente defectuoso y que 10% contienen dos componentes defectuosos. Se elige un lote y de éste se extraen aleatoriamente dos componentes, los cuales se prueban con el resultado de que ninguno está defectuoso.

a) ¿Cuál es la probabilidad de que haya cero componentes defectuosos en el lote?

b) ¿Cuál es la probabilidad de que haya uno defectuoso en el lote?

c) ¿Cuál es la probabilidad de que haya dos defectuosos en el lote?

2.128 Hay una extraña enfermedad que sólo afecta a 1 de cada 500 individuos. Se dispone de una prueba para detectarla, pero desde luego ésta no es infalible. Un resultado correcto positivo (un paciente que realmente tiene la enfermedad) ocurre 95% de las veces; en tanto que un resultado positivo falso (un paciente que no tiene la enfermedad) ocurre 1% de las veces. Si se somete a prueba un individuo elegido al azar, y el resultado es positivo, ¿cuál es la probabilidad de que el individuo tenga la enfermedad?

2.129 Una compañía constructora emplea a 2 ingenieros de ventas. El ingeniero 1 hace el trabajo de estimar costos en 70% de las cotizaciones solicitadas a la empresa. El ingeniero 2 lo hace para 30% de tales cotizaciones. Se sabe que la tasa de error para el ingeniero 1 es tal que 0.02 es la probabilidad de un error cuando éste hace el trabajo; mientras que la probabilidad de un error en el trabajo del ingeniero 2 es 0.04. Suponga que llega una solicitud de cotización y ocurre un error grave al estimar los costos. ¿Qué ingeniero supondría usted que hizo el trabajo? Explique y muestre todo el desarrollo.

2.130 En el rubro del control de la calidad la ciencia estadística a menudo se utiliza para determinar si un proceso está "fuera de control". Suponga que el proceso, de hecho, está fuera de control y que 20% de los artículos producidos están defectuosos.

a) Si tres artículos salen en serie de la línea de proceso, ¿cuál es la probabilidad de que los tres estén defectuosos?

b) Si salen cuatro artículos en serie, ¿cuál es la probabilidad de que tres estén defectuosos?

2.131 En una planta industrial se está realizando un estudio para determinar qué tan rápido los trabajadores lesionados regresan a sus labores después del percance. Los registros demuestran que 10% de todos los trabajadores lesionados llegan al hospital para atención y 15% están de vuelta en su trabajo al día siguiente. Además, los estudios demuestran que 2% llegan al hospital y están de vuelta al trabajo al día siguiente. Si un trabajador se lesiona, ¿cuál es la probabilidad de que llegue al hospital o regrese al trabajo al día siguiente, o ambas?

2.132 Una empresa acostumbra a capacitar operadores que realizan ciertas actividades en la línea de producción. Se sabe que los operadores que asisten al curso de capacitación son capaces de cumplir sus cuotas de producción 90% de las veces. Los nuevos operarios que no toman el curso de capacitación sólo cumplen con sus cuotas 65% de las veces. Cincuenta por ciento de los nuevos operadores asisten al curso. Dado que un nuevo operador cumple con su cuota de producción, ¿cuál es la probabilidad de que él (o ella) haya asistido al curso?

2.133 Una encuesta aplicada a quienes usan un software estadístico específico indica que 10% no quedaron satisfechos. La mitad de quienes no quedaron satisfechos compraron el sistema al vendedor A. Se sabe que 20% de los encuestados compraron al vendedor A. Dado que el paquete de software se compró del vendedor A, ¿cuál es la probabilidad de que ese usuario específico haya quedado insatisfecho?

2.134 Durante las crisis económicas, se despide a obreros y a menudo se les reemplaza con máquinas. Se revisa la historia de 100 trabajadores cuya pérdida del empleo se atribuye a los avances tecnológicos. Por cada uno de esos individuos se determinó si el o ella recibieron un empleo alternativo dentro de la misma compañía, si encontraron un empleo en otra compañía pero trabajando en la misma área, si encontraron trabajo en una nueva área o si llevan desempleados más de un año. Además, se registró el estatus sindical de cada trabajador. La siguiente tabla resume los resultados.

	Sindicalizado	**No sindicalizado**
Sigue en la misma compañía	40	15
Está en otra compañía (en la misma área)	13	10
Está en una nueva área	4	11
Está desempleado	2	5

a) Si los trabajadores seleccionados encontraron empleo en una nueva compañía en la misma área, ¿cuál es la probabilidad de que el trabajador sea miembro de un sindicato?

b) Si el trabajador es miembro de un sindicato, ¿cuál es la probabilidad de que esté desempleado desde hace un año?

2.135 Hay una probabilidad de 50-50 de que la reina tenga el gen de la hemofilia. Si lo tiene, entonces cada uno de los príncipes tiene una probabilidad de 50-50 de tener hemofilia independientemente. Si la reina no tiene el gen, el príncipe no tendrá la enfermedad. Suponga que la reina tuvo tres príncipes que no padecen la enfermedad, ¿cuál es la probabilidad de que la reina tenga el gen?

2.136 ¿Cuál es la probabilidad de que dos estudiantes no tengan la misma fecha de cumpleaños en un grupo de 60 alumnos? (Véase el ejercicio 2.50.)

Capítulo 3

Variables aleatorias y distribuciones de probabilidad

3.1 Concepto de variable aleatoria

La estadística realiza inferencias acerca de las poblaciones y sus características. Se llevan a cabo experimentos cuyos resultados se encuentran sujetos al azar. La prueba de un número de componentes electrónicos es un ejemplo de **experimento estadístico**, que es un concepto que se utiliza para describir cualquier proceso mediante el cual se generan varias observaciones al azar. Con frecuencia es importante asignar una descripción numérica al resultado. Por ejemplo, el espacio muestral que ofrece una descripción detallada de cada posible resultado, cuando se prueban tres componentes electrónicos, se escribe como

$$S = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\},$$

donde N denota "no defectuoso"; y D, "defectuoso". Evidentemente, nos interesa el número de defectuosos que se presenten. De esta forma, a cada punto en el espacio muestral se le *asignará un valor numérico* de 0, 1, 2 o 3. Estos valores son, por supuesto, cantidades aleatorias *determinadas por el resultado del experimento*. Se pueden ver como valores que toma la *variable aleatoria* X, es decir, el número de artículos defectuosos cuando se prueban tres componentes electrónicos.

Definición 3.1: Una **variable aleatoria** es una función que asocia un número real con cada elemento del espacio muestral.

Utilizaremos una letra mayúscula, digamos X, para denotar una variable aleatoria; y su correspondiente letra minúscula, x en este caso, para uno de sus valores. En el ejemplo de la prueba de componentes electrónicos, observamos que la variable aleatoria X toma el valor 2 para todos los elementos en el subconjunto

$$E = \{DDN, DND, NDD\}$$

del espacio muestral S. Esto es, cada valor posible de X representa un evento que es un subconjunto del espacio muestral para el experimento dado.

Ejemplo 3.1: Se sacan 2 bolas de manera sucesiva sin reemplazo, de una urna que contiene 4 bolas rojas y 3 negras. Los posibles resultados y los valores y de la variable aleatoria Y, donde Y es el número de bolas rojas, son

Espacio Muestral	y
RR	2
RB	1
BR	1
BB	0

Ejemplo 3.2: El empleado de un almacén regresa tres cascos de seguridad al azar a tres trabajadores de un taller siderúrgico que ya los habían probado. Si Smith, Jones y Brown, en ese orden, reciben uno de los tres cascos, liste los puntos muestrales para los posibles órdenes de regreso de los cascos, y encuentre el valor m de la variable aleatoria M que representa el número de asociaciones correctas.

Solución: Si S, J y B representan, respectivamente, los cascos de Smith, Jones y Brown, entonces los posibles arreglos en los cuales se pueden regresar los cascos y el número de asociaciones correctas son

Espacio Muestral	m
SJB	3
SBJ	1
BJS	1
JSB	1
JBS	0
BSJ	0

En cada uno de los dos ejemplos anteriores, el espacio muestral contiene un número finito de elementos. Por otro lado, cuando se lanza un dado hasta que salga un 5, obtenemos un espacio muestral con una secuencia de elementos interminable,

$$S = \{F,\ NF,\ NNF,\ NNNF, \dots\},$$

donde F y N representan, respectivamente, la ocurrencia y la no ocurrencia de un 5. Sin embargo, incluso en este experimento el número de elementos puede ser igual a todos los números enteros, de manera que hay un primer elemento, un segundo, un tercero y así sucesivamente, y en este sentido se pueden contar.

Hay casos en que la variable aleatoria es categórica por naturaleza y se utilizan las llamadas variables *ficticias o indicadoreas*. Un buen ejemplo de ello es el caso en que la variable aleatoria es binaria por naturaleza, como se indica en el siguiente ejemplo.

Ejemplo 3.3: Considere la condición en que los componentes llegan de la línea de ensamble y se les clasifica como *defectuosos* o *no defectuosos*. Defina la variable aleatoria X mediante

$$X = \begin{cases} 1, & \text{si el componente está defectuoso.} \\ 0, & \text{si el componente no está defectuoso.} \end{cases}$$

Evidentemente la asignación de 1 o 0 es arbitraria, aunque bastante conveniente, lo cual se volverá más claro conforme avancemos en los siguientes capítulos. La variable aleatoria en la que se eligen 0 y 1 para describir dos posibles valores se denomina **variable aleatoria de Bernoulli**.

Veremos más casos de variables aleatorias en los siguientes cuatro ejemplos.

Ejemplo 3.4: Los estadísticos utilizan **planes de muestreo** ya sea para aceptar o para rechazar lotes de materiales. Suponga que uno de los planes de muestreo implica el muestreo independiente de 10 artículos de un lote de 100 de ellos, donde 12 están defectuosos.

Sea X la variable aleatoria definida como el número de artículos que están defectuososo en la muestra de 10. En este caso, la variable aleatoria toma los valores 0, 1, 2, . . . , 9, 10.

Ejemplo 3.5: Suponga que un plan de muestreo implica el muestreo de artículos de un proceso hasta que se encuentre uno defectuoso. La evaluación del proceso dependerá de cuántos artículos consecutivos se observen. En ese aspecto, sea X una variable aleatoria que se define como el número de artículos observados antes de que salga uno defectuoso. Se asigna N a no defectuoso, y D a defectuoso; los espacios muestrales son $S = (D)$ dado que $X = 1$, $S = (ND)$ dado que $X = 2$, $S = (NND)$ dado que $X = 3$, y así sucesivamente.

Ejemplo 3.6: El interés se centra en la proporción de personas que responden a cierta encuesta enviada por correo. Sea X tal proporción. X es una variable aleatoria que toma todos los valores de x para los cuales $0 \leq x \leq 1$.

Ejemplo 3.7: Sea X la variable aleatoria definida como el tiempo de espera, en horas, entre conductores sucesivos que exceden los límites de velocidad detectados por una unidad de radar. La variable aleatoria X toma todos los valores de x tales que $x \geq 0$.

Definición 3.2: Si un espacio muestral contiene un número finito de posibilidades, o una serie interminable con tantos elementos como números enteros existen, se llama **espacio muestral discreto**.

Los resultados de algunos experimentos estadísticos no pueden ser ni finitos ni contables. Es el caso, por ejemplo, cuando se realiza una investigación para medir las distancias que recorre cierta marca de automóvil, en una ruta de prueba preestablecida, con cinco litros de gasolina. Supongamos que la distancia es una variable que se mide con algún grado de precisión, entonces claramente tenemos un número infinito de distancias posibles en el espacio muestral, que no se pueden igualar a todos los números enteros. También, si se registrara el tiempo requerido para que ocurra una reacción química, una vez más los posibles intervalos de tiempo que forman nuestro espacio muestral son un número infinito e incontable. Vemos ahora que no todos los espacios muestrales necesitan ser discretos.

Definición 3.3: Si un espacio muestral contiene un número infinito de posibilidades igual al número de puntos en un segmento de línea, se le llama **espacio muestral continuo**.

Una variable aleatoria se **llama variable aleatoria discreta** si se puede contar su conjunto de resultados posibles. En los ejemplos 3.1 a 3.5 las variables aleatorias

son discretas. Sin embargo, una variable aleatoria cuyo conjunto de valores posibles es un intervalo completo de números no es discreta. Cuando una variable aleatoria puede tomar valores en una escala continua, se le denomina **variable aleatoria continua**. A menudo los posibles valores de una variable aleatoria continua son precisamente los mismos valores que contiene el espacio muestral continuo. Evidentemente en los ejemplos 3.6 y 3.7 se trata de variables aleatorias continuas.

En la mayoría de los problemas prácticos, las variables aleatorias continuas representan datos *medidos*, como serían todos los posibles pesos, alturas, temperaturas, distancias o periodos de vida; en tanto que las variables aleatorias discretas representan datos *por conteo*, como el número de artículos defectuosos en una muestra de k artículos o el número de accidentes de carretera por año en una entidad específica. Observe que las variables aleatorias Y y M de los ejemplos 3.1 y 3.2 representan ambas datos por conteo: Y el número de bolas rojas y M el número de asignaciones de cascos correctas.

3.2 Distribuciones discretas de probabilidad

Una variable aleatoria discreta toma cada uno de sus valores con cierta probabilidad. Al lanzar una moneda tres veces, la variable X, que representa el número de caras, toma el valor 2 con probabilidad 3/8, pues 3 de los 8 puntos muestrales igualmente probables tienen como resultado dos caras y una cruz. Si se suponen pesos iguales para los eventos simples del ejemplo 3.2, la probabilidad de que ningún empleado obtenga de vuelta su casco correcto, es decir, la probabilidad de que M tome el valor cero, es 1/3. Los valores posibles m de M y sus probabilidades son

m	0	1	3
$P(M = m)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

Note que los valores de m agotan todos los casos posibles y por ello las probabilidades suman 1.

Con frecuencia es conveniente representar todas las probabilidades de una variable aleatoria X usando una fórmula, la cual necesariamente sería una función de los valores numéricos x que denotaremos con $f(x)$, $g(x)$, $r(x)$, y así sucesivamente. Por lo tanto, escribimos $f(x) = P(X = x)$; es decir, $f(3) = P(X = 3)$. El conjunto de pares ordenados $(x, f(x))$ se llama **función de probabilidad** o **distribución de probabilidad** de la variable aleatoria discreta X.

Definición 3.4: El conjunto de pares ordenados $(x, f(x))$ es una **función de probabilidades**, una **función de masa de probabilidad** o una **distribución de probabilidad** de la variable aleatoria discreta X si, para cada resultado posible x,

1. $f(x) \geq 0$,
2. $\sum_x f(x) = 1$,
3. $P(X = x) = f(x)$.

Ejemplo 3.8: Un embarque de 8 microcomputadoras similares para una tienda al detalle contiene 3 que están defectuosas. Si una escuela hace una compra al azar de dos de estas computadoras, encuentre la distribución de probabilidad para el número de defectuosas.

Solución: Sea X una variable aleatoria cuyos valores x son los números posibles de computadoras defectuosas que la escuela compra. Entonces, x puede ser cualquiera de los números 0, 1 y 2. Así,

$$f(0) = P(X = 0) = \frac{\binom{3}{0}\binom{5}{2}}{\binom{8}{2}} = \frac{10}{28},$$
$$f(1) = P(X = 1) = \frac{\binom{3}{1}\binom{5}{1}}{\binom{8}{2}} = \frac{15}{28},$$
$$f(2) = P(X = 2) = \frac{\binom{3}{2}\binom{5}{0}}{\binom{8}{2}} = \frac{3}{28}.$$

De manera que la distribución de probabilidad de X es

x	0	1	2
$f(x)$	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$

Ejemplo 3.9: Si una agencia automotriz vende 50% de su inventario de cierto vehículo extranjero equipado con bolsas de aire, encuentre una fórmula para la distribución de probabilidad del número de automóviles con bolsas de aire entre los siguientes 4 vehículos que venda la agencia.

Solución: Como la probabilidad de vender un automóvil con bolsas de aire es 0.5, los $2^4 = 16$ puntos del espacio muestral tienen la misma probabilidad de ocurrencia. Por lo tanto, el denominador para todas las probabilidades, y también para nuestra función, es 16. Para obtener el número de formas de vender tres modelos con bolsas de aire, necesitamos considerar el número de formas de dividir 4 resultados en dos celdas con 3 modelos con bolsas de aire asignadas a una celda, y el modelo sin bolsas de aire asignado a la otra. Esto se puede hacer de $\binom{4}{3} = 4$ formas. En general, el evento de vender x modelos con bolsas de aire y $4 - x$ modelos sin bolsas de aire puede ocurrir de $\binom{4}{x}$ formas, donde x puede ser 0, 1, 2, 3 o 4. Entonces, la distribución de probabilidad $f(x) = P(X = x)$ es

$$f(x) = \frac{\binom{4}{x}}{16}, \quad \text{para } x = 0, 1, 2, 3, 4.$$

Hay muchos problemas donde queremos calcular la probabilidad de que el valor observado de una variable aleatoria X sea menor o igual que algún número real x. Al escribir $F(x) = P(X \leq x)$ para cualquier número real x, definimos $F(x)$ como la **función de la distribución acumulada** de la variable aleatoria X.

Definición 3.5: La **función de la distribución acumulada** $F(x)$ de una variable aleatoria discreta X con distribución de probabilidad $f(x)$ es

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{para } -\infty < x < \infty.$$

Para la variable aleatoria M, el número de asociaciones correctas en el ejemplo 3.2, tenemos

$$F(2) = P(M \leq 2) = f(0) + f(1) = \frac{1}{3} + \frac{1}{2} = \frac{5}{6}.$$

La distribución acumulada de M es

$$F(m) = \begin{cases} 0, & \text{para } m < 0, \\ \frac{1}{3}, & \text{para } 0 \leq m < 1, \\ \frac{5}{6}, & \text{para } 1 \leq m < 3, \\ 1, & \text{para } m \geq 3. \end{cases}$$

Se debería notar en particular el hecho de que la distribución acumulada es una función no decreciente monótona que se define no sólo para los valores que toma la variable aleatoria dada, sino para todos los números reales.

Ejemplo 3.10: Encuentre la función de la distribución acumulada de la variable aleatoria X del ejemplo 3.4. Mediante el uso de $F(x)$, verifique que $f(2) = 3/8$.

Solución: El cálculo directo de la distribución de probabilidad del ejemplo 3.4 da $f(0) = 1/16$, $f(1) = 1/4$, $f(2) = 3/8$, $f(3) = 1/4$ y $f(4) = 1/16$. Por lo tanto,

$$\begin{aligned} F(0) &= f(0) = \frac{1}{16}, \\ F(1) &= f(0) + f(1) = \frac{5}{16}, \\ F(2) &= f(0) + f(1) + f(2) = \frac{11}{16}, \\ F(3) &= f(0) + f(1) + f(2) + f(3) = \frac{15}{16}, \\ F(4) &= f(0) + f(1) + f(2) + f(3) + f(4) = 1. \end{aligned}$$

De aquí,

$$F(x) = \begin{cases} 0, & \text{para } x < 0, \\ \frac{1}{16}, & \text{para } 0 \leq x < 1, \\ \frac{5}{16}, & \text{para } 1 \leq x < 2, \\ \frac{11}{16}, & \text{para } 2 \leq x < 3, \\ \frac{15}{16}, & \text{para } 3 \leq x < 4, \\ 1 & \text{para } x \geq 4. \end{cases}$$

Entonces,

$$f(2) = F(2) - F(1) = \frac{11}{16} - \frac{5}{16} = \frac{3}{8}.$$

A menudo es útil ver una distribución de probabilidad en forma gráfica. Se pueden graficar los puntos $(x, f(x))$ del ejemplo 3.9 para obtener la figura 3.1. Al unir los puntos al eje x, ya sea con una línea punteada o con una sólida, obtenemos lo que, por lo general, se denomina como **gráfica de barras**. La figura 3.1 permite ver fácilmente qué valores de X tienen más probabilidad de ocurrencia, y también indica, en este caso, una situación perfectamente simétrica.

En vez de graficar los puntos $(x, f(x))$, más a menudo construimos rectángulos, como en la figura 3.2. Aquí los rectángulos se construyen de manera que sus bases de igual ancho se centren en cada valor x, y sus alturas sean iguales a las probabi-

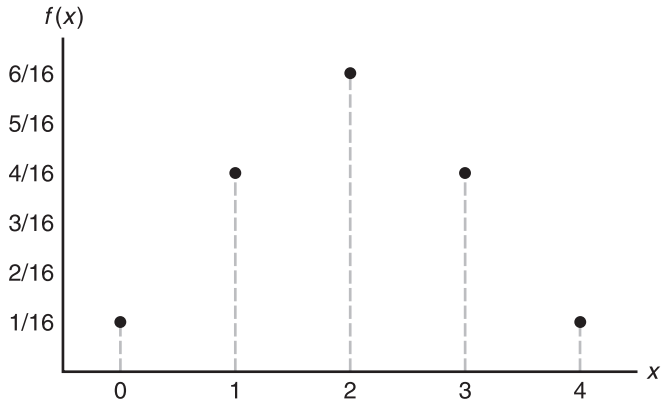


Figura 3.1: Gráfica de barras.

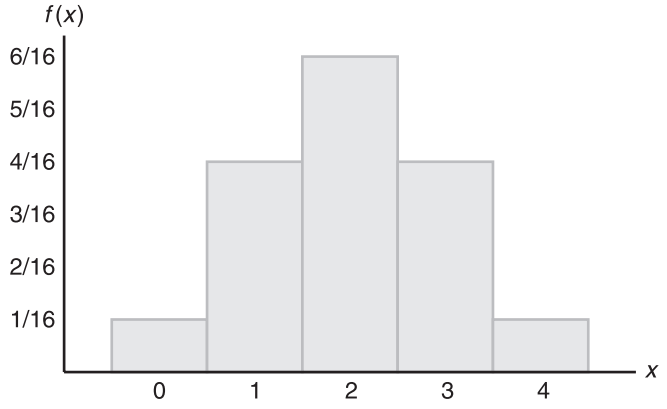


Figura 3.2: Histograma de probabilidad.

lidades correspondientes dadas por $f(x)$. Las bases se construyen de forma tal que no dejen espacios entre los rectángulos. La figura 3.2 se denomina **histograma de probabilidad**.

En la figura 3.2 como cada base tiene ancho unitario, $P(X = x)$ es igual al área del rectángulo centrado en x. Incluso si las bases no fueran de ancho unitario, podríamos ajustar las alturas de los rectángulos para que las áreas tuvieran probabilidades iguales a X de tomar cualquiera de sus valores x. Este concepto de utilizar áreas para representar probabilidades es necesario para nuestra consideración de la distribución de probabilidad de una variable aleatoria continua.

La gráfica de la distribución acumulada del ejemplo 3.9, que aparece como una función escalonada en la figura 3.3, se obtiene al graficar los puntos $(x, F(x))$.

Ciertas distribuciones de probabilidad se aplican a más de una situación física. La distribución de probabilidad del ejemplo 3.9 también se aplica a la variable aleatoria Y, donde Y es el número de caras cuando se lanza 4 veces una moneda, o a la variable aleatoria W, donde W es el número de cartas rojas que resultan cuando se

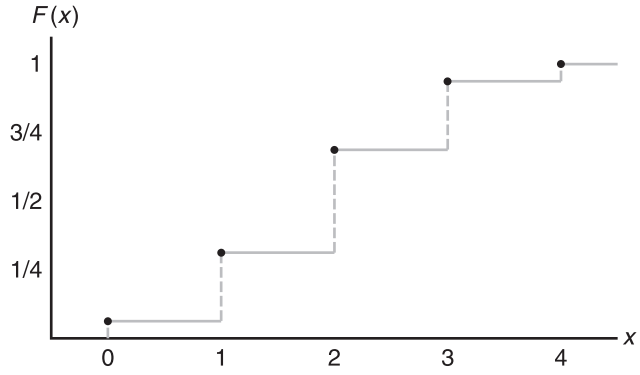


Figura 3.3: Distribución acumulada discreta.

sacan sucesivamente 4 cartas al azar, de una baraja con el reemplazo de cada carta y barajando antes de sacar la siguiente. En el capítulo 5 se considerarán distribuciones discretas especiales que se aplican a diversas situaciones experimentales.

3.3 Distribuciones continuas de probabilidad

Una variable aleatoria continua tiene una probabilidad cero de tomar *exactamente* cualquiera de sus valores. En consecuencia, su distribución de probabilidad no se puede dar en forma tabular. En un principio esto parecería sorprendente; no obstante, se vuelve más convincente si consideramos un ejemplo específico. Consideremos una variable aleatoria cuyos valores son las alturas de toda la gente mayor de 21 años de edad. Entre cualesquiera dos valores, digamos 163.5 y 164.5 centímetros, o incluso entre 163.99 y 164.01 centímetros, hay un número infinito de alturas, una de las cuales es 164 centímetros. Es remota la probabilidad de seleccionar al azar a alguien que tenga exactamente 164 centímetros de estatura y no sea del conjunto infinitamente grande de estaturas tan cercanas a 164 centímetros, que humanamente no es posible medir la diferencia; por ello, asignamos una probabilidad cero a tal evento. Éste no es el caso, sin embargo, si nos referimos a la probabilidad de seleccionar a una persona que, al menos, mida 163 centímetros pero no más de 165 centímetros de estatura. Tratamos ahora con un intervalo en vez de un valor puntual de nuestra variable aleatoria.

Trataremos el cálculo de probabilidades para varios intervalos de variables aleatorias continuas como $P(a < X < b)$, $P(W \geq c)$, etcétera. Observe que cuando X es continua,

$$P(a < X \leq b) = P(a < X < b) + P(X = b) = P(a < X < b).$$

Es decir, no importa si incluimos o no un extremo del intervalo. Esto no es cierto, sin embargo, cuando X es discreta.

Aunque la distribución de probabilidad de una variable aleatoria continua no se puede representar de forma tabular, sí se establece como una fórmula, la cual necesariamente será función de los valores numéricos de la variable aleatoria continua X y como tal se representará mediante la notación funcional $f(x)$. Al tratar con variables continuas, $f(x)$, por lo general, se llama **función de densidad de**

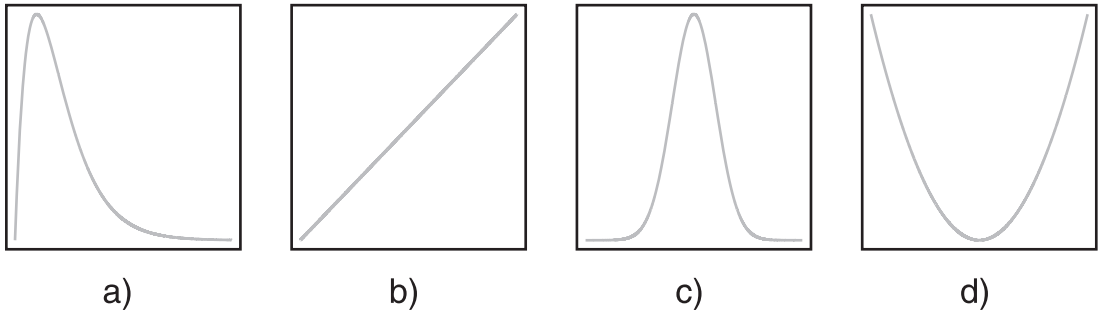


Figura 3.4: Funciones de densidad típicas.

probabilidad, o simplemente **función de densidad** de X. Como X se define sobre un espacio muestral continuo, es posible que $f(x)$ tenga un número finito de discontinuidades. Sin embargo, la mayoría de las funciones de densidad que tienen aplicaciones prácticas en el análisis de datos estadísticos son continuas y sus gráficas pueden tomar cualquiera de varias formas, algunas de las cuales se presentan en la figura 3.4. Como se utilizarán áreas para representar probabilidades y éstas son valores numéricos positivos, la función de densidad debe estar completamente por arriba del eje x.

Una función de densidad de probabilidad se construye de manera que el área bajo su curva limitada por el eje x sea igual a 1, cuando se calcula en el rango de X para el que se define $f(x)$. Si este rango de X es un intervalo finito, siempre es posible extender el intervalo para incluir a todo el conjunto de números reales al definir $f(x)$ como cero en todos los puntos de las partes extendidas del intervalo. En la figura 3.5, la probabilidad de que X tome un valor entre a y b es igual al área sombreada bajo la función de densidad entre las ordenadas en $x = a$ y $x = b$, y del cálculo integral está dada por

$$P(a < X < b) = \int_a^b f(x)\, dx.$$

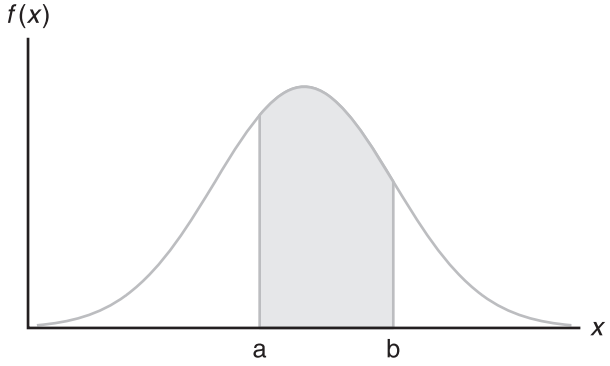


Figura 3.5: $P(a < X < b)$.

Definición 3.6: La función $f(x)$ es una **función de densidad de probabilidad (fdp)** para la variable aleatoria continua X, definida en el conjunto de números reales R, si

1. $f(x) \geq 0$, para toda $x \in R$.

2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

3. $P(a < X < b) = \int_a^b f(x)\, dx$.

Ejemplo 3.11: Suponga que el error en la temperatura de reacción, en °C, para un experimento de laboratorio controlado, es una variable aleatoria continua X, que tiene la función de densidad de probabilidad

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique la condición 2 de la definición 3.6.

b) Encuentre $P(0 < \mathrm{X} \leq 1)$.

Solución: *a*) $\int_{-\infty}^{\infty} f(x)\, dx = \int_{-1}^{2} \frac{x^2}{3} dx = \frac{x^3}{9}|_{-1}^{2} = \frac{8}{9} + \frac{1}{9} = 1$.

b) $P(0 < X \leq 1) = \int_0^1 \frac{x^2}{3} dx = \left.\frac{x^3}{9}\right|_0^1 = \frac{1}{9}$.

Definición 3.7: La **función de distribución acumulada** $F(x)$ de una variable aleatoria continua X con función de densidad $f(x)$ es

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt, \text{ para } -\infty < x < \infty.$$

Como consecuencia inmediata de la definición 3.7 se escriben los dos resultados,

$$P(a < X < b) = F(b) - F(a), \quad \text{y} \qquad f(x) = \frac{dF(x)}{dx},$$

si existe la derivada.

Ejemplo 3.12: Para la función de densidad del ejemplo 3.11 encuentre $F(x)$, y utilícela para evaluar $P(0 < X \leq 1)$.

Solución: Para $-1 < x < 2$,

$$F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-1}^{x} \frac{t^2}{3} dt = \left.\frac{t^3}{9}\right|_{-1}^{x} = \frac{x^3 + 1}{9}.$$

Por lo tanto,

$$F(x) = \begin{cases} 0, & x < -1, \\ \frac{x^3+1}{9}, & -1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

La distribución acumulada $F(x)$ se expresa de forma gráfica en la figura 3.6. Así,

$$P(0 < X \leq 1) = F(1) - F(0) = \frac{2}{9} - \frac{1}{9} = \frac{1}{9},$$

que concuerda con el resultado que se obtuvo al utilizar la función de densidad en el ejemplo 3.11.

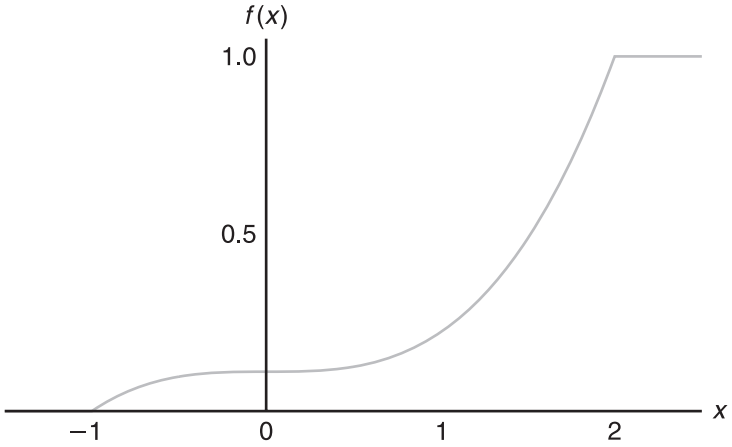


Figura 3.6: Función de distribución acumulada continua.

Ejemplo 3.13: El Departamento de Energía (DE) asigna proyectos mediante licitación y, por lo general, estima lo que debería ser una licitación razonable. Sea b el estimado. El DE determinó que la función de densidad de la licitación ganadora (baja) es

$$f(y) = \begin{cases} \frac{5}{8b}, & \frac{2}{5}b \leq y \leq 2\text{b}, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre $F(y)$ y utilícela para determinar la probabilidad de que la licitación ganadora sea menor que la estimación b preliminar del DE.

Solución: Para $\frac{2}{5}b \leq y \leq 2b$,

$$F(y) = \int_{2b/5}^{y} \frac{5}{8b} dy = \left. \frac{5t}{8b} \right|_{2b/5}^{y} = \frac{5y}{8b} - \frac{1}{4}.$$

De manera que

$$F(y) = \begin{cases} 0, & y < \frac{2}{5}b, \\ \frac{5y}{8b} - \frac{1}{4}, & \frac{2}{5}b \leq y < 2b, \\ 1, & y \geq 2b. \end{cases}$$

Para determinar la probabilidad de que la licitación ganadora sea menor que la estimación b preliminar de la licitación, tenemos

$$P(Y \leq b) = F(b) = \frac{5}{8} - \frac{1}{4} = \frac{3}{8}.$$

Ejercicios

3.1 Clasifique las siguientes variables aleatorias como discretas o continuas:

X: el número de accidentes automovilísticos por año en Virginia.

Y: el tiempo para jugar 18 hoyos de golf.

M: la cantidad de leche que una vaca específica produce anualmente.

N: el número de huevos que una gallina pone mensualmente.

P: el número de permisos para construcción que emiten cada mes en una ciudad.

Q: el peso del grano producido por acre.

3.2 Un embarque foráneo de cinco automóviles extranjeros contiene 2 que tienen ligeras manchas de pintura. Si una agencia recibe 3 de estos automóviles al azar, liste los elementos del espacio muestral S con las letras B y N para "manchado" y "sin mancha", respectivamente; luego a cada punto muestral asigne un valor x de la variable aleatoria X que representa el número de automóviles que la agencia compra con manchas de pintura.

3.3 Sea W la variable aleatoria que da el número de caras menos el número de cruces en tres lanzamientos de una moneda. Liste los elementos del espacio muestral S para los tres lanzamientos de la moneda y asigne un valor w de W a cada punto muestral.

3.4 Se lanza una moneda hasta que ocurren 3 caras sucesivamente. Liste sólo aquellos elementos del espacio muestral que requieren 6 o menos lanzamientos. ¿Es un espacio muestral discreto? Explique.

3.5 Determine el valor c de modo que cada una de las siguientes funciones sirva como distribución de probabilidad de la variable aleatoria discreta X:

a) $f(x) = c(x^2 + 4)$, para $x = 0, 1, 2, 3$;

b) $f(x) = c\binom{2}{x}\binom{3}{3-x}$, para $x = 0, 1, 2$.

3.6 La vida útil, en días, para frascos de cierta medicina de prescripción es una variable aleatoria que tiene la función de densidad

$$f(x) = \begin{cases} \frac{20{,}000}{(x+100)^3}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la probabilidad de que un frasco de esta medicina tenga una vida útil de

a) al menos 200 días;

b) cualquier lapso entre 80 a 120 días.

3.7 El número total de horas, medidas en unidades de 100 horas, que una familia utiliza una aspiradora en un periodo de un año es una variable aleatoria continua X que tiene la función de densidad

$$f(x) = \begin{cases} x, & 0 < x < 1, \\ 2 - x, & 1 \le x < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la probabilidad de que en un periodo de un año, una familia utilice su aspiradora

a) menos de 120 horas;

b) entre 50 y 100 horas.

3.8 Encuentre la distribución de probabilidad de la variable aleatoria W del ejercicio 3.3; suponga que la moneda está cargada de manera que una cara tenga doble de probabilidad de ocurrir que una cruz.

3.9 La proporción de personas que responden a cierta encuesta enviada por correo es una variable aleatoria continua X que tiene la función de densidad

$$f(x) = \begin{cases} \frac{2(x+2)}{5}, & 0 < x < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Muestre que $P(0 < X < 1) = 1$.

b) Encuentre la probabilidad de que más de 1/4 pero menos de 1/2 de las personas contactadas respondan a este tipo de encuesta.

3.10 Encuentre una fórmula para la distribución de probabilidad de la variable aleatoria X que represente el resultado cuando se lanza una vez un solo dado.

3.11 Un embarque de 7 televisores contiene 2 unidades defectuosas. Un hotel realiza una compra azar de 3 de los televisores. Si x es el número unidades defectuosas que compra el hotel, encuentre la distribución de probabilidad de X. Exprese los resultados de forma gráfica como un histograma de probabilidad.

3.12 Una firma de inversiones ofrece a sus clientes bonos municipales que vencen después de varios años. Dado que la función de distribución acumulada de T, el número de años de vencimiento para un bono que se elige al azar, es

$$F(t) = \begin{cases} 0, & t < 1, \\ \frac{1}{4}, & 1 \le t < 3, \\ \frac{1}{2}, & 3 \le t < 5, \\ \frac{3}{4}, & 5 \le t < 7, \\ 1, & t \ge 7. \end{cases}$$

encuentre

a) $P(T = 5)$;

b) $P(T > 3)$;

c) $P(1.4 < T < 6)$.

3.13 La distribución de probabilidad de X, el número de imperfecciones por 10 metros de una tela sintética en rollos continuos de ancho uniforme, está dada por

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

Construya la función de distribución acumulada de X.

3.14 El tiempo de espera, en horas, entre conductores sucesivos que exceden los límites de velocidad detectados por un radar es una variable aleatoria continua con distribución acumulada

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-8x}, & x \geq 0. \end{cases}$$

Encuentre la probabilidad de esperar menos de 12 minutos entre conductores sucesivos que exceden los límites de velocidad

a) usando la función de distribución acumulada de X;

b) utilizando la función de densidad de probabilidad de X.

3.15 Encuentre la función de distribución acumulada de la variable aleatoria X que represente el número de unidades defectuosas en el ejercicio 3.11. Con $F(x)$, encuentre

a) $P(X = 1)$;

b) $P(0 < \mathrm{X} \leq 2)$.

3.16 Construya una gráfica de la función de distribución acumulada del ejercicio 3.15.

3.17 Una variable aleatoria continua X que puede tomar valores entre $x = 1$ y $x = 3$ tiene una función de densidad dada por $f(x) = 1/2$.

a) Muestre que el área bajo la curva es igual a 1.

b) Encuentre $P(2 < X < 2.5)$.

c) Encuentre $P(X \leq 1.6)$.

3.18 Una variable aleatoria continua X que puede tomar valores entre $x = 2$ y $x = 5$ tiene una función de densidad dada por $f(x) = 2(1 + x)/27$. Encuentre

a) $P(X < 4)$;

b) $P(3 \leq X < 4)$.

3.19 Para la función de densidad del ejercicio 3.17, encuentre $F(x)$. Utilícela para evaluar $P(2 < X < 2.5)$.

3.20 Para la función de densidad del ejercicio 3.18, encuentre $F(x)$, y utilícela para evaluar $P(3 \leq X < 4)$.

3.21 Considere la función de densidad

$$f(x) = \begin{cases} k\sqrt{x}, & 0 < x < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Evalúe k.

b) Encuentre $F(x)$ y utilícela para evaluar

$$P(0.3 < X < 0.6).$$

3.22 De una baraja se sacan tres cartas sucesivamente sin reemplazo. Encuentre la distribución de probabilidad para el número de espadas.

3.23 Encuentre la función de distribución acumulada de la variable aleatoria W del ejercicio 3.8. Usando $F(w)$, encuentre

a) $P(W > 0)$;

b) $P(-1 \leq W < 3)$.

3.24 Encuentre la distribución de probabilidad para el número de discos compactos de jazz, cuando se seleccionan cuatro CD al azar de una colección que consiste en cinco de jazz, dos de música clásica y tres de rock. Exprese sus resultados utilizando una fórmula.

3.25 Se seleccionan aleatoriamente 3 monedas sin reemplazo de una caja que contiene 4 de diez centavos y 2 de cinco centavos. Encuentre la distribución de probabilidad para el total T de las tres monedas. Exprese la distribución de probabilidad de forma gráfica como un histograma de probabilidad.

3.26 Se sacan 3 bolas sucesivamente de una caja que contiene 4 bolas negras y 2 verdes; cada bola se regresa a la caja antes de sacar siguiente. Encuentre la distribución de probabilidad para el número de bolas verdes.

3.27 El tiempo de operación antes del fallo, en horas, de una pieza importante de equipo electrónico que se utiliza para la fabricación de un reproductor de DVD tiene la función de densidad

$$f(x) = \begin{cases} \frac{1}{2000} \exp(-x/2000), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

a) Encuentre $F(x)$.

b) Determine la probabilidad de que el componente (y, por lo tanto, el reproductor de DVD) funcionen durante más de 1000 horas antes de que necesite reemplazarse el componente.

c) Determine la probabilidad de que el componente falle antes de 2000 horas.

3.28 Un productor de cereales está consciente de que en la caja el peso del producto varía ligeramente entre una caja y otra. De hecho, datos históricos suficientes han permitido determinar la función de densidad que

describe la estructura de probabilidad para el peso (en onzas). Entonces, si X es el peso, en onzas, de la variable aleatoria, la función de densidad se describe como

$$f(x) = \begin{cases} \frac{2}{5}, & 23.75 \leq x \leq 26.25, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique que sea una función de densidad válida.

b) Determine la probabilidad de que el peso sea menor que 24 onzas.

c) La compañía busca que un peso mayor que 26 onzas sea un caso extraordinariamente raro. ¿Cuál será la probabilidad de que este "caso extraordinariamente raro" en verdad ocurra?

3.29 Un factor importante en el combustible sólido para proyectiles es la distribución del tamaño de las partículas. Ocurren problemas significativos cuando las partículas son demasiado grandes. A partir de datos de producción históricos, se determinó que la distribución del tamaño (en micras) de las partículas se caracteriza por

$$f(x) = \begin{cases} 3x^{-4}, & x > 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique que sea una función de densidad válida.

b) Evalúe $F(x)$.

c) ¿Cuál es la probabilidad de que una partícula tomada al azar del combustible fabricado sea mayor que 4 micras.

3.30 En los sistemas científicos las medidas siempre están sujetas a variación, algunas veces más que otras. Hay muchas estructuras para los errores de medición y los estadísticos pasan una gran cantidad de tiempo modelando tales errores. Suponga que el error de medición X de cierta cantidad física está determinado por la función de densidad

$$f(x) = \begin{cases} k(3 - x^2), & -1 \leq x \leq 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Determine k que representa $f(x)$, una función de densidad válida.

b) Encuentre la probabilidad de que un error aleatorio en la medición sea menor que 1/2.

c) Para esta medida específica, resulta indeseable si la *magnitud* del error (es decir, $|x|$) excede 0.8. ¿Cuál es la probabilidad de que esto ocurra?

3.31 Con base en pruebas de gran alcance, el fabricante de una lavadora determinó que el tiempo Y (en años) antes de que se requiera una reparación mayor se obtiene de la función de densidad de probabilidad

$$f(x) = \begin{cases} \frac{1}{4}e^{-y/4}, & y \geq 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Los críticos, en efecto, considerarían que el producto es una ganga si es improbable que requiera una reparación mayor antes del sexto año. Comente sobre esto determinando $P(Y > 6)$.

b) ¿Cuál es la probabilidad de que ocurra una reparación mayor durante el primer año?

3.32 La proporción del presupuesto para cierta clase de compañía industrial que se asigna a controles ambientales y de contaminación ha estado bajo escrutinio. Un proyecto de recopilación de datos determina que la distribución de tales proporciones está dada por

$$f(y) = \begin{cases} 5(1 - y)^4, & 0 \leq y \leq 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique que la densidad anterior sea válida.

b) ¿Cuál es la probabilidad de que una compañía elegida al azar gaste menos del 10% de su presupuesto en controles ambientales y de contaminación?

c) ¿Cuál es la probabilidad de que una compañía seleccionada al azar gaste más del 50% en controles ambientales y de la contaminación?

3.33 Suponga que un tipo especial de empresa de procesamiento de datos pequeña está tan especializada que algunas tienen dificultades para obtener utilidades durante su primer año de operación. La función de densidad de probabilidad que caracteriza la proporción Y que obtiene utilidades está dada por

$$f(x) = \begin{cases} ky^4(1 - y)^3, & 0 \leq y \leq 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) ¿Cuál es el valor de k que hace de la anterior una función de densidad válida?

b) Encuentre la probabilidad de que al menos el 50% de las empresas tenga utilidades durante el primer año.

c) Encuentre la probabilidad de que al menos 80% de las empresas tenga utilidades durante el primer año.

3.34 Los tubos de magnetrón se producen en una línea de ensamble automatizada. Periódicamente se utiliza un plan de muestreo para evaluar la calidad en la longitud de los tubos; no obstante, dicha medida está sujeta a incertidumbre. Se considera que la probabilidad de que un tubo elegido al azar cumpla con las especificaciones de longitud es 0.99. Se utiliza un plan de muestreo en el cual se mide la longitud de 5 tubos elegidos al azar.

a) Muestre que la función de probabilidad de Y, el tubo de cada 5 que cumple con las especificaciones de longitud, está dado por la siguiente función de probabilidad discreta

$$f(y) = \frac{5!}{y!(5 - y)!}(0.99)^y(0.01)^{5-y},$$

para $y = 0, 1, 2, 3, 4, 5$.

b) Suponga que las selecciones aleatorias se toman de la línea y 3 no cumplen con las especificaciones. Utilice la $f(x)$ anterior ya sea para apoyar o para refutar la conjetura de que la probabilidad de un solo tubo cumpla con las especificaciones es 0.99.

3.35 Suponga que a partir de gran cantidad de datos históricos se sabe que X, el número de automóviles que llegan a una intersección específica durante un periodo de 20 segundos, está determinada por la siguiente función de probabilidad discreta

$$f(x) = e^{-6}\frac{6^x}{x!}, \quad x = 0, 1, 2, \ldots.$$

a) Encuentre la probabilidad de que en un periodo específico de 20 segundos, más de 8 automóviles lleguen a la intersección.

b) Encuentre la probabilidad de que sólo lleguen 2.

3.36 En una tarea de laboratorio, cuando el equipo está operando la función de densidad del resultado observado, X, es

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Calcule $P(X \leq 1/3)$.

b) ¿Cuál es la probabilidad de que X excederá 0.5?

c) Dado que $X \geq 0.5$, ¿cuál es la probabilidad de que X será menor que 0.75?

3.4 Distribuciones de probabilidad conjunta

En las secciones anteriores nuestro estudio de variables aleatorias y sus distribuciones de probabilidad se restringió a espacios muestrales unidimensionales, donde registramos los resultados de un experimento como los valores que toma una sola variable aleatoria. No obstante, habrá situaciones en que encontraremos deseable registrar los resultados simultáneos de diversas variables aleatorias. Por ejemplo, podríamos medir la cantidad de precipitado P y volumen V de gas liberado en un experimento químico controlado, que dan lugar a un espacio muestral bidimensional que consiste en los resultados (p, v); o podríamos interesarnos en la dureza H y en la resistencia a la tensión T de cobre estirado en frío que conducen a los resultados (h, t). En un estudio para determinar la probabilidad de éxito en la universidad, que se basa en los datos del nivel preparatoria, se puede utilizar un espacio muestral tridimensional y registrar, para cada individuo, su calificación de la prueba de aptitudes, su clasificación de clase en preparatoria y el promedio en puntos al final del primer año en la universidad.

Si X y Y son dos variables aleatorias discretas, la distribución de probabilidad para sus ocurrencias simultáneas se representa mediante una función con valores $f(x, y)$, para cualquier par de valores (x, y) dentro del rango de las variables aleatorias X y Y. Se acostumbra referirse a esta función como la **distribución de probabilidad conjunta** de X y Y.

De aquí, en el caso discreto,

$$f(x, y) = P(X = x, Y = y);$$

es decir, los valores $f(x, y)$ dan la probabilidad de que ocurran al mismo tiempo los resultados x y y. Por ejemplo, si se le va a dar servicio a un televisor, y X representa la edad de la unidad al año más cercano y Y representa el número de bulbos defectuosos en el televisor, entonces $f(5, 3)$ es la probabilidad de que el televisor tenga 5 años de edad y necesite 3 bulbos nuevos.

Definición 3.8: La función $f(x,y)$ es una **distribución de probabilidad conjunta** o **función de masa de probabilidad** de las variables aleatorias discretas X y Y, si

1. $f(x,y) \geq 0$ para toda (x,y),
2. $\sum_x \sum_y f(x,y) = 1$,
3. $P(X=x, Y=y) = f(x,y)$.

Para cualquier región A en el plano xy, $P[(X,Y) \in A] = \sum\sum_A f(x,y)$.

Ejemplo 3.14: Se seleccionan al azar 2 repuestos para un bolígrafo de una caja que contiene 3 repuestos azules, 2 rojos y 3 verdes. Si X es el número de repuestos azules y Y es el número de repuestos rojos seleccionados, encuentre

a) la función de probabilidad conjunta $f(x, y)$,

b) $P[(X, Y) \in A]$, donde A es la región $\{(x, y) | x + y \leq 1\}$.

Solución: *a*) Los posibles pares de valores (x, y) son (0, 0), (0, 1), (1, 0), (1, 1), (0, 2) y (2, 0). Así, $f(0,1)$, por ejemplo, representa la probabilidad de que se seleccionen un repuesto rojo y uno verde. El número total de formas igualmente probables de seleccionar cualesquiera 2 repuestos de los 8 es $\binom{8}{2} = 28$. El número de formas de seleccionar 1 rojo de 2 repuestos rojos y 1 verde de 3 repuestos verdes es $\binom{2}{1}\binom{3}{1} = 6$. De aquí, $f(0, 1) = 6/28 = 3/14$. Cálculos similares dan las probabilidades para los otros casos, que se presentan en la tabla 3.1. Observa que las probabilidades suman 1. En el capítulo 4 quedará claro que la distribución de probabilidad conjunta de la tabla 3.1 se puede representar con la fórmula

$$f(x,y) = \frac{\binom{3}{x}\binom{2}{y}\binom{3}{2-x-y}}{\binom{8}{2}},$$

para $x = 0, 1, 2$; $y = 0, 1, 2$; y $0 \leq x + y \leq 2$.

b)
$$\begin{aligned} P[(X,Y) \in A] &= P(X+Y \leq 1) = f(0,0) + f(0,1) + f(1,0) \\ &= \frac{3}{28} + \frac{3}{14} + \frac{9}{28} = \frac{9}{14}. \end{aligned}$$

Tabla 3.1: Distribución de probabilidad conjunta para el ejemplo 3.14

$f(x,y)$		x = 0	x = 1	x = 2	Totales por renglón
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Totales por columna		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Cuando X y Y son variables aleatorias continuas, la **función de densidad conjunta** $f(x, y)$ es una superficie sobre el plano xy, y $P[(X, Y) \in A]$, donde A es

cualquier región en el plano xy, es igual al volumen del cilindro recto limitado por la base A y la superficie.

Definición 3.9: La función $f(x, y)$ es una **función de densidad conjunta** de las variables aleatorias continuas X y Y si

1. $f(x, y) \geq 0$, para toda (x, y),
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$,
3. $P[(X, Y) \in A] = \int \int_A f(x, y)\, dx\, dy$,

para cualquier región A en el plano xy.

Ejemplo 3.15: Una fábrica de dulces distribuye cajas de chocolates con un surtido de cremas, chiclosos y nueces cubiertas con chocolate claro y oscuro. Para una caja seleccionada al azar, sean X y Y, respectivamente, las proporciones de chocolates claro y oscuro que son cremas y suponga que la función de densidad conjunta es

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique la condición 2 de la definición 3.9.

b) Encuentre $P[(X, Y \in = A]$, donde $A = \{(x, y) | 0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}\}$.

Solución: *a*)

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy &= \int_0^1 \int_0^1 \frac{2}{5}(2x + 3y)\, dx\, dy \\ &= \int_0^1 \left(\frac{2x^2}{5} + \frac{6xy}{5} \right) \Big|_{x=0}^{x=1} dy \\ &= \int_0^1 \left(\frac{2}{5} + \frac{6y}{5} \right) dy = \left(\frac{2y}{5} + \frac{3y^2}{5} \right) \Big|_0^1 = \frac{2}{5} + \frac{3}{5} = 1. \end{aligned}$$

b)

$$\begin{aligned} P[(X, Y) \in A] &= P(0 < X < \frac{1}{2}, \frac{1}{4} < Y < \frac{1}{2}) \\ &= \int_{1/4}^{1/2} \int_0^{1/2} \frac{2}{5}(2x + 3y)\, dx\, dy = \int_{1/4}^{1/2} \left(\frac{2x^2}{5} + \frac{6xy}{5} \right) \Big|_{x=0}^{x=1/2} dy \\ &= \int_{1/4}^{1/2} \left(\frac{1}{10} + \frac{3y}{5} \right) dy = \left(\frac{y}{10} + \frac{3y^2}{10} \right) \Big|_{1/4}^{1/2} \\ &= \frac{1}{10} \left[\left(\frac{1}{2} + \frac{3}{4} \right) - \left(\frac{1}{4} + \frac{3}{16} \right) \right] = \frac{13}{160}. \end{aligned}$$

Dada la distribución de probabilidad conjunta $f(x, y)$ de las variables aleatorias discretas X y Y, la distribución de probabilidad $g(x)$ de X sola se obtiene al sumar $f(x, y)$ sobre los valores de Y. De manera similar, la distribución de probabilidad $h(y)$ de Y sola se obtiene al sumar $f(x, y)$ sobre los valores de X. Definimos $g(x)$ y $h(y)$ como **distribuciones marginales** de X y Y, respectivamente. Cuando X y Y son variables aleatorias continuas, las sumatorias se reemplazan por integrales. Ahora podemos establecer la siguiente definición general.

Definición 3.10: Las **distribuciones marginales** de X sola y de Y sola son

$$g(x) = \sum_y f(x,y) \quad \text{y} \quad h(y) = \sum_x f(x,y),$$

Para el caso discreto, y

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \quad \text{y} \quad h(y) = \int_{-\infty}^{\infty} f(x,y)\, dx,$$

para el caso continuo.

El término *marginal* se utiliza aquí porque, en el caso discreto, los valores de $g(x)$ y $h(y)$ son precisamente los totales marginales de las columnas y los renglones respectivos, cuando los valores de $f(x, y)$ se muestran en una tabla rectangular.

Ejemplo 3.16: Muestre que los totales de columnas y renglones de la tabla 3.1 dan las distribuciones marginales de X sola y Y sola.

Solución: Para la variable aleatoria X, vemos que

$$g(0) = f(0,0) + f(0,1) + f(0,2) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14},$$
$$g(1) = f(1,0) + f(1,1) + f(1,2) = \frac{9}{28} + \frac{3}{14} + 0 = \frac{15}{28},$$

y

$$g(2) = f(2,0) + f(2,1) + f(2,2) = \frac{3}{28} + 0 + 0 = \frac{3}{28},$$

que son precisamente los totales por columna de la tabla 3.1. De manera similar podemos mostrar que los valores de $h(y)$ están dados por los totales de los renglones. En forma tabular, estas distribuciones marginales se pueden escribir como sigue:

x	0	1	2
$g(x)$	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$

y	0	1	2
$h(y)$	$\frac{15}{28}$	$\frac{3}{7}$	$\frac{1}{28}$

Ejemplo 3.17: Encuentre $g(x)$ y $h(y)$ para la función de densidad conjunta del ejemplo 3.15.

Solución: Por definición,

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_0^1 \frac{2}{5}(2x+3y)\, dy = \left(\frac{4xy}{5} + \frac{6y^2}{10}\right)\Big|_{y=0}^{y=1} = \frac{4x+3}{5},$$

para $0 \le x \le 1$, y $g(x) = 0$ en cualquier otro caso. De manera similar,

$$h(y) = \int_{-\infty}^{\infty} f(x,y)\, dx = \int_0^1 \frac{2}{5}(2x+3y)\, dx = \frac{2(1+3y)}{5},$$

para $0 \le y \le 1$, y $h(y) = 0$ en cualquier otro caso.

El hecho de que las distribuciones marginales $g(x)$ y $h(y)$ sean en realidad las distribuciones de probabilidad de las variables individuales X y Y solas se puede

verificar al mostrar que se satisfacen las condiciones de la definición 3.4 o de la definición 3.6. Por ejemplo, en el caso continuo

$$\int_{-\infty}^{\infty} g(x)\ dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ dy\ dx = 1,$$

y

$$\begin{aligned} P(a < X < b) &= P(a < X < b, -\infty < Y < \infty) \\ &= \int_a^b \int_{-\infty}^{\infty} f(x,y)\ dy\ dx = \int_a^b g(x)\ dx. \end{aligned}$$

En la sección 3.1 establecimos que el valor x de la variable aleatoria X representa un evento que es un subconjunto del espacio muestral. Si utilizamos la definición de probabilidad condicional que se estableció en el capítulo 2,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0,$$

donde A y B son ahora los eventos definidos por $X = x$ y $Y = y$, respectivamente, entonces,

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f(x,y)}{g(x)}, \quad g(x) > 0,$$

donde X y Y son variables aleatorias discretas.

No es difícil mostrar que la función $f(x, y)/g(x)$, que es estrictamente una función de y con x fija, satisface todas las condiciones de una distribución de probabilidad. Esto también es cierto cuando $f(x, y)$ y $g(x)$ son la densidad conjunta y la distribución marginal, respectivamente, de variables aleatorias continuas. Como resultado es muy importante que utilicemos el tipo especial de distribución de la forma $f(x, y)/g(x)$ con la finalidad de ser capaces de calcular probabilidades condicionales de manera eficaz. Este tipo de distribución se llama **distribución de probabilidad condicional**; la definición condicional es la siguiente.

Definición 3.11: Sean X y Y dos variables aleatorias, discretas o continuas. La **distribución condicional** de la variable aleatoria Y, dado que $X = x$, es

$$f(y|x) = \frac{f(x,y)}{g(x)}, \quad g(x) > 0.$$

De manera similar, la distribución condicional de la vaiable aleatoria X, dado que $Y = y$, es

$$f(x|y) = \frac{f(x,y)}{h(y)}, \quad h(y) > 0.$$

Si deseamos encontrar la probabilidad de que la variable aleatoria discreta X caiga entre a y b cuando se sabe que la variable discreta $Y = y$, evaluamos

$$P(a < X < b|Y = y) = \sum_{a<x<b} f(x|y),$$

donde la sumatoria se extiende a todos los valores de X entre a y b. Cuando X y Y son continuas, evaluamos

$$P(a < X < b|Y = y) = \int_a^b f(x|y)\ dx.$$

Ejemplo 3.18: Con referencia al ejemplo 3.14, encuentre la distribución condicional de X, dado que $Y = 1$, y utilícela para determinar $P(X = 0|Y = 1)$.

Solución: Necesitamos encontrar $f(x|y)$, donde $y = 1$. Primero, encontramos que

$$h(1) = \sum_{x=0}^{2} f(x, 1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}.$$

Ahora,

$$f(x|1) = \frac{f(x, 1)}{h(1)} = \frac{7}{3} f(x, 1), \quad x = 0, 1, 2.$$

Por lo tanto,

$$f(0|1) = \left(\frac{7}{3}\right) f(0, 1) = \left(\frac{7}{3}\right)\left(\frac{3}{14}\right) = \frac{1}{2},$$
$$f(1|1) = \left(\frac{7}{3}\right) f(1, 1) = \left(\frac{7}{3}\right)\left(\frac{3}{14}\right) = \frac{1}{2},$$
$$f(2|1) = \left(\frac{7}{3}\right) f(2, 1) = \left(\frac{7}{3}\right)(0) = 0,$$

y la distribucion condicional de X, dado que $Y = 1$, es

x	0	1	2
$f(x\|1)$	$\frac{1}{2}$	$\frac{1}{2}$	0

Finalmente,

$$P(X = 0|Y = 1) = f(0|1) = \frac{1}{2}.$$

Por lo tanto, si se sabe que 1 de los 2 repuestos seleccionados es rojo, tenemos una probabilidad igual a 1/2 de que el otro repuesto no sea azul.

Ejemplo 3.19: La densidad conjunta para las variables aleatorias (X, Y), donde X es el cambio de temperatura unitario y Y es la proporción de desplazamiento espectral que produce cierta partícula atómica es

$$f(x, y) = \begin{cases} 10xy^2, & 0 < x < y < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre las densidades marginales $g(x)$, $h(y)$ y la densidad condicional $f(y|x)$.

b) Encuentre la probabilidad de que el espectro se desplace más de la mitad de las observaciones totales, dado que la temperatura aumenta a 0.25 unidades.

Solución: *a*) Por definición,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{x}^{1} 10xy^2\, dy$$
$$= \frac{10}{3} xy^3 \Big|_{y=x}^{y=1} = \frac{10}{3} x(1 - x^3),\ 0 < x < 1,$$
$$h(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_{0}^{y} 10xy^2\, dx = 5x^2y^2\big|_{x=0}^{x=y} = 5y^4,\ 0 < y < 1.$$

Entonces,

$$f(y|x) = \frac{f(x,y)}{g(x)} = \frac{10xy^2}{\frac{10}{3}x(1-x^3)} = \frac{3y^2}{1-x^3},\ 0 < x < y < 1.$$

b) Por lo tanto,

$$P\left(Y > \frac{1}{2}\middle| X = 0.25\right) = \int_{1/2}^{1} f(y|x = 0.25)dy$$
$$= \int_{1/2}^{1} \frac{3y^2}{1-0.25^3}\ dy = \frac{8}{9}.$$

Ejemplo 3.20: Dada la función de densidad conjunta

$$f(x,y) = \begin{cases} \frac{x(1+3y^2)}{4}, & 0 < x < 2,\ 0 < y < 1, \\ 0, & \text{en cualquier otro caso} \end{cases}$$

encuentre $g(x)$, $h(y)$, $f(x\,|\,y)$, y evalúe $P(\frac{1}{4} < X < \frac{1}{2}|Y = \frac{1}{3})$.

Solución: Por definición,

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\ dy = \int_{0}^{1} \frac{x(1+3y^2)}{4}dy$$
$$= \left(\frac{xy}{4} + \frac{xy^3}{4}\right)\Big|_{y=0}^{y=1} = \frac{x}{2}, \quad 0 < x < 2,$$

y

$$h(y) = \int_{-\infty}^{\infty} f(x,y)\ dx = \int_{0}^{2} \frac{x(1+3y^2)}{4}dx$$
$$= \left(\frac{x^2}{8} + \frac{3x^2y^2}{8}\right)\Big|_{x=0}^{x=2} = \frac{1+3y^2}{2}, \quad 0 < y < 1.$$

Por lo tanto,

$$f(x|y) = \frac{f(x,y)}{h(y)} = \frac{x(1+3y^2)/4}{(1+3y^2)/2} = \frac{x}{2}, \quad 0 < x < 2,$$

y

$$P\left(\frac{1}{4} < X < \frac{1}{2}\middle| Y = \frac{1}{3}\right) = \int_{1/4}^{1/2} \frac{x}{2}dx = \frac{3}{64}.$$

Independencia estadística

Si $f(x\,|\,y)$ no depende de y, como en el caso del ejemplo 3.20, entonces $f(x\,|\,y) = g(x)$ y $f(x,\, y) = g(x)h(y)$. La demostración se tiene al sustituir

$$f(x,y) = f(x|y)h(y)$$

en la distribución marginal de X. Es decir,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\ dy = \int_{-\infty}^{\infty} f(x|y)h(y)\ dy.$$

Si $f\ (x|y)$ no depende de y, escribimos

$$g(x) = f(x|y) \int_{-\infty}^{\infty} h(y)\ dy.$$

Entonces,

$$\int_{-\infty}^{\infty} h(y)\ dy = 1,$$

ya que $h(y)$ es la función de densidad de probabilidad de Y. Por lo tanto,

$$g(x) = f(x\,|\,y) \text{ y, entonces, } f(x,\ y) = g(x)h(y).$$

Debería tener sentido para el lector que si $f(x\,|\,y)$ no depende de y, entonces, por supuesto, el resultado de la variable aleatoria Y no tiene impacto en el resultado de la variable aleatoria X. En otras palabras, decimos que X y Y son variables aleatorias independientes. Ofrecemos ahora la siguiente definición formal de independencia estadística.

Definición 3.12: Sean X y Y dos variables aleatoria, discretas o continuas, con distribución de probabilidad conjunta $f(x,\ y)$ y distribuciones marginales $g(x)$ y $h(y)$, respectivamente. Se dice que las variables aleatorias X y Y son **estadísticamente independiente** si y solo si

$$f(x,\ y) = g(x)h(y)$$

para toda $(x,\ y)$ dentro de sus rangos.

Las variables aleatorias continuas del ejemplo 3.20 son estadísticamente independientes, pues el producto de las dos distribuciones marginales da la función de densidad conjunta. Evidentemente éste no es el caso; sin embargo, para las variables continuas del ejemplo 3.19. La comprobación de la independencia estadística de variables aleatorias discretas requiere una investigación más profunda, ya que es posible que el producto de las distribuciones marginales sea igual a la distribución de probabilidad conjunta para algunas —aunque no para todas— de las combinaciones de $(x,\ y)$. Si puede encontrar algún punto $(x,\ y)$ para el que $f(x,\ y)$ se define de manera que $f(x,\ y) \neq g(x)h(y)$, las variables discretas X y Y no son estadísticamente independientes.

Ejemplo 3.21: Muestre que las variables aleatorias del ejemplo 3.14 no son estadísticamente independientes.

Prueba: Consideremos el punto (0, 1). De la tabla 3.1 encontramos que las tres probabilida-

des $f(0, 1)$, $g(0)$ y $h(1)$ son

$$f(0,1) = \frac{3}{14},$$
$$g(0) = \sum_{y=0}^{2} f(0,y) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14},$$
$$h(1) = \sum_{x=0}^{2} f(x,1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}.$$

Claramente,

$$f(0,\ 1) \neq g(0)h(1),$$

y, por lo tanto, X y Y no son estadísticamente independientes.

Todas las definiciones anteriores respecto a dos variables aleatorias se pueden generalizar al caso de n variables aleatorias. Sea $f(x_1,\ x_2, \ldots, x_n)$ la función de probabilidad conjunta de las variables aleatorias $X_1,\ X_2, \ldots, X_n$. La distribución marginal de X_1, por ejemplo, es

$$g(x_1) = \sum_{x_2} \cdots \sum_{x_n} f(x_1, x_2, \ldots, x_n)$$

para el caso discreto, y

$$g(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\ dx_2\ dx_3 \cdots dx_n$$

para el caso continuo. Ahora obtenemos **distribuciones marginales conjuntas** como $g(x_1,\ x_2)$, donde

$$g(x_1, x_2) = \begin{cases} \sum_{x_3} \cdots \sum_{x_n} f(x_1, x_2, \ldots, x_n), & \text{(caso discreto)}, \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\ dx_3\ dx_4 \cdots dx_n, & \text{(caso continuo)}. \end{cases}$$

Podemos considerar numerosas distribuciones condicionales. Por ejemplo, la **distribución condicional conjunta** de X_1, X_2 y X_3, dado que $X_4 = x_4$, $X_5 = x_5, \ldots,$ $X_n = x_n$, se escribe como

$$f(x_1, x_2, x_3 | x_4, x_5, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n)}{g(x_4, x_5, \ldots, x_n)},$$

donde $g(x_4,\ x_5, \ldots, x_n)$ es la distribución marginal conjunta de las variables aleatorias $X_4,\ X_5, \ldots, X_n$.

Una generalización de la definición 3.12 nos lleva a la siguiente definición para la independencia estadística mutua de las variables $X_1,\ X_2, \ldots, X_n$.

Definición 3.13: Sean $X_1, X_2, \ldots, X_n$ n variables aleatorias, discretas o continuas, con distribución de probabilidad conjunta $f(x_1, x_2, \ldots, x_n)$ y distribuciones marginales $f_1(x_1)$, $f_2(x_2), \ldots, f_n(x_n)$, respectivamente. Se dice que las variables aleatorias $X_1, X_2, \ldots, X_n$ son **estadísticamente independientes** mutuamente si y sólo si

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n)$$

para toda $(x_1, x_2, \ldots, x_n)$ dentro de sus rangos.

Ejemplo 3.22: Suponga que el tiempo de vida en anaquel, en años, de cierto producto alimenticio perecedero empacado en cajas de cartón es una variable aleatoria cuya función de densidad de probabilidad está dada por

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Sean X_1, X_2, y X_3 los tiempos de vida en anaquel para tres de estas cajas seleccionadas de forma independiente y encuentre $P(X_1 < 2, 1 < X_2 < 3, X_3 > 2)$.

Solución: Como las cajas se seleccionan de forma independiente, suponemos que las variables aleatorias X_1, X_2 y X_3 son estadísticamente independientes y que tienen la densidad de probabilidad conjunta

$$f(x_1, x_2, x_3) = f(x_1)f(x_2)f(x_3) = e^{-x_1}e^{-x_2}e^{-x_3} = e^{-x_1-x_2-x_3},$$

para $x_1 > 0$, $x_2 > 0$, $x_3 > 0$, y $f(x_1, x_2, x_3) = 0$ en cualquier otro caso. De aquí

$$\begin{aligned} P(X_1 < 2, 1 < X_2 < 3, X_3 > 2) &= \int_2^{\infty}\int_1^3\int_0^2 e^{-x_1-x_2-x_3}\, dx_1\, dx_2\, dx_3 \\ &= (1 - e^{-2})(e^{-1} - e^{-3})e^{-2} = 0.0372. \end{aligned}$$

¿Por qué son importantes las características de las distribuciones de probabilidad y de dónde vienen?

En este texto es un punto importante ofrecer al lector una transición hacia los siguientes tres capítulos. En los ejemplos y los ejercicios hemos trabajados casos de situaciones prácticas de ingeniería y ciencias, en las cuales las distribuciones de probabilidad y sus propiedades se utilizan para resolver problemas importantes. Tales distribuciones de probabilidad, ya sean discretas o continuas, se presentaron mediante frases como "se sabe que", "suponga que" o, incluso en ciertos casos, "la evidencia histórica sugiere que". Se trata de situaciones en las que la naturaleza de la distribución e incluso una estimación óptima de la estructura de la probabilidad se pueden determinar utilizando datos históricos, datos tomados de estudios a largo plazo o hasta de grandes cantidades de datos planeados. El lector debería tener presente la discusión del uso de histogramas del capítulo 1 y, por consiguiente, recordar la manera en que las distribuciones de frecuencias se estiman a partir de histogramas. Sin embargo, no todas las funciones de probabilidad y de densidad de probabilidad se derivan de cantidades grandes de datos históricos. Hay un número significativo de situaciones en las cuales la naturaleza del escenario científico sugiere un tipo de

distribución. De hecho, varias de ellas se reflejan en los ejercicios de los capítulos 2 y 3. Cuando observaciones repetidas independientes son binarias por naturaleza (es decir, "defectuoso o no", "funciona o no", "alérgico o no") con observaciones de 0 o 1, la distribución que cubre esta situación se llama **distribución binomial** y la función de probabilidad se conoce y se demostrará en sus principios en el capítulo 5. El ejercicio 3.34 de la sección 3.3 y el ejercicio 3.82 constituyen ejemplos y son de otro tipo que el lector también debería reconocer. El escenario de una distribución continua en "tiempo de operación antes del fallo", como en el ejercicio de repaso 3.71 o en el ejercicio 3.27 de la página 89, a menudo sugiere una clase de distribución denominada **distribución exponencial**. Tales tipos de ejemplos son tan sólo dos de los muchos llamados de distribuciones estándar que se utilizan ampliamente en situaciones del mundo real, ya que el escenario científico que ocasiona cada uno de ellos es reconocible y a menudo sucede en la práctica. Los capítulos 5 y 6 cubren muchos de tales tipos junto con alguna teoría inherente respecto de su uso.

La segunda parte de esta trancisión al material de los futuros capítulos tiene que ver con la noción de **parámetros de la población** o **parámetros distributivos**. Recuerde que en el capítulo 1 vimos la necesidad de utilizar datos para ofrecer información sobre dichos parámetros. Nos concentramos en estudiar las nociones de **media** y de **varianza**, y aprendimos sobre éstas en el contexto de una población. De hecho, la media y la varianza de la población son fáciles de encontrar a partir de la función de probabilidad para el caso discreto, o de la función de densidad de probabilidad para el caso continuo. Tales parámetros y su importancia en la solución de muchas clases de problemas de la vida real nos proporcionarán mucho del material de los capítulos 8 a 17.

Ejercicios

3.37 Determine el valor de c tal que las siguientes funciones representen distribuciones de probabilidad conjunta de las variables aleatorias X y Y:

a) $f(x, y) = cxy$, para $x = 1, 2, 3$; y $= 1, 2, 3$;

b) $f(x, y) = c|x - y|$, para $x = -2, 0, 2$; $y = -2, 3$.

3.38 Si la distribución de probabilidad conjunta de X y Y está dada por

$$f(x,y) = \frac{x+y}{30}, \quad \text{para } x = 0, 1, 2, 3;\ y = 0, 1, 2,$$

encuentre

a) $P(X \leq 2, Y = 1)$;

b) $P(X > 2, Y \leq 1)$;

c) $P(X > Y)$;

d) $P(X + Y = 4)$.

3.39 De un saco de frutas que contiene 3 naranjas, 2 manzanas y 3 plátanos se selecciona una muestra aleatoria de 4 frutas. Si X es el número de naranjas y Y el de manzanas en la muestra, encuentre

a) la distribución de probabilidad conjunta de X y Y;

b) $P[(X, Y) \in A]$, donde A es la región dada por $\{(x, y) \mid x + y \leq 2\}$.

3.40 Una vinatería de un particular opera instalaciones para atención en el automóvil y para atender a quien llega caminando. En un día seleccionado al azar, sean X y Y, respectivamente, las proporciones del tiempo que se utiliza cada instalación, y suponga que la función de densidad conjunta de estas variables aleatorias es

$$f(x,y) = \begin{cases} \frac{2}{3}(x + 2y), & 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre la densidad marginal de X.

b) Encuentre la densidad marginal de Y.

c) Encuentre la probabilidad de que las instalaciones para atención en en el automóvil estén ocupadas menos de la mitad del tiempo.

3.41 Una compañía dulcera distribuye cajas de chocolates con un surtido de cremas, chiclosos y envinados. Suponga que el peso de cada caja es 1 kilogramo; pero que los pesos individuales de cremas, chiclosos y envinados varían de una caja a otra. Para una caja se-

leccionada al azar, sean X y Y los pesos de las cremas y los chiclosos, respectivamente, y suponga que la función de densidad conjunta de estas variables es

$$f(x,y)=\begin{cases}24xy, & 0\le x\le 1,\ 0\le y\le 1,\\ & x+y\le 1,\\ 0, & \text{en cualquier otro caso.}\end{cases}$$

a) Encuentre la probabilidad de que en una caja dada los envinados representen más de la mitad del peso.

b) Encuentre la densidad marginal para el peso de las cremas.

c) Encuentre la probabilidad de que el peso de los chiclosos en una caja sea menor que 1/8 de kilogramo, si se sabe que las cremas constituyen 3/4 del peso.

3.42 Sean X y Y la duración de la vida, en años, de dos componentes en un sistema electrónico. Si la función de densidad conjunta de estas variables es

$$f(x,y)=\begin{cases}e^{-(x+y)}, & x>0,\ y>0,\\ 0, & \text{en cualquier otro caso.}\end{cases}$$

Encuentre $P(0<X<1\mid Y=2)$.

3.43 Sea X el tiempo de reacción, en segundos, a cierto estimulante, y Y la temperatura (°F) a la cual cierta reacción comienza a suceder. Suponga que dos variales aleatorias X y Y tienen la densidad conjunta

$$f(x,y)=\begin{cases}4xy, & 0<x<1,\ 0<y<1,\\ 0, & \text{en cualquier otro caso.}\end{cases}$$

Encuentre

a) $P(0\le X\le \frac{1}{2}$ y $\frac{1}{4}\le Y\le\frac{1}{2})$;

b) $P(X<Y)$.

3.44 Se supone que cada rueda trasera de un avión experimental se llena a una presión de 40 libras por pulgada cuadrada (psi). Sea X la presión real del aire para la rueda derecha y Y la presión real del aire de la rueda izquierda. Suponga que X y Y son variables aleatorias con la densidad conjunta

$$f(x,y)=\begin{cases}k(x^2+y^2), & 30\le x<50;\\ & 30\le y<50,\\ 0, & \text{en cualquier otro caso.}\end{cases}$$

a) Encuentre k.

b) Encuentre $P(30\le X\le 40$ y $40\le Y<50)$.

c) Encuentre la probabilidad de que ambas ruedas o estén insuficientemente llenas.

3.45 Sea X el diámetro de un cable eléctrico blindado y Y el diámetro del molde cerámico que hace el cable. Tanto X como Y tienen una escala tal que están entre 0 y 1. Suponga que X y Y tienen la densidad conjunta

$$f(x,y)=\begin{cases}\frac{1}{y}, & 0<x<y<1,\\ 0, & \text{en cualquier otro caso.}\end{cases}$$

Encuentre $P(X+Y>1/2)$.

3.46 Con referencia al ejercicio 3.38, encuentre

a) la distribución marginal de X;

b) la distribución marginal de Y.

3.47 La cantidad de queroseno, en miles de litros, en un tanque al principio de cualquier día es una cantidad aleatoria Y, de la que una cantidad aleatoria X se vende durante el día. Suponga que el tanque no se abastece durante el día, por lo que $x\le y$, y suponga que la función de densidad conjunta de estas variables es

$$f(x,y)=\begin{cases}2, & 0<x<y<1,\\ 0, & \text{en cualquier otro caso.}\end{cases}$$

a) Determine si X y Y son independientes.

b) Encuentre $P(1/4<X<1/2\mid Y=3/4)$.

3.48 Refiérase al ejercicio 3.39 y encuentre

a) $f(y\mid 2)$ para todos los valores de y;

b) $P(Y=0\mid X=2)$.

3.49 Sea X el número de veces que fallará cierta máquina de control numérico: l, 2 0 3 veces en un día dado. Sea Y el número de veces que se llama a un técnico para una emergencia. Su distribución de probabilidad conjunta está dada como

	$f(x,y)$	x = 1	2	3
	1	0.05	0.05	0.1
y	2	0.05	0.1	0.35
	3	0	0.2	0.1

a) Evalúe la distribución marginal de X.

b) Evalúe la distribución marginal de Y.

c) Encuentre $P(Y=3\mid X=2)$.

3.50 Suponga que X y Y tienen la siguiente distribución de probabilidad conjunta:

	$f(x,y)$	x = 2	4
	1	0.10	0.15
y	3	0.20	0.30
	5	0.10	0.15

a) Encuentre la distribución marginal de X.

b) Encuentre la distribución marginal de Y.

3.51 Considere un experimento que consiste en 2 lanzamientos de un dado balanceado. Si X es el número de "cuatros" y Y es el número de "cincos" que se obtienen en los 2 lanzamientos del dado, encuentre

a) la distribución de probabilidad conjunta de X y Y;

b) $P[(X, Y) \in A]$, donde A es la región $\{(x, y) \mid 2x + y < 3\}$.

3.52 Sea X el número de caras y Y el número de caras menos el número de cruces cuando se lanzan 3 monedas. Encuentre la distribución de probabilidad conjunta de X y Y.

3.53 Se sacan tres cartas sin reemplazo de las 12 cartas mayores (jacks, reinas y reyes) de una baraja ordinaria de 52 cartas. Sea X el número de reyes que se seleccionan y Y el número de jacks. Encuentre

a) la distribución de probabilidad conjunta de X y Y;

b) $P[(X, Y) \in A]$, donde A es la región dada por $\{(x, y) \mid x + y \geq 2\}$.

3.54 Se lanza dos veces una moneda. Sea Z el número de caras en el primer lanzamiento y W el número total de caras en los 2 lanzamientos. Si la moneda no está balanceada y una cara tiene una probabilidad de ocurrencia de 40%, encuentre

a) la distribución de probabilidad conjunta de W y Z;

b) la distribución marginal de W;

c) la distribución marginal de Z;

d) la probabilidad de que ocurra al menos 1 cara.

3.55 Dada la función de densidad conjunta

$$f(x,y) = \begin{cases} \frac{6-x-y}{8}, & 0 < x < 2,\ 2 < y < 4, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre $P(1 < Y < 3 \mid X = 1)$.

3.56 Determine si las dos variables aleatorias del ejercicio 3.49 son dependientes o independientes.

3.57 Determine si las dos variables aleatorias del ejercicio 3.50 son dependientes o independientes.

3.58 La función de densidad conjunta de las variables aleatorias X y Y es

$$f(x,y) = \begin{cases} 6x, & 0 < x < 1,\ 0 < y < 1 - x, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Muestre que X y Y no son independientes.

b) Encuentre $P(X > 0.3 \mid Y = 0.5)$.

3.59 Si X, Y y Z tienen la función de densidad de probabilidad conjunta

$$f(x,y,z) = \begin{cases} kxy^2z, & 0 < x, y < 1;\ 0 < z < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre k.

b) Encuentre $P(X < \frac{1}{4}, Y > \frac{1}{2}, 1 < Z < 2)$.

3.60 Determine si las dos variables aleatorias del ejercicio 3.43 son dependientes o independientes.

3.61 Determine si las dos variables aleatorias del ejercicio 3.44 son dependientes o independientes.

3.62 La función de densidad de probabilidad conjunta de las variables aleatorias X, Y y Z es

$$f(x,y,z) = \begin{cases} \frac{4xyz^2}{9}, & 0 < x, y < 1;\ 0 < z < 3, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre

a) la función de densidad marginal conjunta de Y y Z;

b) la densidad marginal de Y;

c) $P(\frac{1}{4} < X < \frac{1}{2},\ Y > \frac{1}{3},\ 1 < Z < 2)$;

d) $P(0 < X < \frac{1}{2} \mid Y = \frac{1}{4},\ Z = 2)$.

Ejercicios de repaso

3.63 Una compañía tabacalera produce mezclas de tabaco, y cada mezcla contiene varias proporciones de tabaco turco, tabaco de la región y otros. Las proporciones de turco y de la región en una mezcla son variables aleatorias con función de densidad conjunta (X = turco y Y = de la región)

$$f(x,y) = \begin{cases} 24xy, & 0 \leq x, y \leq 1;\ x + y \leq 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre la probabilidad de que en una caja dada el tabaco turco represente más de la mitad de la mezcla.

b) Encuentre la función de densidad marginal para la proporción del tabaco de la región.

c) Encuentre la probabilidad de que la parte de tabaco turco sea menos de 1/8, si se sabe que la mezcla contiene 3/4 de tabaco de la región.

3.64 Una compañía de seguros ofrece a sus asegurados varias opciones diferentes de pago de la prima. Para

un asegurado seleccionado al azar, sea X el número de meses entre pagos sucesivos. La función de distribución acumulada de X es

$$F(x) = \begin{cases} 0, & \text{si } x < 1, \\ 0.4, & \text{si } 1 \le x < 3, \\ 0.6, & \text{si } 3 \le x < 5, \\ 0.8, & \text{si } 5 \le x < 7, \\ 1.0, & \text{si } x \ge 7. \end{cases}$$

a) ¿Cuál es la función de masa de probabilidad de X?

b) Calcule $P(4 < X \le 7)$.

3.65 Dos componentes electrónicos de un sistema de proyectiles funcionan en conjunto para el éxito de todo el sistema. Sean X y Y la vida en horas de los dos componentes. La densidad conjunta de X y Y es

$$f(x,y) = \begin{cases} ye^{-y(1+x)}, & x, y \ge 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Determine las funciones de densidad marginal para ambas variables aleatorias.

b) ¿Cuál es la probabilidad de que ambos compontes duren más de dos horas?

3.66 Una instalación de servicio opera con dos líneas. En un día seleccionado al azar, sea X la proporción de tiempo que la primera línea está en uso; mientras que Y es la proporción de tiempo en que la segunda línea está en uso. Suponga que la función de densidad de probabilidad conjunta para (X, Y) es

$$f(x,y) = \begin{cases} \frac{3}{2}(x^2 + y^2), & 0 \le x, y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Calcule la probabilidad de que ninguna línea esté ocupada más de la mitad del tiempo.

b) Encuentre la probabilidad de que la primera línea esté ocupada más del 75% del tiempo.

3.67 Sea el número de llamadas telefónicas que recibe el conmutador durante un intervalo de 5 minutos una variable aleatoria X con función de probabilidad

$$f(x) = \frac{e^{-2}2^x}{x!}, \quad \text{para } x = 0, 1, 2, \ldots.$$

a) Determine la probabilidad de que X sea igual a 0, 1, 2, 3, 4, 5 y 6.

b) Grafique la función de masa de probabilidad para estos valores de x.

c) Determine la función de distribución acumulada para estos valores de X.

3.68 Considere las variables aleatorias X y Y con función de densidad conjunta

$$f(x,y) = \begin{cases} x + y, & 0 \le x, y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre las distribuciones marginales de X y Y.

b) Encuentre $P(X > 0.5, Y > 0.5)$.

3.69 De un proceso industrial se elaboran artículos que se clasifican como defectuosos o no defectuosos. La probabilidad de que un artículo esté defectuoso es 0.1. Se lleva a cabo un experimento, en el cual se sacan 5 artículos al azar del proceso. Sea la variable aleatoria X el número de defectuosos en esta muestra de 5. ¿Cuál es la función de masa de probabilidad de X?

3.70 Considere la siguiente función de densidad de probabilidad conjunta de las variables aleatorias X y Y:

$$f(x,y) = \begin{cases} \frac{3x-y}{9}, & 1 < x < 3,\ 1 < y < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre las funciones de densidad marginal de X y Y.

b) ¿Son independientes X y Y?

c) Encuentre $P(X > 2)$.

3.71 La duración en horas de un componente eléctrico es una variable aleatoria con función de distribución acumulada

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{50}}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Determine su función de densidad de probabilidad.

b) Determine la probabilidad de que la vida útil de tal componente exceda 70 horas.

3.72 Ciertas instalaciones industriales producen pantalones. Un grupo de 10 trabajadores los "verifican". Los trabajadores inspecionan pantalones que se toman aleatoriamente de la línea de producción. A cada inspector se le asigna un número del 1 al 10. Un comprador selecciona un pantalón para adquirirlo. Sea la variable aleatoria X el número del inspector.

a) Determine una función de masa de probabilidad razonable para X.

b) Grafique la función de distribución acumulada para X.

3.73 La vida en anaquel de un producto es una variable aleatoria que se relaciona con la aceptación por parte del consumidor. Resulta que la vida en anaquel Y, en días, de cierta clase de artículo de panadería tiene una función de densidad

$$f(y) = \begin{cases} \frac{1}{2}e^{-y/2}, & 0 \le y < \infty, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

¿Qué fracción de estos panes que se exhiben hoy estarán a la venta dentro de 3 días?

3.74 El congestionamiento de pasajeros es un problema en servicio de los aeropuertos. Dentro de éstos se instalan trenes para reducir la congestión. Usando el tren, el tiempo X, en minutos, que toma viajar desde la terminal principal hasta una explanada específica tiene una función de densidad

$$f(x) = \begin{cases} \frac{1}{10}, & 0 \le x \le 10, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Muestre que la función de densidad de probabilidad anterior es válida.

b) Encuentre la probabilidad de que el tiempo que toma a un pasajero viajar desde la terminal principal hasta la explanada no excederá los 7 minutos.

3.75 Las impurezas en el lote del producto final de un proceso químico a menudo reflejan un grave problema. A partir de una cantidad considerable datos recabados en la planta, se sabe que la proporción de Y de las impurezas en un lote tiene una función de densidad dada por

$$f(y) = \begin{cases} 10(1-y)^9, & 0 \le y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique que la función de densidad anterior sea válida.

b) Se considera que un lote no es vendible y, por consiguiente, no es aceptable si el porcentaje de impurezas supera 60%. Con la calidad del proceso actual, ¿cuál es el porcentaje de lotes que no son aceptables?

3.76 El tiempo Z en minutos entre llamadas a un sistema de alimentación eléctrica tiene la función de densidad de probabilidad

$$f(z) = \begin{cases} \frac{1}{10} e^{-z/10}, & 0 < z < \infty, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) ¿Cuál es la probabilidad de que no haya llamadas en un lapso de 20 minutos?

b) ¿Cuál es la probabilidad de que la primera llamada entre en los primeros 10 minutos?

3.77 Un sistema químico que surge a partir de una reacción química tiene dos componentes importantes, entre otros, en una mezcla. La distribución conjunta que describe la proporción X_1 y X_2 de estos dos componentes está dada por

$$f(x_1, x_2) = \begin{cases} 2, & 0 < x_1 < x_2 < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Determine la distribución marginal de X_1.

b) Determine la distribución marginal de X_2.

c) ¿Cuál es la probabilidad de que las proporciones del componente generen los resultados $X_1 < 0.2$ y $X_2 > 0.5$?

d) Determine la distribución condicional $f_{X_1|X_2}(x_1|x_2)$.

3.78 Considere la situación del ejercicio de repaso 3.77; pero suponga que la distribución conjunta de las dos proporciones está dada por

$$f(x_1, x_2) = \begin{cases} 6x_2, & 0 < x_2 < x_1 < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Determine la distribución marginal $fx_1(x_1)$ de la proporción X_1 y verifique que sea una función de densidad válida.

b) ¿Cuál es la probabilidad de que la proporción X_2 sea menor que 0.5 dado que X_1 es 0.7?

3.79 Considere que las variables aleatorias X y Y representan el número de vehículos que llegan a dos esquinas de calles separadas durante cierto periodo de 2 minutos. Estas esquinas de las calles están bastante cerca una de la otra, de manera que es importante que los ingenieros de tráfico se ocupen de ellas de manera conjunta si es necesario. Se sabe que la distribución conjunta de X y Y es

$$f(x, y) = \frac{9}{16} \cdot \frac{1}{4^{(x+y)}},$$

para $x = 0, 1, 2, \ldots,$ y para $y = 0, 1, 2, \ldots$.

a) ¿Son independientes las dos variables aleatorias X y Y? Explique por qué.

b) ¿Cuál es la probabilidad de que durante el periodo en cuestión menos de 4 vehículos lleguen a las dos esquinas.

3.80 El comportamiento de series de componentes juega un papel importante en problemas de confiabilidad científicos y de ingeniería. Ciertamente la confiabilidad de todo el sistema no es mejor que el componente más débil de las series. En un sistema de series los componentes funcionan independientemente entre sí. En un sistema particular de tres componentes la probabilidad de cumplir con la especificación para los componentes 1, 2 y 3, respectivamente, son 0.95, 0.99 y 0.92. ¿Cuál es la probabilidad de que todo el sistema funcione?

3.81 Otro tipo de sistema que se utiliza en trabajos de ingeniería es un grupo de componentes en paralelo o sistema paralelo. En este enfoque más conservador, la probabilidad de que el sistema funcione es mayor que la probabilidad de que culaquier componente funcione. El sistema fallará sólo cuando todo el sistema falle. Considere una situación en que haya 4 componentes independientes en un sistema paralelo con la probabilidad de operación dada por

Componente 1: 0.95; Componente 2: 0.94;

Componente 3: 0.90; Componente 4: 0.97.

¿Cuál es la probabilidad de que no falle el sistema?

3.82 Considere un sistema de componentes en que haya cinco componentes independientes, cada uno de los cuales tiene una probabilidad de operación de 0.92. De hecho, el sistema tiene una redundancia preventiva en el sistema diseñada para que éste no falle mientras 3 de sus 5 componentes estén en funcionamiento. ¿Cuál es la probabilidad de que funcione todo el sistema?

3.5 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

En futuros capítulos será evidente que las distribuciones de probabilidad representan la estructura mediante la cual las probabilidades que se calculan ayudan a evaluar y a comprender un proceso. En el ejercicio de repaso 3.67, por ejemplo, la distribución de probabilidad que cuantifica la probabilidad de que haya una carga excesiva durante ciertos periodos podría ser muy útil en la planeación de cualesquiera cambios en el sitema. El ejercicio de repaso 3.71 describe un escenario donde se estudia el periodo de vida útil de un componente electrónico. Conocer la estructura de la probabilidad para el componente contribuirá de manera significativa con el entendimiento de la confiabilidad de un sistema mayor del cual forma parte el componente. Además, comprender la naturaleza general de las distribuciones de probabilidad reforzará el conocimiento del concepto **valor-*P***, que se estudió brevemente en el capítulo 1, jugará un papel destacado al inicio del capítulo 10 y continuará por todo el texto.

Los capítulos 4, 5 y 6 dependen en mucho del material cubierto en este capítulo. En el capítulo 4 estudiaremos el significado de **parámetros** importantes en las distribuciones de probabilidad. Tales parametros cuantifican las nociones de **tendencia central** y **variabilidad** en un sistema. De hecho, el conocimiento mismo de tales cantidades, al margen de la distribución completa, puede ofrecer información sobre la naturaleza del sistema. En los capítulos 5 y 6 se examinarán escenarios de ingeniería, biológicos y de ciencia en general, que identifican tipos de distribuciones especiales. Por ejemplo, la estructura de la función de probabilidad en el ejercicio de repaso 3.67 se identificará fácilmente bajo ciertas suposiciones estudiadas en el capítulo 5. Lo mismo ocurre en el contexto del ejercicio de repaso 3.71. Éste es un caso especial de problema sobre **tiempo de operación antes del fallo**, cuya función de densidad de probabilidad se estudiará en el capítulo 6.

En lo que concierne a los riegos potenciales de utilizar el material de este capítulo, la "advertencia" para el lector sería no leer el material más allá de lo que sea evidente. La naturaleza general de la distribución de probabilidad para un fenómeno científico determinado no es obvia a partir de lo que se estudio aquí. La finalidad de este capítulo es aprender a manipular una distribución de probabilidad, no saber cómo identificar un tipo específico. Los capítulos 5 y 6 recorren un largo trecho hacia la identficación respecto de la naturaleza general del sistema científico.

Capítulo 4

Esperanza matemática

4.1 Media de una variable aleatoria

Si dos monedas se lanzan 16 veces y X es el número de caras que ocurre por cada lanzamiento, entonces los valores de X pueden ser 0, 1 y 2. Suponga que en el experimento salen cero caras, una cara y dos caras, respectivamente, un total de 4, 7 y 5 veces. El número promedio de caras por lanzamiento de las dos monedas es, entonces,

$$\frac{(0)(4)+(1)(7)+(2)(5)}{16} = 1.06.$$

Éste es un valor promedio y no necesariamente es un resultado posible del experimento. Por ejemplo, el ingreso mensual promedio de un vendedor probablemente no sea igual a alguno de sus cheques de pago mensuales.

Reestructuremos ahora nuestro cálculo del número promedio de caras, de manera que tengamos la siguiente forma equivalente:

$$(0)\left(\frac{4}{16}\right)+(1)\left(\frac{7}{16}\right)+(2)\left(\frac{5}{16}\right) = 1.06.$$

Los números 4/16, 7/16 y 5/16 son las fracciones de los lanzamientos totales que resultan, respectivamente, en 0, 1 y 2 caras. Tales fracciones también son las frecuencias relativas de los diferentes valores de X en nuestro experimento. En efecto, entonces, calculamos la media o promedio de un conjunto de datos utilizando el conocimiento de los distintos valores que ocurren y sus frecuencias relativas, sin un conocimiento del número total de observaciones en nuestro conjunto de datos. Por lo tanto, si 4/16 o 1/4 de los lanzamientos tienen como resultado cero caras, 7/16 de los lanzamientos tienen como resultado una cara y 5/16 de éstos tienen dos caras, el número medio de caras por lanzamiento sería 1.06, sin importar si el número total de lanzamientos fue 16, 1000 o incluso 10,000.

Utilicemos ahora este método de frecuencias relativas para calcular el número promedio de caras por el lanzamiento de dos monedas que esperaríamos en el largo plazo. Nos referiremos a este valor promedio como la **media de la variable aleatoria** X o **la media de la distribución de probabilidad de** X, y la denotamos con μ_x o simplemente como μ cuando esté claro a qué variable aleatoria nos referimos.

También es común entre los estadísticos referirse a esta media como la esperanza matemática o el valor esperado de la variable aleatoria X y denotarla como $E(X)$.

Suponga que se lanzan monedas legales, y encontramos que el espacio muestral para nuestro experimento es

$$S = \{HH, HT, TH, TT\}.$$

Como los 4 puntos muestrales son igualmente probables, se sigue que

$$P(X = 0) = P(TT) = \frac{1}{4}, \quad P(X = 1) = P(TH) + P(HT) = \frac{1}{2},$$

y

$$P(X = 2) = P(HH) = \frac{1}{4},$$

donde un elemento típico, digamos TH, indica que el primer lanzamiento tuvo como resultado una cruz seguida de una cara en el segundo lanzamiento. Así, estas probabilidades son precisamente las frecuencias relativas para los eventos dados en el largo plazo. Por lo tanto,

$$\mu = E(X) = (0)\left(\frac{1}{4}\right) + (1)\left(\frac{1}{2}\right) + (2)\left(\frac{1}{4}\right) = 1.$$

Este resultado quiere decir que una persona que lance 2 monedas una y otra vez, en promedio, obtendrá 1 cara por cada lanzamiento.

El método descrito antes para calcular el número esperado de caras en el lanzamiento de 2 monedas sugiere que la media, o el valor esperado de cualquier variable aleatoria, discreta se puede obtener multiplicando cada uno de los valores $x_1, x_2, \ldots, x_n$ de la variable aleatoria X por su probabilidad correspondiente $f(x_1), f(x_2), \ldots, f(x_n)$ y sumando los productos. Esto es cierto, sin embargo, sólo si la variable aleatoria es discreta. En el caso de variables aleatorias continuas, la definición de un valor esperado es esencialmente la misma, pero con integrales que reemplazan las sumatorias.

Definición 4.1: Sea X una variable aleatoria con distribución de probabilidad $f(x)$. La **media** o **valor esperado** de X es

$$\mu = E(X) = \sum_x x f(x)$$

si X es discreta, y

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

si X es continua.

Ejemplo 4.1: Un inspector de calidad muestrea un lote que contiene 7 componentes; el lote contiene 4 componentes buenos y 3 defectuosos. El inspector toma una muestra de 3 componentes. Encuentre el valor esperado del número de componentes buenos en esta muestra.

Solución: Sea X el número de componentes buenos en la muestra. La distribución de probabilidad de X es

$$f(x) = \frac{\binom{4}{x}\binom{3}{3-x}}{\binom{7}{3}}, \qquad x = 0, 1, 2, 3.$$

Unos cálculos sencillos dan $f(0) = 1/35$, $f(1) = 12/35$, $f(2) = 18/35$ y $f(3) = 4/35$. Por lo tanto,

$$\mu = E(X) = (0)\left(\frac{1}{35}\right) + (1)\left(\frac{12}{35}\right) + (2)\left(\frac{18}{35}\right) + (3)\left(\frac{4}{35}\right) = \frac{12}{7} = 1.7.$$

De esta manera, si se selecciona al azar una muestra de tamaño 3 una y otra vez de un lote de 4 componentes buenos y 3 defectuosos, contendría, en promedio, 1.7 componentes buenos.

Ejemplo 4.2: En un juego de azar se pagarán \$5 a una persona si le salen puras caras o puras cruces cuando se lanzan tres monedas, y ella pagará \$3 si salen una o dos caras. ¿Cuál es su ganancia esperada?

Solución: El espacio muestral para los posibles resultados cuando se lanzan de manera simultánea tres monedas o, de manera equivalente, si se lanza tres veces 1 moneda, es

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

Se podría argumentar que cada una de estas posibilidades es igualmente probable y que ocurre con probabilidad de 1/8. Un método alternativo sería aplicar la regla de la multiplicación de probabilidades para eventos independientes a cada elemento de S. Por ejemplo,

$$P(HHT) = P(H)P(H)P(T) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8}.$$

La variable aleatoria de interés es Y, el monto que el jugador puede ganar; y los valores posibles de Y son \$5 si ocurre el evento $E_1 = \{HHH, TTT\}$ y −\$3 si ocurre el evento

$$E_2 = \{HHT, HTH, THH, HTT, THT, TTH\}.$$

Como E_1 y E_2 ocurren con probabilidades 1/4 y 3/4, respectivamente, se sigue que

$$\mu = E(Y) = (5)\left(\frac{1}{4}\right) + (-3)\left(\frac{3}{4}\right) = -1.$$

En este juego la persona perderá, en promedio, \$1 por lanzamiento de las tres monedas. Un juego se considera "equitativo" si el jugador, en promedio, sale empatado. Por lo tanto, una ganancia esperada de cero define un juego equitativo.

Los ejemplos 4.1 y 4.2 se diseñaron para permitir al lector lograr una mejor comprensión de lo que queremos decir por valor esperado de una variable aleatoria. En ambos casos, las variables aleatorias son discretas. Seguimos con un ejemplo de variable aleatoria continua, donde un ingeniero se interesa en la *vida media* de cierto tipo de dispositivo electrónico. Ésta es una ilustración del problema de *tiempo de*

operación antes del fallo que a menudo ocurre en la práctica. El valor esperado de la vida del dispositivo es un parámetro importante para su evaluación.

Ejemplo 4.3: Sea X la variable aleatoria que denota la vida en horas de cierto dispositivo electrónico. La función de densidad de probabilidad es

$$f(x) = \begin{cases} \frac{20{,}000}{x^3}, & x > 100, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la vida esperada para esta clase de dispositivo.

Solución: Con la definición 4.1, tenemos

$$\mu = E(X) = \int_{100}^{\infty} x\frac{20{,}000}{x^3}dx = \int_{100}^{\infty} \frac{20{,}000}{x^2}dx = 200.$$

Por lo tanto, esperamos que, *en promedio*, este tipo de dispositivo dure 200 horas.

Consideremos ahora una nueva variable aleatoria $g(X)$, la cual depende de X; es decir, cada valor de $g(X)$ está determinado al conocer los valores de X. Por ejemplo, $g(X)$ podría ser X^2 o $3X - 1$, de manera que siempre que X tome el valor 2, $g(X)$ toma el valor $g(2)$. En particular, si X es una variable aleatoria discreta con distribución de probabilidad $f(x)$, $x = -1$, 0, l, 2 y $g(X) = X^2$, entonces,

$$\begin{aligned} P[g(X) = 0] &= P(X = 0) = f(0), \\ P[g(X) = 1] &= P(X = -1) + P(X = 1) = f(-1) + f(1), \\ P[g(X) = 4] &= P(X = 2) = f(2), \end{aligned}$$

de manera que la distribución de probabilidad de $g(X)$ se escribe como

$g(x)$	0	1	4
$P[g(X) = g(x)]$	$f(0)$	$f(-1) + f(1)$	$f(2)$

Por definición del valor esperado de una variable aleatoria, obtenemos

$$\begin{aligned} \mu_{g(X)} &= E[g(x)] = 0f(0) + 1[f(-1) + f(1)] + 4f(2) \\ &= (-1)^2 f(-1) + (0)^2 f(0) + (1)^2 f(1) + (2)^2 f(2) = \sum_{x} g(x)f(x). \end{aligned}$$

Este resultado se generaliza en el teorema 4.1 para variables aleatorias discretas y continuas.

Teorema 4.1:

Sea X una variable aleatoria con distribución de probabilidad $f(x)$. El valor esperado de la variable aleatoria $g(X)$ es

$$\mu_{g(X)} = E[g(X)] = \sum_{x} g(x)f(x)$$

si X es discreta, y

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\ dx$$

si X es continua.

Ejemplo 4.4: Suponga que el número de automóviles X que pasa por un autolavado entre 4:00 P.M. y 5:00 P.M. en cualquier viernes soleado tiene la siguiente distribución de probabilidad:

x	4	5	6	7	8	9
$P(X=x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$

Sea $g(X) = 2X - 1$ la cantidad de dinero en dólares, que el administrador paga al dependiente. Encuentre las ganancias que espera el dependiente en este periodo específico.

Solución: Por el teorema 4.1, el dependiente puede esperar recibir

$$\begin{aligned} E[g(X)] =& E(2X-1) = \sum_{x=4}^{9}(2x-1)f(x) \\ =& (7)\left(\frac{1}{12}\right) + (9)\left(\frac{1}{12}\right) + (11)\left(\frac{1}{4}\right) + (13)\left(\frac{1}{4}\right) \\ &+ (15)\left(\frac{1}{6}\right) + (17)\left(\frac{1}{6}\right) = \$12.67. \end{aligned}$$

Ejemplo 4.5: Sea X una variable aleatoria con función de densidad

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el valor esperado de $g(X) = 4X + 3$.

Solución: Por el teorema 4.1, tenemos

$$E(4X+3) = \int_{-1}^{2} \frac{(4x+3)x^2}{3} dx = \frac{1}{3}\int_{-1}^{2} (4x^3 + 3x^2)\, dx = 8.$$

Debemos extender ahora nuestro concepto de esperanza matemática al caso de dos variables aleatorias X y Y con distribución de probabilidad conjunta $f(x, y)$.

Definición 4.2: Sean X y Y variables aleatorias con distribución de probabilidad conjunta $f(x, y)$. La media o valor esperado de la variable aleatoria $g(X, Y)$ es

$$\mu_{g(X,Y)} = E[g(X,Y)] = \sum_x \sum_y g(x,y)f(x,y)$$

si X y Y son discretas, y

$$\mu_{g(X,Y)} = E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\, dx\, dy$$

si X y Y son continuas.

Es evidente la generalización de la definición 4.2 para el cálculo de la esperanza matemática de funciones de diversas variables aleatorias.

Ejemplo 4.6: Sean X y Y variables aleatorias con la distribución de probabilidad conjunta que se indica en la tabla 3.1 de la página 92. Encuentre el valor esperado de $g(X, Y) = XY$. Por conveniencia se repite aquí la tabla.

$f(x,y)$		x = 0	x = 1	x = 2	Totales por renglón
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Totales por columna		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Solución: Por la definición 4.2, escribimos

$$\begin{aligned} E(XY) &= \sum_{x=0}^{2}\sum_{y=0}^{2} xyf(x,y) = (0)(0)f(0,0) + (0)(1)f(0,1) \\ &\quad + (1)(0)f(1,0) + (1)(1)f(1,1) + (2)(0)f(2,0) \\ &= f(1,1) = \frac{3}{14}. \end{aligned}$$

Ejemplo 4.7: Encuentre (Y/X) para la función de densidad

$$f(x,y) = \begin{cases} \frac{x(1+3y^2)}{4}, & 0 < x < 2,\ 0 < y < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Solución: Tenemos

$$E\left(\frac{Y}{X}\right) = \int_0^1 \int_0^2 \frac{y(1+3y^2)}{4}\, dx\, dy = \int_0^1 \frac{y+3y^3}{2} dy = \frac{5}{8}.$$

Observe que si $g(X, Y) = X$ en la definición 4.2, tenemos

$$E(X) = \begin{cases} \sum_x \sum_y xf(x,y) = \sum_x xg(x), & \text{(caso discreto)}, \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\, dy\, dx = \int_{-\infty}^{\infty} xg(x)\, dx, & \text{(caso continuo)}, \end{cases}$$

donde $g(x)$ es la distribución marginal de X. Por lo tanto, para calcular $E(X)$ en un espacio bidimensional, se puede utilizar tanto la distribución de probabilidad conjunta de X y Y, como la distribución marginal de X. De manera similar, definimos

$$E(Y) = \begin{cases} \sum_y \sum_x yf(x,y) = \sum_y yh(y), & \text{(caso discreto)}, \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\, dx\, dy = \int_{-\infty}^{\infty} yh(y)\, dy, & \text{(caso continuo)}, \end{cases}$$

donde $h\ (y)$ es la distribución marginal de la variable aleatoria Y.

Ejercicios

4.1 Suponga que dos variables aleatorias (X, Y) se distribuyen de manera uniforme en un círculo con radio a. La función densidad de probabilidad conjunta es, entonces,

$$f(x,y) = \begin{cases} \frac{1}{\pi a^2}, & x^2 + y^2 \leq a^2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el valor esperado de X, μx.

4.2 La distribución de probabilidad de la variable aleatoria discreta X es

$$f(x) = \binom{3}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{3-x}, \quad x = 0, 1, 2, 3.$$

Encuentre la media de X.

4.3 Encuentre la media de la variable aleatoria T que representa el total de las tres monedas del ejercicio 3.25 de la página 89.

4.4 Una moneda está cargada de manera que la probabilidad de ocurrencia de una cara es tres veces mayor que la de una cruz. Encuentre el número esperado de cruces cuando se lanza dos veces esta moneda.

4.5 La distribución de probabilidad de X, el número de imperfecciones por cada 10 metros de una tela sintética, en rollos continuos de ancho uniforme, está dada en el ejercicio 3.13 de la página 89 como

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

Encuentre el número promedio de imperfecciones en 10 metros de esta tela.

4.6 A un dependiente de un autolavado se le paga de acuerdo con el número de automóviles que lava. Suponga que las probabilidades son 1/12, 1/12, 1/4, 1/4, 1/6, y 1/6, respectivamente, de que el dependiente reciba \$7, \$9, \$11, \$13, \$15 o \$17 entre 4:00 P.M. y 5:00 P.M. en cualquier viernes soleado. Encuentre las ganancias que esperada el dependiente para este periodo específico.

4.7 Al invertir en unas accioness particulares, en un año un individuo puede obtener una ganancia de \$4000 con probabilidad de 0.3, o tener una pérdida de \$1000 con probabilidad de 0.7. ¿Cuál es la ganancia esperada por esta persona?

4.8 Suponga que un distribuidor de joyería antigua se interesa en comprar un collar de oro, para el que las probabilidades son 0.22, 0.36, 0.28 y 0.14, respectivamente, de que pueda venderlo con una ganancia de \$250, venderlo con una ganancia de \$150, venderlo al costo o venderlo con una pérdida de \$150. ¿Cuál es su ganancia esperada?

4.9 En un juego de azar, a una mujer se le pagan \$3 si saca un jack o una reina, y \$5 si saca un rey o un as de una baraja ordinaria de 52 cartas. Pierde si saca cualquier otra carta. ¿Cuánto debería pagar si el juego es justo?

4.10 Dos expertos en calidad de neumáticos examinan lotes de éstos y asignan puntuaciones de calidad a cada neumático en una escala de tres puntos. Sea X la puntuación dada por el experto A y Y la del experto B. La siguiente tabla presenta la distribución conjunta para X y Y.

		y		
$f(x,y)$		1	2	3
	1	0.10	0.05	0.02
x	2	0.10	0.35	0.05
	3	0.03	0.10	0.20

Encuentre μ_X y μ_Y.

4.11 Un piloto privado desea asegurar su avión por \$200,000. La compañía de seguros estima que puede ocurrir una pérdida total con probabilidad de 0.002, una pérdida de 50% con probabilidad de 0.01 y una pérdida de 25% con probabilidad de 0.1. Si se ignoran todas las demás pérdidas parciales, ¿qué prima debería cobrar cada año la compañía de seguros para tener una utilidad promedio de \$500?

4.12 Si la ganancia de un distribuidor, en unidades de \$5000, para un automóvil nuevo se puede ver como una variable aleatoria X que tiene la función de densidad

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

encuentre la ganancia promedio por automóvil.

4.13 La función de densidad de las mediciones codificadas del diámetro de paso de los hilos de un encaje es

$$f(x) = \begin{cases} \frac{4}{\pi(1+x^2)}, & 0 < x < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el valor esperado de X.

4.14 ¿Qué proporción de individuos se puede esperar que respondan a cierta encuesta que se envía por correo, si la proporción X tiene la función de densidad

$$f(x) = \begin{cases} \frac{2(x+2)}{5}, & 0 < x < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

4.15 La función de densidad de la variable aleatoria continua X, el número total de horas, en unidades de 100 horas, que una familia utiliza una aspiradora en un periodo de un año, se da en el ejercicio 3.7 de la página 88 como

$$f(x) = \begin{cases} x, & 0 < x < 1, \\ 2 - x, & 1 \le x < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el número promedio de horas por año que las familias utilizan sus aspiradoras.

4.16 Suponga que usted inspecciona un lote de 1000 bombillas de luz, entre los cuales hay 20 defectuosos. Elija al azar dos bombillas del lote sin reemplazo. Sean

$$X_1 = \begin{cases} 1, & \text{si la primera bombilla está defectuosa,} \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{si la primera bombilla está defectuosa,} \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la probabilidad de que al menos una bombilla esté defectuosa. [*Sugerencia:* Calcule $P(X_1 + X_2 = 1)$.]

4.17 Sea X una variable aleatoria con la siguiente distribución de probabilidad:

x	-3	6	9
$f(x)$	1/6	1/2	1/3

Encuentre $\mu_{g(X)}$, donde $g(X) = (2X + 1)^2$.

4.18 Encuentre el valor esperado de la variable aleatoria $g(X) = X^2$, donde X tenga la distribución de probabilidad del ejercicio 4.2.

4.19 Una empresa industrial grande compra varios procesadores de palabras nuevos al final de cada año; el número exacto depende de la frecuencia de reparaciones en el año anterior. Suponga que el número de procesadores de palabras, X, que se compran cada año tiene la siguiente distribución de probabilidad:

x	0	1	2	3
$f(x)$	1/10	3/10	2/5	1/5

Si el costo del modelo que se desea permanecerá fijo en \$1200 a lo largo de este año y se obtiene un descuento de $50X^2$ dólares en cualquier compra, ¿cuánto espera gastar esta empresa en nuevos procesadores de palabras al final de este año?

4.20 Una variable aleatoria continua X tiene la función de densidad

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el valor esperado de $g(X) = e^{2X/3}$.

4.21 ¿Cuál es la ganancia promedio por automóvil del distribuidor, si la ganancia en cada uno está dada por $g(X) = X^2$, donde X es una variable aleatoria que tiene la función de densidad del ejercicio 4.12?

4.22 El periodo de hospitalización, en días, para pacientes que siguen el tratamiento para cierto tipo de trastorno renal es una variable aleatoria $Y = X + 4$, donde X tiene la función de densidad

$$f(x) = \begin{cases} \frac{32}{(x+4)^3}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el número promedio de días que un individuo permanece hospitalizada para seguir el tratamiento para dicha enfermedad.

4.23 Suponga que X y Y tienen la siguiente función de probabilidad conjunta:

		x	
$f(x, y)$		2	4
	1	0.10	0.15
y	3	0.20	0.30
	5	0.10	0.15

a) Encuentre el valor esperado de $g(X, Y) = XY^2$.

b) Encuentre μ_X y μ_Y.

4.24 Con referencia a las variables aleatorias cuya disribución de probabilidad conjunta se da en el ejercicio 3.39 de la página 101,

a) encuentre $E(X^2Y - 2XY)$;

b) encuentre $\mu_X - \mu_Y$.

4.25 Refiérase a las variables aleatorias cuya distribución de probabilidad conjunta se da en el ejercicio 5.53 de la página 103, y encuentre la media para el número total de jacks y reyes cuando se sacan 3 cartas sin reemplazo de las 12 cartas mayores de una baraja ordinaria de 52 cartas.

4.26 Sean X y Y variables aleatorias con función de densidad conjunta

$$f(x, y) = \begin{cases} 4xy, & 0 < x,\ y < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el valor esperado de $Z = \sqrt{X^2 + Y^2}$.

4.27 En el ejercicio 3.27 de la página 89, una función de densidad está dada por el tiempo de operación antes del fallo de un componente importante de un reproductor de DVD. Encuentre el número medio en horas antes del fallo del compnente y, por lo tanto, del DVD.

4.28 Considere la información del ejercicio 3.28 de la página 89. El problema tiene que ver con los pesos, en onzas, del producto en una caja de cereal con

$$f(x) = \begin{cases} \frac{2}{5}, & 23.75 \leq x \leq 26.25, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Grafique la función de densidad.

b) Calcule el valor esperado o peso medio en onzas.

c) ¿Se sorprende de su respuesta en *b*)? Explique.

4.29 En el ejercicio 3.29 de la página 90, tratamos con una importante distribución del tamaño de las partículas, en que la distribución del tamaño de las partículas está caracterizada por

$$f(x) = \begin{cases} 3x^{-4}, & x > 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Grafique la función de densidad.

b) Determine el tamaño medio de la partícula.

4.30 En el ejercicio 3.31 de la página 90, la distribución del tiempo antes de una reparación mayor de una lavadora estuvo dada como

$$f(y) = \begin{cases} \frac{1}{4}e^{-y/4}, & y \geq 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

¿Cuál es el "tiempo para reparación" medio de la población?

4.31 Considere el ejercicio 3.32 de la página 90.

a) ¿Cuál es la proporción media del presupuesto asignado a los controles ambiental y de la contaminación?

b) ¿Cuál es la probabilidad de que elegida al azar tendrá una proporción asignada a los controles ambiental y de la contaminación que exceda la media de la población dada en *a*)?

4.32 En el ejercicio 3.13 de la página 89, la distribución del número de imperfecciones por 10 metros de tela sintética está dada por

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

a) Grafique la función de probabilidad.

b) Encuentre el número de imperfecciones esperado $E(X) = \mu$.

c) Encuentre $E(X^2)$.

4.2 Varianza y covarianza de variables aleatorias

La media o valor esperado de una variable aleatoria X es de especial importancia en estadística, ya que describe el lugar donde se centra la distribución de probabilidad. Por sí misma, sin embargo, la media no ofrece una descripción adecuada de la forma de la distribución. Necesitamos caracterizar la variabilidad en la distribución. En la figura 4.1 tenemos los histogramas de dos distribuciones de probabilidad discretas con la misma media $\mu = 2$, que difieren de manera considerable en la variabilidad o dispersión de sus observaciones alrededor de la media.

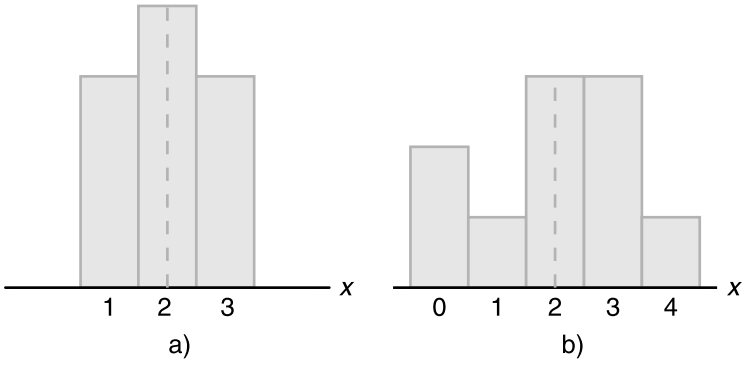


Figura 4.1: Distribuciones con medias iguales y dispersiones diferentes.

La medida de variabilidad más importante de una variable aleatoria X se obtiene al aplicar el teorema 4.1 en $g(X) = (X - \mu)^2$. A causa de su importancia en estadística, se le denomina **varianza de la variable aleatoria** X o **varianza de la distribución de probabilidad de** X y se denota como $Var(X)$ o con el símbolo σ_X^2, o simplemente σ^2 cuando queda claro a qué variable aleatoria nos referimos.

Definición 4.3: Sea X una variable aleatoria con distribución de probabilidad $f(x)$ y media μ. La varianza de X es

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x), \qquad \text{si } X \text{ es discreta, y}$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx, \qquad \text{si } X \text{ es continua.}$$

La raíz cuadrada positiva de la varianza, σ, se llama **desviación estándar** de X.

La cantidad $x - \mu$ en la definición 4.3 se llama **desviación de una observación** respecto a su media. Como estas desviaciones se elevan al cuadrado y después se promedian, σ^2 será mucho menor para un conjunto de valores x que estén cercanos a μ, que para un conjunto de valores que varíe de forma considerable de μ.

Ejemplo 4.8: Sea la variable aleatoria X el número de automóviles que se utilizan con propósitos de negocios oficiales en un día de trabajo dado. La distribución de probabilidad para la compañía A [figura 4.1a)] es

x	1	2	3
$f(x)$	0.3	0.4	0.3

y para la compañía B [figura 4.1b)] es

x	0	1	2	3	4
$f(x)$	0.2	0.1	0.3	0.3	0.1

Muestre que la varianza de la distribución de probabilidad para la compañía B es mayor que la de la compañía A.

Solución: Para la compañía A, encontramos que

$$\mu_A = E(X) = (1)(0.3) + (2)(0.4) + (3)(0.3) = 2.0,$$

y entonces

$$\sigma_A^2 = \sum_{x=1}^{3} (x-2)^2 = (1-2)^2(0.3) + (2-2)^2(0.4) + (3-2)^2(0.3) = 0.6.$$

Para la compañía B, tenemos

$$\mu_B = E(X) = (0)(0.2) + (1)(0.1) + (2)(0.3) + (3)(0.3) + (4)(0.1) = 2.0,$$

por lo que

$$\sigma_B^2 = \sum_{x=0}^{4} (x-2)^2 f(x) = (0-2)^2(0.2) + (1-2)^2(0.1) + (2-2)^2(0.3) + (3-2)^2(0.3) + (4-2)^2(0.1) = 1.6.$$

De forma clara, la varianza del número de automóviles que se utilizan con propósitos de negocios oficiales es mayor para la compañía B que para la compañía A.

Una fórmula alternativa que se prefiere para encontrar σ^2, que a menudo simplifica los cálculos, se establece en el siguiente teorema.

Teorema 4.2: La varianza de una variable aleatoria X es

$$\sigma^2 = E(X^2) - \mu^2.$$

Prueba: Para el caso discreto escribimos

$$\begin{aligned}\sigma^2 &= \sum_x (x-\mu)^2 f(x) = \sum_x (x^2 - 2\mu x + \mu^2) f(x) \\ &= \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x).\end{aligned}$$

Como $\mu = \sum_x x f(x)$ por definición, y $\sum_x f(x) = 1$ para cualquier distribución de probabilidad discreta, se sigue que

$$\sigma^2 = \sum_x x^2 f(x) - \mu^2 = E(X^2) - \mu^2.$$

Para el caso continuo la demostración es la misma paso a paso, si se reemplazan las sumatorias por integrales.

Ejemplo 4.9: Sea la variable aleatoria X el número de partes defectuosas de una máquina, cuando se muestrean y se prueban tres partes de una línea de producción. La siguiente es la distribución de probabilidad de X.

x	0	1	2	3
$f(x)$	0.51	0.38	0.10	0.01

Con el teorema 4.2, calcule σ^2.

Solución: Primero, calculamos

$$\mu = (0)(0.51) + (1)(0.38) + (2)(0.10) + (3)(0.01) = 0.61.$$

Luego,

$$E(X^2) = (0)(0.51) + (1)(0.38) + (4)(0.10) + (9)(0.01) = 0.87.$$

Por lo tanto,

$$\sigma^2 = 0.87 - (0.61)^2 = 0.4979.$$

Ejemplo 4.10: La demanda semanal de Pepsi, en miles de litros, de una cadena local de tiendas al menudeo, es una variable aleatoria continua X que tiene la densidad de probabilidad

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la media y la varianza de X.

Solución:

$$\mu = E(X) = 2\int_1^2 x(x-1)\,dx = \frac{5}{3},$$

y

$$E(X^2) = 2\int_1^2 x^2(x-1)\,dx = \frac{17}{6}.$$

Por lo tanto,

$$\sigma^2 = \frac{17}{6} - \left(\frac{5}{3}\right)^2 = \frac{1}{18}.$$

Hasta el momento, la varianza o la desviación estándar sólo tiene significado cuando comparamos dos o más distribuciones que tienen las mismas unidades de medida. Por lo tanto, podemos comparar las varianzas de las distribuciones de contenidos, medidos en litros, para dos compañías que embotellan jugo de naranja y el valor más grande indicaría la compañía cuyo producto es más variable o menos uniforme. No tendría caso comparar la varianza de una distribución de alturas con la varianza de una distribución de puntuaciones de aptitud. En la sección 4.4 mostramos cómo se utiliza la desviación estándar para describir una sola distribución de observaciones.

Extenderemos ahora nuestro concepto de varianza de una variable aleatoria X para incluir también variables aleatorias relacionadas con X. Para la variable aleatoria $g(X)$, la varianza se denotará con $\sigma^2_{g(X)}$ y se calculará empleando el siguiente teorema.

Teorema 4.3: Sea X una variable aleatoria con distribución de probabilidad $f(x)$. La varianza de la variable aleatoria $g(X)$ es

$$\sigma^2_{g(X)} = E\{[g(X) - \mu_{g(X)}]^2\} = \sum_x [g(x) - \mu_{g(X)}]^2 f(x)$$

si X es discreta, y

$$\sigma^2_{g(X)} = E\{[g(X) - \mu_{g(X)}]^2\} = \int_{-\infty}^{\infty} [g(x) - \mu_{g(X)}]^2 f(x)\,dx$$

si X es continua.

Prueba: Como $g(X)$ es en sí misma una variable aleatoria con media $\mu_{g(X)}$, Como se define en el teorema 4.1, de la definición 4.3 se sigue que

$$\sigma^2_{g(X)} = E\{[g(X) - \mu_{g(X)}]\}.$$

Así, cuando se aplica el teorema 4.1 nuevamente a la variable aleatoria $[g(X) - \mu_{g(X)}]^2$, la demostración queda completa.

Ejemplo 4.11: Calcule la varianza de $g(X) = 2X + 3$, donde X es una variable aleatoria con distribución de probabilidad

x	0	1	2	3
$f(x)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{8}$

Solución: Primero encontramos la media de la variable aleatoria $2X + 3$. De acuerdo con el teorema 4.1,

$$\mu_{2X+3} = E(2X+3) = \sum_{x=0}^{3} (2x+3) f(x) = 6.$$

Ahora, con el teorema 4.3, tenemos

$$\sigma^2_{2X+3} = E\{[(2X+3) - \mu_{2x+3}]^2\} = E[(2X+3-6)^2]$$
$$= E(4X^2 - 12X + 9) = \sum_{x=0}^{3}(4x^2 - 12x + 9)f(x) = 4.$$

Ejemplo 4.12: Sea X una variable aleatoria que tiene la función de densidad dada en el ejemplo 4.5 de la página 111. Encuentre la varianza de la variable aleatoria $g(X) = 4X + 3$.

Solución: En el ejemplo 4.5 encontramos que $\mu_{4x\,+\,3} = 8$. Ahora, usando el teorema 4.3,

$$\sigma^2_{4X+3} = E\{[(4X+3) - 8]^2\} = E[(4X-5)^2]$$
$$= \int_{-1}^{2}(4x-5)^2\frac{x^2}{3}dx = \frac{1}{3}\int_{-1}^{2}(16x^4 - 40x^3 + 25x^2)\,dx = \frac{51}{5}.$$

Si $g(X, Y) = (X - \mu_X)(Y - \mu_Y)$, donde $\mu_X = E(X)$ y $\mu_Y = E(Y)$, la definición 4.2 da un valor esperado que se llama covarianza de X y Y, que denotamos como σ_{XY} o $\text{cov}(X, Y)$.

Definición 4.4: Sean X y Y variables aleatorias con distribución de probabilidad conjunta $f(x, y)$. La covarianza de X y Y es

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_y)f(x, y)$$

si X y Y son discretas, y

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \mu_X)(y - \mu_y)f(x, y)\,dx\,dy$$

si X y Y son continuas.

La covarianza entre dos variables aleatorias es una medida de la naturaleza de la asociación entre ambas. Si valores grandes de X a menudo tienen como resultado valores grandes de Y, o valores pequeños de X tienen como resultado valores pequeños de Y, $X - \mu_X$ positiva con frecuencia tendrá como resultado $Y - \mu_Y$ positiva, y $X - \mu_X$ negativa a menudo tendrá como resultado $Y - \mu_Y$ negativa. De esta forma, el producto $(X - \mu_X)(Y - \mu_Y)$ tenderá a ser positivo. Por otro lado, si con frecuencia valores grandes de X tienen como resultado valores pequeños de Y, entonces el producto $(X - \mu_X)(Y - \mu_Y)$ tenderá a ser negativo. Así, el *signo* de la covarianza indica si la relación entre dos variables aleatorias dependientes es positiva o negativa. Cuando X y Y son estadísticamente independientes, se puede mostrar que la covarianza es cero (véase el corolario 4.5). Lo opuesto, sin embargo, por lo general, no es cierto. Dos variables pueden tener covarianza cero e incluso así no ser estadísticamente independientes. Observe que la covarianza sólo describe la relación *lineal* entre dos variables aleatorias. Por consiguiente, si una covarianza entre X y Y es cero, X y Y quizá tengan una relación no lineal, lo cual significa que no necesariamente son independientes.

La fórmula alternativa que se prefiere para σ_{XY} se establece en el teorema 4.4.

Teorema 4.4: La covarianza de dos variables aleatorias X y Y con medias μ_X y μ_Y, respectivamente, está dada por

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y.$$

Prueba: Para el caso discreto escribimos

$$\begin{aligned}\sigma_{XY} &= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)f(x,y)\\ &= \sum_x \sum_y (xy - \mu_X y - \mu_Y x + \mu_X\mu_Y)f(x,y)\\ &= \sum_x \sum_y xyf(x,y) - \mu_X \sum_x \sum_y yf(x,y)\\ &\quad - \mu_Y \sum_x \sum_y xf(x,y) + \mu_X\mu_Y \sum_x \sum_y f(x,y).\end{aligned}$$

Como

$$\mu_X = \sum_x xf(x,y) \quad \text{y} \quad \mu_Y = \sum_y yf(x,y)$$

por definición y, además,

$$\sum_x \sum_y f(x,y) = 1$$

para cualquier distribución discreta conjunta, se sigue que

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y.$$

Para el caso continuo la prueba es idéntica, pero con las sumatorias reemplazadas por integrales.

Ejemplo 4.13: En el ejemplo 3.14 de la página 92 se describió una situación con el número de repuestos azules X y el número de repuestos rojos Y. Cuando se seleccionan al azar dos repuestos para bolígrafo de cierta caja, se tiene la siguiente distribución de probabilidad conjunta,

		x			
$f(x,y)$		0	1	2	$h(y)$
	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
y	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
	$g(x)$	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Encuentre la covarianza de X y Y.

Solución: Del ejemplo 4.6, vemos que $E(XY) = 3/14$. Entonces,

$$\mu_X = \sum_{x=0}^{2} xg(x) = (0)\left(\frac{5}{14}\right) + (1)\left(\frac{15}{28}\right) + (2)\left(\frac{3}{28}\right) = \frac{3}{4},$$

y

$$\mu_Y = \sum_{y=0}^{2} yh(y) = (0)\left(\frac{15}{28}\right) + (1)\left(\frac{3}{7}\right) + (2)\left(\frac{1}{28}\right) = \frac{1}{2}.$$

Por lo tanto,

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = \frac{3}{14} - \left(\frac{3}{4}\right)\left(\frac{1}{2}\right) = -\frac{9}{56}.$$

Ejemplo 4.14: La fracción X de corredores y la fracción Y de corredoras que compiten en la maratón se describen mediante la función de densidad conjunta

$$f(x,y) = \begin{cases} 8xy, & 0 \le y \le x \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la covarianza de X y Y.

Solución: Debemos calcular primero las funciones de densidad marginal. Éstas son

$$g(x) = \begin{cases} 4x^3, & 0 \le x \le 1, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

y

$$h(y) = \begin{cases} 4y(1-y^2), & 0 \le y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

De las funciones de densidad marginal dadas arriba, calculamos

$$\mu_X = E(X) = \int_0^1 4x^4\, dx = \frac{4}{5}, \quad \text{y} \quad \mu_Y = \int_0^1 4y^2(1-y^2)\, dy = \frac{8}{15}.$$

De las funciones de densidad conjunta dadas, tenemos

$$E(XY) = \int_0^1 \int_y^1 8x^2y^2\, dx\, dy = \frac{4}{9}.$$

Entonces,

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = \frac{4}{9} - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) = \frac{4}{225}.$$

Aunque la covarianza entre dos variables aleatorias brinda información respecto de la naturaleza de la relación, la magnitud de σ_{XY} *no indica nada respecto a la fuerza de la relación*, ya que σ_{XY} depende de la escala. Su magnitud dependerá de las unidades que se miden para X y Y. Hay una versión de la covarianza libre de la escala, que se denomina **coeficiente de correlación** y que se utiliza ampliamente en estadística.

Definición 4.5: Sean X y Y variables aleatorias con covarianza σ_{XY} y desviación estándar σ_X y σ_Y, respectivamente. El coeficiente de correlación X y Y es

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Debería quedar claro para el lector que p_{XY} es independiente de las unidades de X y Y. El coeficiente de correlación satisface la desigualdad $-1 \le p_{XY} \le 1$. Toma un

valor de cero cuando $\sigma_{XY} = 0$. Donde hay una dependencia lineal exacta, digamos, $Y \equiv a + bX$, $\rho_{XY} = 1$ si $b > 0$ y $\rho_{XY} = -1$ si $b < 0$. (Véase el ejercicio 4.48.) El coeficiente de correlación es tema que amerita un estudio más amplio en el capítulo 12, donde examinamos la regresión lineal.

Ejercicios

4.33 Use la definición 4.3 de la página 116 para encontrar la varianza de la variable aleatoria X del ejercicio 4.7 de la página 113.

4.34 Sea X una variable aleatoria con la siguiente distribución de probabilidad:

x	-2	3	5
$f(x)$	0.3	0.2	0.5

Encuentre la desviación estándar de X.

4.35 La variable aleatoria X, que representa el número de errores por 100 líneas de código de programación, tiene la siguiente distribución de probabilidad:

x	2	3	4	5	6
$f(x)$	0.01	0.25	0.4	0.3	0.04

Aplique el teorema 4.2 y encuentre la varianza de X.

4.36 Suponga que las probabilidades son 0.4, 0.3, 0.2 y 0.1, respectivamente, de que 0, 1, 2 o 3 fallas de energía eléctrica afecten cierta subdivisión en cualquier año dado. Encuentre la media y la varianza de la variable aleatoria X que representa el número de fallas de energía que afectan esta subdivisión.

4.37 La ganancia de un distribuidor, en unidades de \$5000, para un automóvil nuevo es una variable aleatoria X que tiene la función de densidad que se presenta en el ejercicio 4.2 de la página 113. Encuentre la varianza de X.

4.38 La proporción de individuos que responden cierta encuesta que se manda por correo es una variable aleatoria X, que tiene la función de densidad que se da en el ejercicio 4.14 de la página 113. Encuentre la varianza de X.

4.39 El número total de horas, en unidades de 100 horas, que una familia utiliza una aspiradora en un periodo de un año es una variable aleatoria X, cuya función de densidad se da en el ejercicio 4.15 de la página 114. Encuentre la varianza de X.

4.40 Refiérase al ejercicio 4.14 de la página 113, y encuentre $\sigma^2_{g(X)}$ para la función $g(X) = 3X^2 + 4$.

4.41 Encuentre la desviación estándar de la variable aleatoria $g(X) = (2X + 1)^2$ del ejercicio 4.17 de la página 114.

4.42 Utilice los resultados del ejercicio 4.21 de la página 114, y encuentre la varianza de $g(X) = X^2$, donde X es una variable aleatoria que tiene la función de densidad que se da en el ejercicio 4.12 de la página 113.

4.43 El tiempo, en minutos, para que un avión obtenga vía libre para despegar en cierto aeropuerto es una variable aleatoria $Y = 3X - 2$, donde X tiene la función de densidad

$$f(x) = \begin{cases} \frac{1}{4}e^{-x/4}, & x > 0 \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la media y la varianza de la variable aleatoria Y.

4.44 Encuentre la covarianza de las variables aleatorias X y Y del ejercicio 3.39 de la página 101.

4.45 Encuentre la covarianza de las variables aleatorias X y Y del ejercicio 3.49 de la página 102.

4.46 Encuentre la covarianza de las variables aleatorias X y Y del ejercicio 3.44 de la página 102.

4.47 Refiérase a las variables aleatorias cuya función de densidad conjunta está dada en el ejercicio 3.40 de la página 101, y encuentre la covarianza de X y Y.

4.48 Dada una variable aleatoria X, con desviación estándar σ_X y una variable aleatoria $Y = a + bX$, demuestre que si $b < 0$, el coeficiente de correlación $p_{XY} = -1$, y si $b > 0$, $p_{XY} = 1$.

4.49 Considere la situación del ejercicio 4.32 de la página 115. La distribución del número de imperfecciones por 10 metros de tela sintética está dada por

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

Encuentre la varianza y la desviación estándar del número de imperfecciones.

4.50 En una tarea de laboratorio, si el equipo está funcionando, la función de densidad del resultado observado, X, es

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la varianza y la desviación estándar de X.

4.3 Medias y varianzas de combinaciones lineales de variables aleatorias

Desarrollemos ahora algunas propiedades útiles que simplificarán los cálculos de medias y las varianzas de variables aleatorias que aparecen en los siguientes capítulos. Estas propiedades nos permitirán tratar con las esperanzas matemáticas, en relación con otros parámetros que ya se conocen o que se calculan con facilidad. Todos los resultados que presentamos aquí son válidos para variables aleatorias tanto continuas como discretas. Las demostraciones se dan sólo para el caso continuo. Comenzamos con un teorema y dos corolarios que deberían ser, de forma intuitiva, razonables para el lector.

Teorema 4.5: Si a y b son constantes, entonces,

$$E(aX + b) = aE(X) + b.$$

Prueba: Por la definición de un valor esperado,

$$E(aX + b) = \int_{-\infty}^{\infty} (ax + b)f(x)\, dx = a\int_{-\infty}^{\infty} xf(x)\, dx + b\int_{-\infty}^{\infty} f(x)\, dx.$$

La primera integral de la derecha es $E(X)$; y la segunda integral es igual a 1. Por lo tanto,

$$E(aX + b) = aE(X) + b.$$

Corolario 4.1: Al hacer $a = 0$, vemos que $E(b) = b$.

Corolario 4.2: Al hacer $b = 0$, vemos que $E(aX) = aE(X)$.

Ejemplo 4.15: Al aplicar el teorema 4.5 a la variable aleatoria discreta $f(X) = 2X - 1$, resuelva de nuevo el ejemplo 4.4.

Solución: De acuerdo con el teorema 4.5, escribimos

$$E(2X - 1) = 2E(X) - 1.$$

Así,

$$\mu = E(X) = \sum_{x=4}^{9} xf(x)$$
$$= (4)\left(\frac{1}{12}\right) + (5)\left(\frac{1}{12}\right) + (6)\left(\frac{1}{4}\right) + (7)\left(\frac{1}{4}\right) + (8)\left(\frac{1}{6}\right) + (9)\left(\frac{1}{6}\right) = \frac{41}{6}.$$

Por lo tanto,

$$\mu_{2X-1} = (2)\left(\frac{41}{6}\right) - 1 = \$12.67,$$

como antes.

Ejemplo 4.16: Para resolver de nuevo el ejemplo 4.5, aplique el teorema 4.5 a la variable aleatoria continua $g(X) = 4X + 3$.

Solución: En el ejemplo 4.5 utilizamos el teorema 4.5 para escribir

$$E(4X + 3) = 4E(X) + 3.$$

Así,

$$E(X) = \int_{-1}^{2} x \left(\frac{x^2}{3}\right) dx = \int_{-1}^{2} \frac{x^3}{3} dx = \frac{5}{4}.$$

Por lo tanto,

$$E(4X + 3) = (4)\left(\frac{5}{4}\right) + 3 = 8,$$

como antes.

Teorema 4.6: El valor esperado de la suma o diferencia de dos o más funciones de una variable aleatoria X es la suma o diferencia de los valores esperados de las funciones. Es decir,

$$E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)].$$

Prueba: Por definición,

$$\begin{aligned} E[g(X) \pm h(X)] &= \int_{-\infty}^{\infty} [g(x) \pm h(x)] f(x)\ dx \\ &= \int_{-\infty}^{\infty} g(x) f(x)\ dx \pm \int_{-\infty}^{\infty} h(x) f(x)\ dx \\ &= E[g(X)] \pm E[h(X)]. \end{aligned}$$

Ejemplo 4.17: Sea X una variable aleatoria con la siguiente distribución de probabilidad:

x	0	1	2	3
$f(x)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{6}$

Encuentre el valor esperado de $Y = (X - 1)^2$.

Solución: Al aplicar el teorema 4.6 a la función $Y = (X - 1)^2$, escribimos

$$E[(X - 1)^2] = E(X^2 - 2X + 1) = E(X^2) - 2E(X) + E(1).$$

Del corolario 4.1, $E(1) = 1$, y por cálculo directo

$$E(X) = (0)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{2}\right) + (2)(0) + (3)\left(\frac{1}{6}\right) = 1,$$

y

$$E(X^2) = (0)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{2}\right) + (4)(0) + (9)\left(\frac{1}{6}\right) = 2.$$

De aquí,

$$E[(X-1)^2] = 2 - (2)(1) + 1 = 1.$$

Ejemplo 4.18: La demanda semanal de cierta bebida, en miles de litros, en una cadena de tiendas de abarrotes es una variable aleatoria continua $g(X) = X^2 + X - 2$, donde X tiene la función de densidad

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el valor esperado para la demanda semanal de tal bebida.

Solución: Por el teorema 4.6, escribimos

$$E(X^2 + X - 2) = E(X^2) + E(X) - E(2).$$

Del corolario 4.1, $E(2) = 2$ y, por integración directa,

$$E(X) = \int_1^2 2x(x-1)\,dx = 2\int_1^2 (x^2 - x)\,dx = \frac{5}{3},$$

y

$$E(X^2) = \int_1^2 2x^2(x-1)\,dx = 2\int_1^2 (x^3 - x^2)\,dx = \frac{17}{6}.$$

Entonces,

$$E(X^2 + X - 2) = \frac{17}{6} + \frac{5}{3} - 2 = \frac{5}{2},$$

de manera que la demanda semanal promedio de la bebida en esta cadena de tiendas de abarrotes es de 2500 litros.

Suponga que tenemos dos variables aleatorias X y Y con distribución de probabilidad conjunta $f(x, y)$. Dos propiedades adicionales que serán muy útiles en los capítulos siguientes incluyen los valores esperados de la suma, la diferencia y el producto de estas dos variables aleatorias. Primero, sin embargo, demostremos un teorema sobre el valor esperado de la suma o diferencia de funciones de las variables dadas. Éste, por supuesto, es tan sólo una extensión del teorema 4.6.

Teorema 4.7: El valor esperado de la suma o diferencia de dos o más funciones de las variables aleatorias X y Y es la suma o diferencia de los valores esperados de las funciones. Es decir,

$$E[g(X,Y) \pm h(X,Y)] = E[g(X,Y)] \pm E[h(X,Y)].$$

Prueba: Por la definición 4.2,

$$\begin{aligned} E[g(X,Y) \pm h(X,Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [g(x,y) \pm h(x,y)]f(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy \pm \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)\,dx\,dy \\ &= E[g(X,Y)] \pm E[h(X,Y)]. \end{aligned}$$

Corolario 4.3: Al hacer $g(X, Y) = g(X)$ y $h(X, Y) = h(Y)$, vemos que

$$E[g(X) \pm h(Y)] = E[g(X)] \pm E[h(Y)].$$

Corolario 4.4: Al hacer $g(X, Y) = X$ y $h(X, Y) = Y$, vemos que

$$E[X \pm Y] = E[X] \pm E[Y].$$

Si X representa la producción diaria de algún artículo de la máquina A, y Y la producción diaria de la misma clase de artículo de la máquina B, entonces $X + Y$ representa el número total de artículos que ambas máquinas producen diariamente. El corolario 4.4 establece que la producción promedio diaria para ambas máquinas es igual a la suma de la producción promedio diaria de cada máquina.

Teorema 4.8: Sean X y Y dos variables aleatorias independientes. Entonces,

$$E(XY) = E(X)E(Y).$$

Prueba: Por la definición 4.2,

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)\ dx\ dy.$$

Como X y Y son independientes, escribimos

$$f(x, y) = g(x)h(y),$$

donde $g(x)$ y $h(y)$ son, respectivamente, las distribuciones marginales de X y Y. De aquí,

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyg(x)h(y)\ dx\ dy = \int_{-\infty}^{\infty} xg(x)\ dx \int_{-\infty}^{\infty} yh(y)\ dy \\ &= E(X)E(Y). \end{aligned}$$

Para variables discretas, el teorema 4.8 se ilustra lanzando un dado verde y uno rojo. Con la variable aleatoria X representamos el resultado del dado verde, y con la variable aleatoria Y el resultado del dado rojo. Entonces, XY representa el producto de los números que ocurren en el par de dados. A largo plazo, el promedio de los productos de los números es igual al producto del número promedio que ocurre en el dado verde y el número promedio que ocurre en el dado rojo.

Corolario 4.5: Sean X y Y dos variables aleatorias independientes. Entonces, $\sigma_{XY} = 0$.

Prueba: La demostración puede realizarse con los teoremas 4.4 y 4.8.

Ejemplo 4.19: En la producción de microchips de arseniuro de galio, se sabe que la proporción entre galio y arseniuro es independiente de la producción de un alto porcentaje de obleas manejables, que son los principales componentes de los microchips. Denotemos

con X la proporción de galio y arseniuro, y con Y el porcentaje de microobleas manejables producidas durante un periodo de 1 hora. X y Y son variables aleatorias independientes con la siguiente densidad conjunta

$$f(x,y) = \begin{cases} \frac{x(1+3y^2)}{4}, & 0 < x < 2,\ 0 < y < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Ilustre que $E(XY) = E(X)E(Y)$, como sugiere el teorema 4.8.

Solución: Por definición,

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^2 xyf(x,y)\ dx\ dy = \int_0^1 \int_0^2 \frac{x^2y(1+3y^2)}{4}\ dx\ dy \\ &= \int_0^1 \frac{x^3y(1+3y^2)}{12}\bigg|_{x=0}^{x=2} dy = \int_0^1 \frac{2y(1+3y^2)}{3} dy = \frac{5}{6}, \\ E(X) &= \int_0^1 \int_0^2 xf(x,y)\ dx\ dy = \int_0^1 \int_0^2 \frac{x^2(1+3y^2)}{4}\ dx\ dy \\ &= \int_0^1 \frac{x^3(1+3y^2)}{12}\bigg|_{x=0}^{x=2} dy = \int_0^1 \frac{2(1+3y^2)}{3} dy = \frac{4}{3}, \\ E(Y) &= \int_0^1 \int_0^2 yf(x,y)\ dx\ dy = \int_0^1 \int_0^2 \frac{xy(1+3y^2)}{4}\ dx\ dy \\ &= \int_0^1 \frac{x^2y(1+3y^2)}{8}\bigg|_{x=0}^{x=2} dy = \int_0^1 \frac{y(1+3y^2)}{2} dy = \frac{5}{8}, \end{aligned}$$

De aquí,

$$E(X)E(Y) = \left(\frac{4}{3}\right)\left(\frac{5}{8}\right) = \frac{5}{6} = E(XY).$$

Concluimos esta sección con la demostración de dos teoremas que son útiles para calcular varianzas o desviaciones estándar.

Teorema 4.9: Si a y b son constantes, entonces,

$$\sigma^2_{aX+b} = a^2\sigma^2_X = a^2\sigma^2.$$

Prueba: Por definición,

$$\sigma^2_{aX+b} = E\{[(aX+b) - \mu_{aX+b}]^2\}.$$

Entonces,

$$\mu_{aX+b} = E(aX+b) = a\mu + b$$

por el teorema 4.5. Por lo tanto,

$$\sigma^2_{aX+b} = E[(aX+b-a\mu-b)^2] = a^2E[(X-\mu)^2] = a^2\sigma^2.$$

Corolario 4.6: Al hacer $a = 1$, vemos que

$$\sigma^2_{X+b} = \sigma^2_X = \sigma^2.$$

Corolario 4.7: Al hacer $b = 0$, vemos que

$$\sigma^2_{aX} = a^2\sigma^2_X = a^2\sigma^2.$$

El corolario 4.6 establece que la varianza no cambia si se suma o se resta una constante a la variable aleatoria. La suma o resta de una constante simplemente corre los valores de X a la derecha o a la izquierda; pero no cambia su variabilidad. Sin embargo, si una variable aleatoria se multiplica por una constante o se divide entre una constante, entonces el corolario 4.7 establece que la varianza se multiplica por el cuadrado de la constante o se divide entre el cuadrado de la constante.

Teorema 4.10: Si X y Y son variables aleatorias con distribución de probabilidad conjunta $f(x, y)$, entonces,

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_{XY}.$$

Prueba: Por definición,

$$\sigma^2_{aX+bY} = E\{[(aX + bY) - \mu_{aX+bY}]^2\}.$$

Ahora,

$$\mu_{aX+bY} = E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y,$$

utilizando el corolario 4.4 seguido por el corolario 4.2. Por lo tanto,

$$\begin{aligned}\sigma^2_{aX+bY} &= E\{[a(X - \mu_X) + b(Y - \mu_Y)]^2\} \\ &= a^2E[(X - \mu_X)^2] + b^2E[(Y - \mu_Y)^2] + 2abE[(X - \mu_X)(Y - \mu_Y)] \\ &= a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_{XY}.\end{aligned}$$

Corolario 4.8: Si X y Y son variables aleatorias independientes, entonces,

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y.$$

El resultado que se establece en el corolario 4.8 se obtiene a partir del teorema 4.10, que se relaciona con el corolario 4.5.

Corolario 4.9: Si X y Y son variables aleatorias independientes, entonces,

$$\sigma_{aX-bY} = a^2\sigma^2_X + b^2\sigma^2_Y.$$

El corolario 4.9 se obtiene al reemplazar b por $-b$ en el corolario 4.8. Al generalizar a una combinación lineal de n variables aleatorias independientes, escribimos

Corolario 4.10: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes, entonces,

$$\sigma^2_{a_1X_1+a_2X_2+\cdots+a_nX_n} = a_1^2\sigma^2_{X_1} + a_2^2\sigma^2_{X_2} + \cdots + a_n^2\sigma^2_{X_n}.$$

Ejemplo 4.20: Si X y Y son variables aleatorias con varianzas $\sigma_X^2 = 2$, $\sigma_Y^2 = 4$, y covarianza $\sigma_{XY} = -2$, encuentre la varianza de la variable aleatoria $Z = 3X - 4Y + 8$.

Solución:

$$\begin{aligned}\sigma_Z^2 &= \sigma^2_{3X-4Y+8} = \sigma^2_{3X-4Y} && \text{(por el teorema 4.9)}\\ &= 9\sigma_X^2 + 16\sigma_Y^2 - 24\sigma_{XY} && \text{(por el teorema 4.10)}\\ &= (9)(2) + (16)(4) - (24)(-2) = 130.\end{aligned}$$

Ejemplo 4.21: Denotemos con X y Y la cantidad de dos tipos diferentes de impurezas en un lote de cierto producto químico. Suponga que X y Y son variables aleatorias independientes con varianzas $\sigma_X^2 = 2$ y $\sigma_Y^2 = 3$. Encuentre la varianza de la variable aleatoria $Z = 3X - 2Y + 5$.

Solución:

$$\begin{aligned}\sigma_Z^2 &= \sigma^2_{3X-2Y+5} = \sigma^2_{3X-2Y} && \text{(por el teorema 4.9)}\\ &= 9\sigma_x^2 + 4\sigma_y^2 && \text{(por el teorema 4.10)}\\ &= (9)(2) + (4)(3) = 30.\end{aligned}$$

¿Qué sucede en el caso de una función no lineal?

En los apartados anteriores estudiamos propiedades de funciones lineales de variables aleatorias por razones muy importantes. En los capítulos 8 a 15, mucho de lo que se examina e ilustra son problemas prácticos y del mundo real, en los cuales el analista construye un **modelo lineal** para describir un conjunto de datos y, de esta manera, describir o explicar el comportamiento de un fenómeno científico específico. Así, resulta natural que se encuentren los valores esperados y las varianzas de combinaciones lineales de variables aleatorias. No obstante, hay situaciones en que las propiedades de las funciones **no lineales** de variables aleatorias se vuelven importantes. En efecto, hay muchos fenómenos científicos de naturaleza no lineal, donde el modelado estadístico que utiliza funciones no lineales se vuelve muy importante. De hecho, incluso una función simple de variables aleatorias, digamos, $Z = X/Y$, ocurre con bastante frecuencia en la práctica y a diferencia de las reglas dadas anteriormente en esta sección para los valores esperados de combinaciones lineales de variables aleatorias, no hay una simple regla general. Por ejemplo,

$$E(Z) = E(X/Y) \neq E(X)/E(Y),$$

excepto en circunstancias muy especiales.

El material dado por los teoremas 4.5 a 4.10 y varios corolarios son bastante útiles en cuanto a que no hay restricciones en la forma de la densidad o las funciones de probabilidad, aparte de la propiedad de independencia cuando ésta se requiere, como en los corolarios que siguen al teorema 4.10. Para ilustrar, considere el ejemplo 4.21; la varianza de $Z = 3X - 2Y + 5$ no requiere restricciones en las distribuciones de las cantidades X y Y de los dos tipos de impurezas. Sólo se requere la independencia entre X y Y. Por consiguiente, en verdad tenemos a nuestra disposición la capacidad de encontrar $\mu_{g(X)}$ y $\sigma^2_{g(X)}$ para cualquier función $g(\cdot)$ a partir de los principios iniciales de los teoremas 4.1 y 4.3, donde se supone que se **conoce** la distribución

$f(x)$ correspondiente. Los ejercicios 4.40, 4.41 y 4.42, entre otros, ilustran el uso de tales teoremas. De manera que si $g(x)$ es una función no lineal y se conoce la función de densidad (o función de probabilidad en el caso discreto), $\mu_{g(X)}$ y $\sigma^2_{g(X)}$ puede evaluarse con exactitud. No obstante, como en el caso de las reglas dadas para combinaciones lineales, ¿habría reglas para funciones no lineales que puedan utilizarse cuando no se conoce la forma de la distribución de las variables aleatorias pertinentes?

En general, suponga que X es una variable aleatoria y que $Y = g(x)$. La solución general para $E(Y)$ o $Var(Y)$ puede ser difícil y depende de la complejidad de la función $g(\cdot)$. Sin embargo, hay aproximaciones diponibles que dependen de una aproximación lineal de la función $g(x)$. Por ejemplo, suponga que denotamos $E(X)$ como μ y $Var(X) = \sigma^2_X$. Entonces, una aproximación a las series de Taylor de $g(x)$ alrededor de $X = \mu X$ da

$$g(x) = g(\mu_X) + \left.\frac{\partial g(x)}{\partial x}\right|_{x=\mu_X} (x - \mu_X) + \left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=\mu_X} \frac{(x-\mu_X)^2}{2} + \cdots .$$

Como resultado, si truncamos el término lineal y tomamos el valor esperado de ambos lados, obtenemos $E[g(X)] \approx g(\mu_X)$, que ciertamente es intuitivo y en ciertos casos ofrece una aproximación razonable. No obstante, si incluimos el término de segundo orden de las serie de Taylor, entonces tenemos un ajuste de segundo orden para esta *aproximación de primer orden* como

Aproximación de $E[g(X)]$

$$E[g(X)] \approx g(\mu_X) + \left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=\mu_X} \frac{\sigma^2_X}{2}.$$

Ejemplo 4.22: Dada la variable aleatoria X con media μ_X y varianza σ^2_X, determine la aproximación de segundo orden para $E(e^X)$.

Solución: Como $\frac{\partial e^x}{\partial x} = e^x$ y $\frac{\partial^2 e^x}{\partial x^2} = e^x$, obtenemos $E(e^X) \approx e^{\mu_X}(1 + \sigma^2_X/2)$.

De manera similar, podemos desarrollar una aproximación para $Var[g(x)]$ al tomar la varianza de ambos lados de la expansión de la serie de Taylor de primer orden de $g(x)$.

Aproximación de $Var[g(x)]$

$$Var[g(X)] \approx \left[\frac{\partial g(x)}{\partial x}\right]^2_{x=\mu_X} \sigma^2_X.$$

Ejemplo 4.23: Dada la variable aleatoria X como en el ejemplo 4.22, determine una fórmula aproximada para $Var[g(x)]$.

Solución: De nuevo, $\frac{\partial e^x}{\partial x} = e^x$; por lo que $Var(X) \approx e^{2\mu_X}\sigma^2_X$.

Tales aproximaciones pueden extenderse a las funciones no lineales de más de una variable aleatoria.

Dado un conjunto de variables aleatorias independientes $X_1, X_2, \ldots, X_k$ con medias $\mu_1, \mu_2, \ldots, \mu_k$ y varianzas $\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_k$, respectivamente, sea

$$Y = h(X_1, \mathrm{X}_2, \ldots, \mathrm{X}_k)$$

una función no lineal; entonces tenemos las siguientes aproximaciones para $E(Y)$ y

$Var(Y)$:

$$E(Y) \approx h(\mu_1, \mu_2, \ldots, \mu_k) + \sum_{i=1}^{k} \frac{\sigma_i^2}{2} \left[\frac{\partial^2 h(x_1, x_2, \ldots, x_k)}{\partial x_i^2} \right] \Bigg|_{x_i=\mu_i,\ 1\le i\le k},$$

$$Var(Y) \approx \sum_{i=1}^{k} \left[\frac{\partial h(x_1, x_2, \ldots, x_k)}{\partial x_i} \right]^2 \Bigg|_{x_i=\mu_i,\ 1\le i\le k} \sigma_i^2.$$

Ejemplo 4.24: Considere dos variables aleatorias independientes, X y Z, con medias μ_X, μ_Z y varianzas σ_X^2 y σ_Z^2, respectivamente. Considere una variable aleatoria

$$Y = X/Z.$$

Determine aproximaciones para $E(Y)$ y $Var(Y)$.

Solución: Para $E(Y)$, debemos usar $\frac{\partial y}{\partial x} = \frac{1}{z}$ y $\frac{\partial y}{\partial z} = -\frac{x}{z^2}$. Así,

$$\frac{\partial^2 y}{\partial x^2} = 0, \qquad \text{y} \qquad \frac{\partial^2 y}{\partial z^2} = \frac{2x}{z^3}.$$

Como resultado,

$$E(Y) \approx \frac{\mu_X}{\mu_Z} + \frac{\mu_X}{\mu_Z^3}\sigma_Z^2 = \frac{\mu_X}{\mu_Z}\left(1 + \frac{\sigma_Z^2}{\mu_Z^2}\right),$$

y la aproximación para la varianza de Y está dada por

$$Var(Y) \approx \frac{1}{\mu_Z^2}\sigma_X^2 + \frac{\mu_X^2}{\mu_Z^4}\sigma_Z^2 = \frac{1}{\mu_Z^2}\left(\sigma_X^2 + \frac{\mu_X^2}{\mu_Z^2}\sigma_Z^2\right).$$

4.4 Teorema de Chebyshev

En la sección 4.2 establecimos que la varianza de una variable aleatoria nos dice algo acerca de la variabilidad de las observaciones alrededor de la media. Si una variable aleatoria tiene una varianza o desviación estándar pequeña, esperaríamos que la mayoría de los valores se agruparan alrededor de la media. Por lo tanto, la probabilidad de que una variable aleatoria tome un valor dentro de cierto intervalo alrededor de la media es mayor que para una variable aleatoria similar con una desviación estándar mayor. Si pensamos en la probabilidad en términos de un área, esperaríamos una distribución continua con un valor grande de σ que indique una variabilidad mayor y, por lo tanto, esperaríamos que el área esté más extendida, como en la figura 4.2*a*). Sin embargo, una desviación estándar pequeña debería tener la mayor parte de su área cercana a μ, como en la figura 4.2*b*).

Podemos argumentar lo mismo para una distribución discreta. En el histograma de probabilidad de la figura 4.3*b*), el área se extiende mucho más que en la figura 4.3*a*), lo cual indica una distribución más variable de mediciones o resultados.

El matemático ruso P. L. Chebyshev (1821-1894) descubrió que la fracción del área entre cualesquiera dos valores simétricos alrededor de la media está relacionada con la desviación estándar. Como el área bajo una curva de distribución de probabilidad, o en un histograma de probabilidad, suma 1, el área entre cualesquiera dos números es la probabilidad de que la variable aleatoria tome un valor entre estos números.

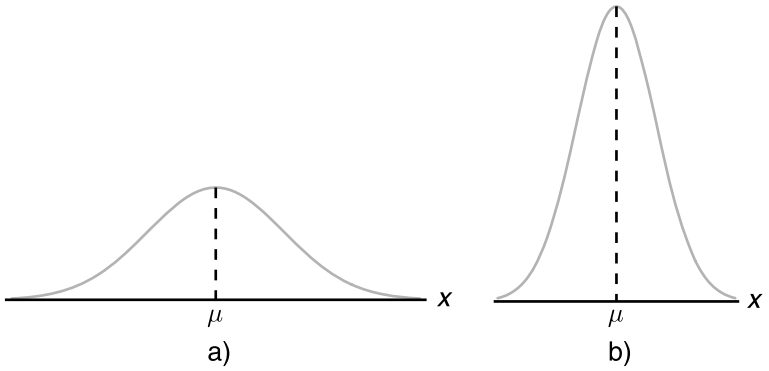


Figura 4.2: Variabilidad de observaciones continuas alrededor de la media.

El siguiente teorema, debido a Chebyshev, da una estimación conservadora de la probabilidad de que una variable aleatoria tome un valor dentro de k desviaciones estándar de su media, para cualquier número real k. Proporcionaremos la demostración sólo para el caso continuo y se deja el caso discreto como ejercicio.

Teorema 4.11:

(Teorema de Chebyshev) La probabilidad de que cualquier variable aleatoria X tome un valor dentro de k desviaciones estándar de la media es al menos $1 - 1/k^2$. Es decir,

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

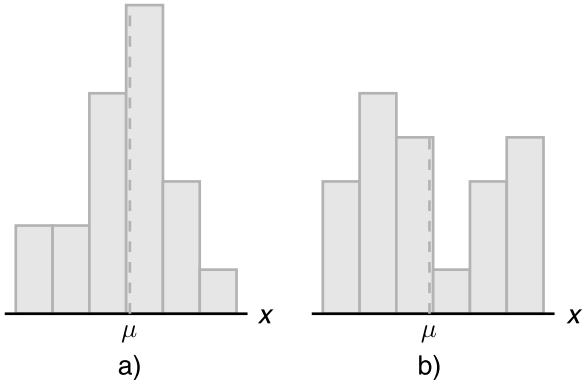


Figura 4.3: Variabilidad de observaciones discretas alrededor de la media.

Prueba: Por nuestra definición anterior de la varianza de X escribimos

$$\begin{aligned}\sigma^2 =& E[(X-\mu)^2] = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx \\ =& \int_{-\infty}^{\mu-k\sigma}(x-\mu)^2 f(x)\,dx + \int_{\mu-k\sigma}^{\mu+k\sigma}(x-\mu)^2 f(x)\,dx \\ &+ \int_{\mu+k\sigma}^{\infty}(x-\mu)^2 f(x)\,dx \\ \geq& \int_{-\infty}^{\mu-k\sigma}(x-\mu)^2 f(x)\,dx + \int_{\mu+k\sigma}^{\infty}(x-\mu)^2 f(x)\,dx,\end{aligned}$$

ya que la segunda de las tres integrales es no negativa. Así, como $|x-\mu| \geq k\sigma$, para cualquier $x \geq \mu + k\sigma$ o $x \leq \mu - k\sigma$, tenemos que $(x-\mu)^2 \geq k^2\sigma^2$ en ambas integrales restantes. Se sigue que

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)\,dx + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x)\,dx,$$

y que

$$\int_{-\infty}^{\mu-k\sigma} f(x)\,dx + \int_{\mu+k\sigma}^{\infty} f(x)\,dx \leq \frac{1}{k^2}.$$

De aquí,

$$P(\mu - k\sigma < X < \mu + k\sigma) = \int_{\mu-k\sigma}^{\mu+k\sigma} f(x)\,dx \geq 1 - \frac{1}{k^2},$$

con lo cual queda establecido el teorema.

Para $k = 2$ el teorema establece que la variable aleatoria X tiene una probabilidad de al menos $1 - 1/2^2 = 3/4$ de caer dentro de dos desviaciones estándar de la media. Es decir, tres cuartos o más de las observaciones de cualquier distribución yacen en el intervalo $\mu \pm 2\sigma$. De manera similar, el teorema indica que al menos ocho novenos de las observaciones de cualquier distribución caen en el intervalo $\mu \pm 3\sigma$.

Ejemplo 4.25: Una variable aleatoria X tiene una media $\mu = 8$, una varianza $\sigma^2 = 9$, y distribución de probabilidad desconocida. Encuentre

a) $P(-4 < X < 20)$,

b) $P(|X-8| \geq 6)$.

Solución: *a*) $P(-4 < X < 20) = P[8-(4)(3) < X < 8+(4)(3)] \geq \frac{15}{16}$.

b) $P(|X-8| \geq 6) = 1 - P(|X-8| < 6) = 1 - P(-6 < X - 8 < 6)$

$$= 1 - P[8-(2)(3) < X < 8+(2)(3)] \leq \frac{1}{4}.$$

El teorema de Chebyshev tiene validez para cualquier distribución de observaciones y, por esta razón, los resultados son generalmente débiles. El valor que el teorema proporciona es sólo un límite inferior. Es decir, sabemos que la probabilidad de una variable aleatoria que cae dentro de dos desviaciones estándar de la media *no*

puede ser menor que 3/4, pero nunca sabemos cuánto podría ser en realidad. Únicamente cuando se conoce la distribución de probabilidad podemos determinar probabilidades exactas. Por esta razón llamamos al teorema resultado de *distribución libre*. Cuando se supongan distribuciones específicas en los siguientes capítulos, los resultados serán menos conservadores. El uso del teorema de Chebyshev se restringe a situaciones donde se desconoce la forma de la distribución.

Ejercicios

4.51 Refiérase al ejercicio 4.35 de la página 122, y encuentre la media y la varianza de la variable aleatoria discreta $Z = 3X - 2$, donde X representa el número de errores por 100 líneas de código.

4.52 Usando los teoremas 4.5 y 4.9, encuentre la media y la varianza de la variable aleatoria $Z = 5X + 3$, donde X tiene la distribución de probabilidad del ejercicio 4.36 en la página 122.

4.53 Suponga que una tienda de abarrotes compra 5 envases de leche descremada al precio de mayoreo de \$1.20 por envase y la vende a \$1.65 por envase. Después de la fecha de caducidad, la leche que no se vende se retira de los anaqueles y el tendero recibe un crédito del distribuidor igual a tres cuartos del precio de mayoreo. Si la distribución de probabilidad de la variable aleatoria X, el número de envases que se venden de este lote, es

x	0	1	2	3	4	5
$f(x)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{3}{15}$

encuentre la utilidad esperada.

4.54 Repita el ejercicio 4.43 de la página 122, con la aplicación de los teoremas 4.5 y 4.9.

4.55 Sea X una variable aleatoria con la siguiente distribución de probabilidad:

x	-3	6	9
$f(x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Encuentre $E(X)$ y $E(X^2)$ y, después, con el uso de estos valores, evalúe $E[(2X + 1)^2]$.

4.56 El tiempo total, medido en unidades de 100 horas, que un adolescente utiliza su estéreo en un periodo de un año es una variable continua X que tiene la función de densidad

$$f(x) = \begin{cases} x, & 0 < x < 1, \\ 2 - x, & 1 \le x < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Utilice el teorema 4.6 para evaluar la media de la variable aleatoria $Y = 60X^2 + 39X$, donde Y es igual al número de kilowatt-hora que gasta al año.

4.57 Si una variable aleatoria X se define de manera que

$$E[(X-1)^2] = 10, \quad E[(X-2)^2] = 6,$$

encuentre μ y σ^2.

4.58 Suponga que X y Y son variables aleatorias independientes que tienen la distribución de probabilidad conjunta

		x	
$f(x,y)$		2	4
y	1	0.10	0.15
	3	0.20	0.30
	5	0.10	0.15

Encuentre

a) $E(2X - 3Y)$;

b) $E(XY)$.

4.59 Use el teorema 4.7 para evaluar $E(2XY^2 - X^2Y)$ para la distribución de probabilidad conjunta que se muestra en la tabla 3.1 de la página 92.

4.60 Se crean 70 nuevos puestos de trabajo en una planta de ensamble automotriz; pero 1000 aspirantes solicitan los 70 puestos. Para seleccionar a los 70 mejores entre los aspirantes, la armadora aplica un examen que cubre habilidad mecánica, destreza manual y capacidad matemática. La calificación media de este examen resulta 60, y las calificaciones tienen una desviación estándar de 6. ¿Una persona que tiene una calificación de 84 puede obtener uno de los trabajos? [*Sugerencia:* Utilice el teorema de Chebyshev.] Suponga que la distribución es simétrica alrededor de la media.

4.61 Una empresa eléctrica fabrica una bombilla de luz de 100 watts que, de acuerdo con las especificaciones escritas en la caja, tiene una vida media de 900 horas con una desviación estándar de 50 horas. A lo más, ¿qué porcentaje de las bombillas no duran al menos 700 horas? Suponga que la distribución es simétrica alrededor de la media.

4.62 Una compañía local fabrica cable telefónico. La longitud promedio del cable es de 52 pulgadas con una desviación estándar de 6.5 pulgadas. A lo más, ¿qué porcentaje del cable telefónico de esta compañía excede 71.5 pulgadas? Suponga que la distribución es simétrica alrededor de la media.

4.63 Suponga que lanza 500 veces un dado balanceado de 10 lados (0, 1, 2,..., 9). Con el teorema de Chebyshev, calcule la probabilidad de que la media de la muestra, X, esté entre 4 y 5.

4.64 Si X y Y son variables aleatorias independientes con varianzas $\sigma_X^2 = 5$ y $\sigma_Y^2 = 3$, encuentre la varianza de la variable aleatoria $Z = -2X + 4Y - 3$.

4.65 Repita el ejercicio 4.64 si X y Y no son independientes y $\sigma_{XY} = 1$.

4.66 Una variable aleatoria X tiene una media $\mu = 12$, una varianza $\sigma^2 = 9$, y una distribución de probabilidad desconocida. Usando el teorema de Chebyshev, estime

a) $P(6 < X < 18)$;

b) $P(3 < X < 21)$.

4.67 Una variable aleatoria X tiene una media $\mu = 10$ y una varianza $\sigma^2 = 4$. Utilizando el teorema de Chebyshev, encuentre

a) $P(|X - 10| \geq 3)$;

b) $P(|X - 10| < 3)$;

c) $P(5 < X < 15)$;

d) el valor de la constante c tal que

$$P(|X - 10| \geq c) \leq 0.04.$$

4.68 Calcule $P(\mu - 2\sigma < X < \mu + 2\sigma)$, donde X tiene la función de densidad

$$f(x) = \begin{cases} 6x(1-x), & 0 < x < 1, \\ 0, & \text{en cualquier otro caso} \end{cases}$$

y compare con el resultado dado por el teorema de Chebyshev.

4.69 Sea X el número que ocurre cuando se lanza un dado rojo y Y el número que sale cuando se lanza un dado verde. Encuentre

a) $E(X + Y)$;

b) $E(X - Y)$;

c) $E(XY)$.

4.70 Suponga que X y Y son variables aleatorias independientes con densidades de probabilidad y

$$g(x) = \begin{cases} \frac{8}{x^3}, & x > 2, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

y

$$h(y) = \begin{cases} 2y, & 0 < y < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre el valor esperado de $Z = XY$.

4.71 Si la función de densidad conjunta de X y Y está dada por

$$f(x,y) = \begin{cases} \frac{2}{7}(x + 2y), & 0 < x < 1,\ 1 < y < 2, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

encuentre el valor esperado de $g(X,Y) = \frac{X}{Y^3} + X^2Y$.

4.72 Sea X el número que ocurre cuando se lanza un dado verde y Y el número que ocurre cuando se lanza un dado rojo. Encuentre la varianza de la variable aleatoria

a) $2X - Y$;

b) $X + 3Y - 5$.

4.73 Considere una variable aleatoria X con función de densidad

$$f(x) = \begin{cases} \frac{1}{5}. & 0 \leq x \leq 5, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre $\mu = E(X)$ y $\sigma^2 = E[(X - \mu)^2]$.

b) Demuestre que el teorema de Chebyshev es válido para $k = 2$ y $k = 3$.

4.74 La potencia P en watts que se disipa en un circuito eléctrico con resistencia R se sabe que está dada por $P = I^2R$, donde I es la corriente en amperes y R es una constante fija en 50 ohms. Sin embargo, I es una variable aleatoria con $\mu_I = 15$ amperes y $\sigma_I^2 = 0.03$ amperes^2. Dé aproximaciones numéricas a la media y a la varianza de la potencia P.

4.75 Considere el ejercicio de repaso 3.79 de la página 105. Las variables aleatorias X y Y representan el número de vehículos que llegan a dos esquinas de calles separadas durante cierto periodo de 2 minutos en el día. La distribución conjunta es

$$f(x,y) = \frac{1}{4^{(x+y)}} \cdot \frac{9}{16},$$

para $x = 0, 1, 2, \ldots$, y $y = 0, 1, 2, \ldots$,

a) Determine $E(X)$, $E(Y)$, $Var(X)$ y $Var(Y)$.

b) Considere $Z = X + Y$, la suma de ambas. Encuentre $E(Z)$ y $Var(Z)$.

4.76 Considere el ejercicio de repaso 3.66 de la página 104. Hay dos líneas de servicio. Las variables aleatorias X y Y son la proporción del tiempo que la línea 1 y la línea 2 están en funcionamiento, respectivamente. La función de densidad de probabilidad conjunta para (X, Y) está dada por

$$f(x,y) = \begin{cases} \frac{3}{2}(x^2 + y^2), & 0 \leq x, y \leq 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Determine si X y Y son independientes o no.

b) Resulta interesante saber algo acerca de la proporción de $Z = X + Y$, la suma de las dos proporciones. Encuentre $E(X + Y)$. También encuentre $E(XY)$.

c) Encuentre $Var(X)$, $Var(Y)$ y $Cov(X,Y)$.

d) Encuentre $Var(X + Y)$.

4.77 El periodo Y en minutos que se requiere para generar un reflejo humano ante el gas lacrimógeno tiene función de densidad

$$f(y) = \begin{cases} \frac{1}{4}e^{-y/4}, & 0 \le y < \infty, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) ¿Cual es el tiempo medio para el reflejo?

b) Encuentre $E(Y^2)$ y $Var(Y)$.

4.78 Una empresa industrial desarrolló una maquina con buen rendimiento de combustible para limpiar alfombras, que deja las alfombras más limpias con mucha rapidez. Interesa una variable aleatoria Y, la cantidad en galones por minuto que ofrece. Se sabe que la función de densidad está dada por

$$f(y) = \begin{cases} 1, & 7 \le y \le 8, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Grafique la función de densidad.

b) Calcule $E(Y)$, $E(Y^2)$ y $Var(Y)$.

4.79 Para la situación que se describe en el ejercicio 4.78, calcule $E(e^Y)$ con el teorema 4.1, es decir, usando

$$E(e^Y) = \int_7^8 e^y f(y)dy.$$

Luego calcule $E(e^Y)$ sin utilizar $f(y)$, sino usando el ajuste de segundo orden para la aproximación de primer orden de $E(e^Y)$. Haga comentarios.

4.80 Considere de nuevo la situación del ejercicio 4.78. Se requiere encontrar $Var(e^Y)$. Utilice los teoremas 4.2 y 4.3, y defina $Z = e^Y$. De manera que, con las condiciones del ejercicio 4.79, encuentre

$$Var(Z) = E(Z^2) - [E(Z)]^2.$$

Luego hágalo sin utilizar $f(y)$, sino más bien empleando la aproximación de las series de Taylor de primer orden para $Var(e^Y)$. ¡Dé sus comentarios!

Ejercicios de repaso

4.81 Demuestre el teorema de Chebyshev cuando X es una variable aleatoria discreta.

4.82 Encuentre la covarianza de las variables aleatorias X y Y que tienen la función de densidad de probabilidad conjunta.

$$f(x,y) = \begin{cases} x + y, & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

4.83 Refiérase a las variables aleatorias cuya función densidad de probabilidad conjunta está dada en el ejercicio 3.47 de la página 102, y encuentre la cantidad promedio de queroseno que queda en el tanque al final del día.

4.84 Suponga que la duración X en minutos de un tipo específico de conversación telefónica es una variable aleatoria con función de densidad de probabilidad

$$f(x) = \begin{cases} \frac{1}{5}e^{-x/5}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Determine la duración media $E(X)$ de este tipo de conversación telefónica.

b) Encuentre la varianza y la desviación estándar de X.

c) Encuentre $E(X + 5)^2$.

4.85 Refiérase a las variables aleatorias cuya función de densidad conjunta está dada en el ejercicio 3.41 de la página 101, y encuentre la covarianza entre el peso de las cremas y el peso de los chiclosos en estas cajas de chocolates.

4.86 Refiérase a las variables aleatorias cuya función de densidad de probabilidad conjunta está dada en el ejercicio 3.41 de la página 101, y encuentre el peso esperado para la suma de las cremas y los chiclosos, si uno compra una caja de tales chocolates.

4.87 Suponga que se sabe que la vida X en horas de un compresor particular tiene la función de densidad

$$f(x) = \begin{cases} \frac{1}{900}e^{-x/900}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre la vida media del compresor.

b) Encuentre $E(X^2)$.

c) Encuentre la varianza y la desviación estándar de la variable aleatoria X.

4.88 Refiérase a las variables aleatorias cuya función de densidad conjunta está dada en el ejercicio 3.40 de la página 101,

a) encuentre μ_X y μ_Y;

b) encuentre $E[(X + Y)/2]$.

4.89 Muestre que $\text{Cov}(aX, bY) = ab\ \text{Cov}(X, Y)$.

4.90 Considere la función de densidad del ejercicio de repaso 4.87. Demuestre que el teorema de Chebyshev es válido para $k = 2$ y $k = 3$.

4.91 Considere la función de densidad conjunta

$$f(x,y) = \begin{cases} \frac{16y}{x^3}, & x > 2,\ 0 < y < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Calcule el coeficiente de correlación p_{XY}.

4.92 Considere las variables aleatorias X y Y del ejercicio 4.65 de la página 135. Calcule p_{XY}.

4.93 La ganancia de un distribuidor en unidades de $5000 en un automóvil nuevo es una variable aleatoria X que tiene la función de densidad

$$f(x) = \begin{cases} 2(1-x), & 0 \le x \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre la varianza de la ganancia del distribuidor.

b) Demuestre que la desigualdad de Chebyshev es válida para $k = 2$ con la función de densidad anterior.

c) ¿Cuál es la probabilidad de que la ganancia exceda $500?

4.94 Considere el ejercicio 4.10 de la página 113. ¿Se puede decir que las calificaciones dadas por los dos expertos son independientes? Explique por qué.

4.95 Los departamentos de marketing y de contabilidad de una compañía determinaron que si la compañía comercializa su producto recientemente desarrollado, la contribución de éste a las utilidades de la empresa durante los próximos 6 meses se describe de la siguiente manera:

Contribuci?n a la utilidad	**Probabilidad**
$5,000 (pérdida)	0.2
$10,000	0.5
$30,000	0.3

¿Cuál es la utilidad esperada de la compañía?

4.96 Un sistema importante funciona como apoyo de un vehículo en el programa espacial. Un solo componente crucial funciona únicamente 85% del tiempo. Para reforzar la confiabilidad del sistema, se decidió que se instalarán 3 componentes paralelos, de manera que el sistema falle sólo si todos fallan. Suponga que los componentes actúan de forma independiente y que son equivalentes en el sentido de que los tres tienen una tasa de éxito de 85%. Considere la variable aleatoria X como el número de componentes de cada tres que fallan.

a) Escriba una función de probabilidad para la variable aleatoria X.

b) ¿Cuál es $E(X)$ (es decir, el número medio de componentes de cada tres que fallan?

c) ¿Cual es $Var(\text{X})$?

d) ¿Cuál es la probabilidad de que el sistema completo sea exitoso?

e) ¿Cuál es la probabilidad de que falle el sistema?

f) Si se desea que el sistema tenga una probabilidad de éxito de 0.99, son suficientes los tres componentes? Si no, ¿cuantos se requerirían?

4.97 En los negocios es importante planear y llevar a cabo investigación para anticipar lo que ocurrirá al final del año. La investigación sugiere que el espectro de utilidades (pérdidas), con sus respectivas probabilidades, es el siguiente:

Utilidades	**Probabilidad**
−$15,000	0.05
$0	0.15
$15,000	0.15
$25,000	0.30
$40,000	0.15
$50,000	0.10
$100,000	0.05
$150,000	0.03
$200,000	0.02

a) ¿Cuál es la utilidad esperada?

b) Determine la desviación estándar de las utilidades.

4.98 Mediante un conjunto de datos y por la amplia investigación se sabe que la cantidad de tiempo, en segundos, que cierto empleado de una compañía llega tarde a trabajar es una variable aleatoria X con función de densidad

$$f(x) = \begin{cases} \frac{3}{4\times 50^3}(50^2 - x^2), & -50 \le x \le 50, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

En otras palabras, él no sólo llega ligeramente retrasado a veces, sino que también puede llegar a trabajar antes de la hora prevista.

a) Encuentre el valor esperado del tiempo en segundos que llega tarde.

b) Encuentre $E(X^2)$.

c) ¿Cuál es la desviación estándar del tiempo en que llega tarde?

4.99 Un camión de carga viaja desde el punto A hasta el punto B, y regresa por la misma ruta diariamente. Hay cuatro semáforos en la ruta. Sea X_1 el número de semáforos en rojo que el camión encuentra cuando va de A a B, y X_2 el número que encuentra en el viaje de

regreso. Los datos recabados durante un periodo largo sugieren que la distribución de probabilidad conjunta para (X_1, X_2) está dada por

	x_2				
x_1	**0**	**1**	**2**	**3**	**4**
0	.01	.01	.03	.07	.01
1	.03	.05	.08	.03	.02
2	.03	.11	.15	.01	.01
3	.02	.07	.10	.03	.01
4	.01	.06	.03	.01	.01

a) Determine la densidad marginal de X_1.

b) Determine la densidad marginal de X_2.

c) Determine la distribución de densidad condicional de X_1 dado que $X_2 = 3$.

d) Determine $E(X_1)$.

e) Determine $E(X_2)$.

f) Determine $E(X_1|X_2 = 3)$.

g) Determine la desviación estándar de X_1.

4.100 Una tienda de abarrotes tiene dos sitios diferentes en sus instalaciones donde los clientes pueden pagar cuando se marchan. Estos dos lugares tienen dos cajas registradoras y dos empleados que atienden a los clientes cuando éstos pagan. Sea X el número de la caja registradora que se utiliza en un momento específico en el sitio 1, y Y en número de la caja registradora que se utiliza en el mismo momento en el sitio 2. La función de probabilidad conjunta está dada por

	y		
x	**0**	**1**	**2**
0	0.12	0.04	0.04
1	0.08	0.19	0.05
2	0.06	0.12	0.30

a) Determine la densidad marginal de X y de Y, así como la distribución de probabilidad de X dado que $Y = 2$.

b) Determine $E(X)$ y $Var(X)$.

c) Determine $E(X|Y = 2)$ y $Var(X|Y = 2)$.

4.101 Considere un transbordador que puede llevar tanto autobuses como automóviles en un recorrido a través de una vía fluvial. Cada viaje cuesta al propietario aproximadamente \$10. La cuota por automóvil es de \$3, y por autobús de \$8. Sean X y Y el número de autobuses y automóviles, respectivamente, que se transportan en un viaje específico. La distribución conjunta de X y Y está dada por

	x		
y	**0**	**1**	**2**
0	0.01	0.01	0.03
1	0.03	0.08	0.07
2	0.03	0.06	0.06
3	0.07	0.07	0.13
4	0.12	0.04	0.03
5	0.08	0.06	0.02

Calcule la utilidad esperada para el viaje del transbordador.

4.102 Como veremos en el capítulo 12, los métodos estadísticos asociados con los modelos lineal y no lineal son muy importantes. De hecho, a menudo las funciones exponenciales se utilizan en una amplia gama de problemas científicos y de ingeniería. Considere un modelo que se ajusta a un conjunto de datos que implica los valores medidos k_1 y k_2, y una respuesta específica Y a las mediciones. El modelo postulado es

$$\hat{Y} = e^{b_0 + b_1 k_1 + b_2 k_2},$$

donde $\hat{Y}$ denota el **valor estimado de** $\boldsymbol{Y}$, k_1 y k_2 son valores fijos y b_0, b_1 y b_2 son **estimados** de constantes y, por lo tanto, son variable aleatorias. Suponga que tales variables aleatorias son independientes y use la fórmula aproximada para la varianza de una función no lineal de más de una variable. Dé una expresión para $Var(\hat{Y})$. Suponga que se conocen las medias de b_0, b_1 y b_2 y son β_0, β_1 y β_2, y también suponga que se conocen las varianzas de b_0, b_1 y b_2 y que son σ_0^2, σ_1^2, y σ_2^2.

4.103 Considere el ejercicio de repaso 3.75 de la página 105, el cual implica Y, la proporción de impurezas en un lote, donde la función de densidad está dada por

$$f(y) = \begin{cases} 10(1-y)^9, & 0 \le y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre el porcentaje esperado de impurezas.

b) Encuentre el valor esperado de la proporción de la calidad del material (es decir, encuentre $E(1 - Y)$).

c) Encuentre la varianza de la variable aleatoria $Z = 1 - Y$.

4.5 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

El material que se cubrió en este capítulo es ampliamente fundamental por naturaleza, muy parecido respecto al del capítulo anterior. Mientras que en el capítulo 3 hicimos una descripción de las características generales de una distribución de probabilidad, en el presente capítulo definimos cantidades importantes o *parámetros* que caracterizan la naturaleza general del sistema. La **media** de una distribución re-

fleja una *tendencia central*, en tanto que la **varianza** o la **desviación estándar** reflejan *variabilidad* en el sistema. Además, una covarianza refleja la tendencia de dos variables aleatorias a "moverse juntas" en un sistema. Estos importantes parámetros serán fundamentales en el estudio de los siguientes capítulos.

El lector debería comprender que el tipo de distribución a menudo está determinado por el contexto científico. Sin embargo, los valores del parámetro necesitan estimarse a partir de datos científicos. Por ejemplo, en el caso del ejercicio de repaso 4.87 el fabricante del compresor podría saber (material que se presentará en el capítulo 6), por su experiencia al conocer el tipo de compresor, que la naturaleza de la distribución es como se indica en el ejercicio. No obstante, la media $\mu = 900$ **se estimaría** a partir de la experimentación con la máquina. Aunque aquí se da por conocido el valor del parámetro de 900, ello no ocurrirá así en situaciones de la vida real sin el uso de datos experimentales. El capítulo 9 se dedica a la **estimación**.

Capítulo 5

Algunas distribuciones de probabilidad discreta

5.1 Introducción y motivación

Sin importar si la distribución de probabilidad discreta se representa de forma gráfica mediante un histograma, en forma tabular o con una fórmula, describe el comportamiento de una variable aleatoria. A menudo, las observaciones que se generan en diferentes experimentos estadísticos tienen el mismo tipo general de comportamiento. En consecuencia, las variables aleatorias discretas asociadas con estos experimentos se pueden describir esencialmente con la misma distribución de probabilidad y, por lo tanto, se representan usando una sola fórmula. De hecho, se necesita sólo un puñado de distribuciones de probabilidad importantes para describir muchas de las variables aleatorias discretas que se encuentran en la práctica.

Tal puñado de distribuciones en realidad describe varios fenómenos aleatorios de la vida real. Por ejemplo, en un estudio sobre la prueba de la eficacia de un nuevo fármaco, el número de pacientes curados entre todos los pacientes que utilizaron tal medicamento sigue aproximadamente una distribución binomial (sección 5.3). En un ejemplo industrial, cuando se probó una muestra de artículos seleccionados de un lote de producción, el número de artículos defectuosos en la muestra, por lo general, puede modelarse como una variable aleatoria hipergeométrica (sección 5.4). En un problema de control estadístico de la calidad, el experimentador señalará un corrimiento en la media del proceso cuando los datos observacionales excedan ciertos límites. El número de muestras requeridas para generar una falsa alarma sigue una distribución geométrica, que es un caso especial de distribución binomial negativa (sección 5.5). Por otro lado, el número de leucocitos de una cantidad fija de una muestra de la sangre de un individuo es comúnmente aleatorio y podría describirse mediante una distribución de Poisson (sección 5.6). En este capítulo, vamos a presentar esas distribuciones que a menudo se utilizan con varios ejemplos.

5.2 Distribución uniforme discreta

La más simple de todas las distribuciones de probabilidad discreta es aquella donde la variable aleatoria toma cada uno de sus valores con una probabilidad idéntica.

Tal distribución de probabilidad se denomina **distribución uniforme discreta**.

Distribución uniforme discreta

Si la variable aleatoria X toma los valores $x_1, x_2, \ldots, x_k$, con idénticas probabilidades, entonces, la distribución uniforme discreta está dada por

$$f(x;k) = \frac{1}{k}, \quad x = x_1, x_2, \ldots, x_k.$$

Utilizamos la notación $f(x; k)$ en vez de $f(x)$ para indicar que la distribución uniforme depende del *parámetro* k.

Ejemplo 5.1: Cuando se selecciona al azar una bombilla de luz de una caja que contiene una bombilla de 40 watts, una de 60, una de 75 y una de 100, cada elemento del espacio muestral $S = \{40, 60, 75, 100\}$ ocurre con probabilidad de 1/4. Por lo tanto, tenemos una distribución uniforme, con

$$f(x;4) = \frac{1}{4}, \quad x = 40, 60, 75, 100.$$

Ejemplo 5.2: Cuando se lanza un dado legal, cada elemento del espacio muestral $S = \{1, 2, 3, 4, 5, 6\}$ ocurre con probabilidad de 1/6. Por lo tanto, tenemos una distribución uniforme, con

$$f(x;6) = \frac{1}{6}, \quad x = 1, 2, 3, 4, 5, 6.$$

La representación gráfica de la distribución uniforme mediante un histograma siempre resulta ser un conjunto de rectángulos con alturas iguales. El histograma para el ejemplo 5.2 se muestra en la figura 5.1.

Teorema 5.1: La media y la varianza de la distribución uniforme discreta $f(x; k)$ son

$$\mu = \frac{1}{k}\sum_{i=1}^{k} x_i, \quad \text{y} \quad \sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \mu)^2.$$

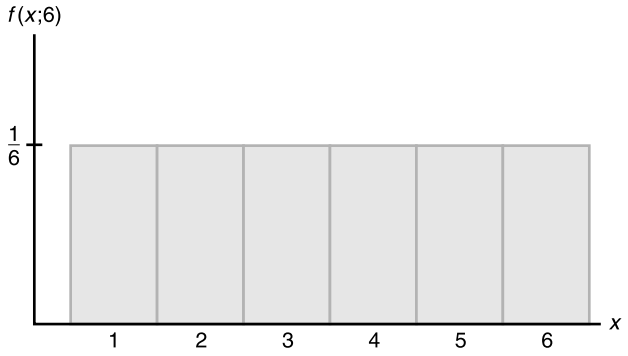


Figura 5.1: Histograma para el lanzamiento de un dado.

Prueba: Por definición,

$$\mu = E(X) = \sum_{i=1}^{k} x_i f(x_i; k) = \sum_{i=1}^{k} \frac{x_i}{k} = \frac{1}{k} \sum_{i=1}^{k} x_i,$$

$$\sigma^2 = E[(X - \mu)^2] = \sum_{i=1}^{k} (x_i - \mu)^2 f(x_i; k) = \frac{1}{k} \sum_{i=1}^{k} (x_i - \mu)^2.$$

Ejemplo 5.3: Con referencia al ejemplo 5.2, encontramos que

$$\mu = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5,$$

y

$$\sigma^2 = \frac{1}{6}[(1 - 3.5)^2 + (2 - 3.5)^2 + \cdots + (6 - 3.5)^2] == \frac{17.5}{6} = \frac{35}{12} = 2.92.$$

5.3 Distribuciones binomial y multinomial

Un experimento a menudo consiste en pruebas repetidas, cada una con dos resultados posibles, los cuales se pueden marcar como **éxito** o **fracaso**. La aplicación más evidente tiene que ver con la prueba de artículos a medida que salen de una línea de ensamble, donde cada prueba o experimento puede indicar si un artículo está defectuoso o no. Podemos elegir definir cualquiera de los resultados como éxito. El proceso se denomina **proceso de Bernoulli**. Cada ensayo se llama **experimento de Bernoulli**. En el ejemplo de extracción de cartas observe que las probabilidades de éxito para los ensayos o pruebas que se repiten cambian si las cartas no se reemplazan. Es decir, la probabilidad de seleccionar una carta de corazones en la primera extracción es 1/4, pero en la segunda es una probabilidad condicional que tiene un valor de 13/51 o 12/51, lo cual depende de si aparece una de corazones en la primera extracción; éste, entonces, ya no se considerará como un conjunto de experimentos de Bernoulli.

El proceso de Bernoulli

Estrictamente hablando, el proceso de Bernoulli debe tener las siguientes propiedades:

1. El experimento consiste en n ensayos que se repiten.
2. Cada ensayo produce un resultado que se puede clasificar como éxito o fracaso.
3. La probabilidad de un éxito, que se denota con p, permanece constante de un ensayo a otro.
4. Los ensayos que se repiten son independientes.

Considere el conjunto de experimentos de Bernoulli donde, de un proceso de ensamble, se seleccionan tres artículos al azar, se inspeccionan y se clasifican como defectuosos o no defectuosos. Un artículo defectuoso se designa como un éxito. El número de éxitos es una variable aleatoria X que toma valores integrales de cero a 3. Los ocho resultados posibles y los valores correspondientes de X son

Resultado	x
NNN	0
NDN	1
NND	1
DNN	1
NDD	2
DND	2
DDN	2
DDD	3

Como los artículos se seleccionan de forma independiente de un proceso que supondremos produce 25% de artículos defectuosos,

$$P(NDN) = P(N)P(D)P(N) = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{9}{64}.$$

Cálculos similares dan las probabilidades para los demás resultados posibles. La distribución de probabilidad de X es, por lo tanto,

x	0	1	2	3
$f(x)$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

El número X de éxitos en n experimentos de Bernoulli se denomina **variable aleatoria binomial**. La distribución de probabilidad de esta variable aleatoria discreta se llama **distribución binomial**, y sus valores se denotarán como $b(x; n, p)$, ya que dependen del número de ensayos y de la probabilidad de éxito en un ensayo dado. Así, para la distribución de probabilidad de X, el número de defectuosos es

$$P(X = 2) = f(2) = b\left(2; 3, \frac{1}{4}\right) = \frac{9}{64}.$$

Generalicemos ahora la ilustración anterior para obtener una fórmula para $b(x; n, p)$. Es decir, deseamos encontrar una fórmula que dé la probabilidad de x éxitos en n ensayos para un experimento binomial. Primero, considere la probabilidad de x éxitos y $n - x$ fracasos en un orden específico. Como los ensayos son independientes, podemos multiplicar todas las probabilidades que corresponden a los diferentes resultados. Cada éxito ocurre con probabilidad p y cada fracaso con probabilidad $q = 1 - p$. Por lo tanto, la probabilidad para el orden específico es $p^x q^{n-x}$. Debemos determinar ahora el número total de puntos muestrales en el experimento que tienen x éxitos y $n - x$ fracasos. Este número es igual al número de particiones de n resultados en dos grupos con x en un grupo y $n - x$ en el otro, y se escribe $\binom{n}{x}$ como se presentó en la sección 2.3. Como estas particiones son mutuamente excluyentes, sumamos las probabilidades de todas las diferentes particiones para obtener la fórmula general o, simplemente, multiplicamos $p^x q^{n-x}$ por $\binom{n}{x}$.

Distribución binomial

Un experimiento de Bernoulli puede tener como resultado un éxito con probabilidad p y un fracaso con probabilidad $q = 1 - p$. Entonces, la distribución de probabilidad de la variable aleatoria binomial X, el número de éxito en n ensayos independientes, es

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

Observe que cuando $n = 3$ y $p = 1/4$, la distribución de probabilidad de X, el número de artículos defectuosos, se escribe como

$$b\left(x; 3, \frac{1}{4}\right) = \binom{3}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{3-x}, \quad x = 0, 1, 2, 3,$$

en vez de la forma tabular de la página 144.

Ejemplo 5.4: La probabilidad de que cierta clase de componente sobreviva a una prueba de choque es 3/4. Encuentre la probabilidad de que sobrevivan exactamente 2 de los siguientes 4 componentes que se prueben.

Solución: Suponga que las pruebas son independientes y como $p = 3/4$ para cada una de las 4 pruebas, obtenemos

$$b\left(2; 4, \frac{3}{4}\right) = \binom{4}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^2 = \left(\frac{4!}{2!\,2!}\right)\left(\frac{3^2}{4^4}\right) = \frac{27}{128}.$$

¿De dónde viene el nombre *binomial*?

La distribución binomial deriva su nombre del hecho de que los $n + 1$ términos en la expansión binomial de $(q + p)^n$ corresponden a los diversos valores de $b(x; n, p)$ para $x = 0, 1, 2, \ldots, n$. Es decir,

$$\begin{aligned}(q+p)^n &= \binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \binom{n}{2}p^2q^{n-2} + \cdots + \binom{n}{n}p^n \\ &= b(0; n, p) + b(1; n, p) + b(2; n, p) + \cdots + b(n; n, p).\end{aligned}$$

Como $p + q = 1$, vemos que

$$\sum_{x=0}^{n} b(x; n, p) = 1,$$

una condición que debe ser válida para cualquier distribución de probabilidad.

Con frecuencia, nos interesamos en problemas donde se necesita encontrar $P(X < r)$ o $P(a \le X \le b)$. Por fortuna, las sumas binomiales

$$B(r; n, p) = \sum_{x=0}^{r} b(x; n, p)$$

están disponibles y se dan en la tabla A.1 del Apéndice para $n = 1, 2, \ldots, 20$, y para valores seleccionados de p entre 0.1 y 0.9. Ilustramos el uso de la tabla A.1 con el siguiente ejemplo.

Ejemplo 5.5: La probabilidad de que un paciente se recupere de una rara enfermedad sanguínea es 0.4. Si se sabe que 15 personas contraen tal enfermedad, ¿cuál es la probabilidad de que *a*) sobrevivan al menos 10, *b*) sobrevivan de 3 a 8, y *c*) sobrevivan exactamente 5?

Solución: Sea X el número de personas que sobreviven.

$$a) \quad P(X \geq 10) = 1 - P(X < 10) = 1 - \sum_{x=0}^{9} b(x; 15, 0.4) = 1 - 0.9662$$
$$= 0.0338.$$
$$b) \quad P(3 \leq X \leq 8) = \sum_{x=3}^{8} b(x; 15, 0.4) = \sum_{x=0}^{8} b(x; 15, 0.4) - \sum_{x=0}^{2} b(x; 15, 0.4)$$
$$= 0.9050 - 0.0271 = 0.8779.$$
$$c) \quad P(X = 5) = b(5; 15, 0.4) = \sum_{x=0}^{5} b(x; 15, 0.4) - \sum_{x=0}^{4} b(x; 15, 0.4)$$
$$= 0.4032 - 0.2173 = 0.1859.$$

Ejemplo 5.6: Una cadena grande de tiendas al detalle compra cierto tipo de dispositivo electrónico de un fabricante. El fabricante indica que la tasa de defectuosos del dispositivo es 3%.

a) El inspector de la cadena elige 20 artículos al azar de un cargamento. ¿Cuál es la probabilidad de que haya al menos 1 artículo defectuoso entre estos 20?

b) Suponga que el detallista recibe 10 cargamentos en un mes y que el inspector aleatoriamente prueba 20 dispositivos por cargamento. ¿Cuál es la probabilidad de que haya 3 cargamentos que contengan al menos un dispositivo defectuoso?

Solución: *a*) Denote con X el número de dispostitivos defectuosos entre los 20. Esta X sigue una distribución $b(x;\ 20,\ 0.30)$. Por consiguiente,

$$P(X \geq 1) = 1 - P(X = 0) = 1 - b(0; 20, 0.03)$$
$$= 1 - 0.03^0(1 - 0.03)^{20-0} = 0.4562.$$

b) En este caso, cada cargamento puede contener al menos un artículo defectuoso o no. Por lo tanto, el hecho de probar el resultado de cada cargamento puede verse como un experimento de Bernoulli con $p = 0.4562$ del inciso *a*). Suponiendo la independencia de un cargamento a otro, y denotando con Y el número de cargamentos que contienen al menos un artículo defectuoso, Y sigue otra distribución binomial $b(y;\ 10,\ 0.4562)$. Por lo tanto, la respuesta a este inciso es

$$P(Y = 3) = \binom{10}{3} 0.4562^3(1 - 0.4562)^7 = 0.1602.$$

Áreas de aplicación

De los ejemplos 5.4, 5.5 y 5.6 debería quedar claro que la distribución binomial encuentra aplicaciones en muchos campos científicos. Un ingeniero industrial está ampliamente interesado en la "proporción de artículos defectuosos" en cierto proceso industrial. A menudo, las mediciones de control de calidad y los esquemas de muestreo para procesos se basan en la distribución binomial, la cual se aplica en cualquier situación industrial donde el resultado de un proceso es dicotómico, y los

resultados del proceso son independientes, y la probabilidad de éxito es constante de una prueba a otra. La distribución binomial también se utiliza de manera extensa en aplicaciones médicas y militares. En ambos casos, un resultado de éxito o de fracaso es importante. Por ejemplo, "cura" o "no cura" es importante en el trabajo farmacéutico; mientras que "dar en el blanco" o "fallar" a menudo es la interpretación del resultado de lanzar un proyectil guiado.

Como la distribución de probabilidad de cualquier variable aleatoria binomial depende sólo de los valores que toman los parámetros n, p y q, parecería razonable suponer que la media y la varianza de una variable aleatoria binomial también dependen de los valores que toman tales parámetros. En realidad, esto es cierto, y en el teorema 5.2 derivamos las fórmulas generales como funciones de n, p y q, que se pueden utilizar para calcular la media y la varianza de cualquier variable aleatoria binomial.

Teorema 5.2: La media y la varianza de la distribución binomial $b(x; n, p)$ son

$$\mu = np \quad \text{y} \quad \sigma^2 = npq.$$

Prueba: Representemos el resultado de la j-ésima prueba mediante la variable aleatoria de Bernoulli I_j, que toma los valores 0 y 1 con probabilidades q y p, respectivamente. Por lo tanto, en un experimento binomial el número de éxitos se escribe como la suma de las n variables indicadoras independientes. De aquí,

$$X = I_1 + I_2 + \cdots + I_n.$$

La media de cualquier I_j es $E(I_j) = (0)(q) + (1)(p) = p$. Por lo tanto, con el corolario 4.4, la media de la distribución binomial es

$$\mu = E(X) = E(I_1) + E(I_2) + \cdots + E(I_n) = \underbrace{p + p + \cdots + p}_{n \text{ términos}} = np.$$

La varianza de cualquier I_j es

$$\sigma_{I_j}^2 = E[(I_j - p)^2] = E(I_j^2) - p^2 = (0)^2(q) + (1)^2(p) - p^2 = p(1-p) = pq.$$

Al extender el corolario 4.10 al caso de n variables independientes, la varianza de la distribución binomial es

$$\sigma_X^2 = \sigma_{I_1}^2 + \sigma_{I_2}^2 + \cdots + \sigma_{I_n}^2 = \underbrace{pq + pq + \cdots + pq}_{n \text{ términos}} = npq.$$

Ejemplo 5.7: Encuentre la media y la varianza de la variable aleatoria binomial del ejemplo 5.5, y después utilice el teorema de Chebyshev (de la página 132) para interpretar el intervalo $\mu \pm 2\sigma$.

Solución: Como el ejemplo 5.5 fue un experimento binomial con $n = 15$ y $p = 0.4$, por el teorema 5.2, tenemos

$$\mu = (15)(0.4) = 6 \quad \text{y} \quad \sigma^2 = (15)\ (0.4)\ (0.6) = 3.6.$$

Al tomar la raíz cuadrada de 3.6, encontramos que $\sigma = 1.897$. Por lo tanto, el intervalo que se requiere es $6 \pm (2)(1.897)$, o de 2.206 a 9.794. El teorema de Chebyshev afirma que el número de recuperaciones entre 15 pacientes sujetos a la enfermedad

mencionada tiene una probabilidad de al menos 3/4 de caer entre 2.206 y 9.794 o, como los datos son discretos, entre 3 y 9 inclusive.

Ejemplo 5.8: Se conjetura que hay impurezas en 30% del total de pozos de agua potable de cierta comunidad rural. Para obtener algún conocimiento del problema, se determina que debería realizarse algún tipo de prueba. Es muy costoso probar todos los pozos del área, por lo que se eligieron 10 aleatoriamente para una prueba.

a) Utilizando la disribución binomial, ¿cuál es la probabilidad de que exactamente tres pozos tengan impurezas considerando que la conjetura es correcta?

b) ¿Cuál es la probabilidad de que más de tres pozos tengan impurezas?

Solución: *a*) Requerimos

$$b(3;10,0.3) = P(X=3) = \sum_{x=0}^{3} b(x;10,0.3) - \sum_{x=0}^{2} b(x;10,0.3)$$
$$= 0.6496 - 0.3828 = 0.2668.$$

b) En este caso necesitamos $P(X > 3) = 1 - 0.6496 = 0.3504$.

Hay soluciones en que el cálculo de probabilidades binomiales nos permitirían obtener inferencias respecto de una población científica después de que se recaban los datos. Se ofrece una ilustración con el siguiente ejemplo.

Ejemplo 5.9: Considere la situación del ejemplo 5.8. La afirmación de que "30% tienen impurezas" es meramente una conjetura del consejo local del agua. Suponga que se eligen 10 pozos de forma aleatoria y se encuentra que 6 contienen impurezas. ¿Qué implica esto respecto de la conjetura? Utilice un enunciado de probabilidad.

Solución: Primero debemos preguntar: "Si la conjetura es correcta, ¿es probable que hubiéramos encontrado 6 o más pozos con impurezas?

$$P(X \geq 6) = \sum_{x=0}^{10} b(x;10,0.3) - \sum_{x=0}^{5} b(x;10,0.3) = 1 - 0.9527 = 0.0473.$$

Como resultado, es muy improbable (4.7% de probabilidad) que 6 o más pozos hubieran resultado impuros si tan sólo 30% de todos ellos son impuros. Esto pone seriamente en duda la conjetura y sugiere que el problema de la impureza es mucho más severo.

Como podrá darse cuenta el lector para este momento, en muchas aplicaciones hay más de dos resultados posibles. Por ejemplo, en el campo de la genética, el color de los conejillos de Indias procreados puede ser rojo, negro o blanco. Con frecuencia, la dicotomía constituida por "defectuoso" y "no defectuoso" en situaciones de ingeniería es en realidad un simplificación excesiva. De hecho, a menudo hay más de dos categorías que caracterizan los artículos o las partes que salen de una línea de ensamble.

Experimentos multinomiales

El experimento binomial se convierte en un **experimento multinomial** si cada prueba tiene más de dos resultados posibles. Por ello, la clasificación de un producto fabricado como ligero, pesado o aceptable, y el registro de los accidentes en cierto

crucero de acuerdo con el día de la semana, constituyen experimentos multinomiales. Extraer una carta de una baraja *con reemplazo* también es un experimento multinomial si los 4 palos son los resultados de interés.

En general, si una prueba dada puede tener como consecuencia cualquiera de los k resultados posibles E_1, $E_2, \ldots,$ E_k con probabilidades p_1, $p_{22}, \ldots,$ p_k, entonces la **distribución multinomial** dará la probabilidad de que E_1 ocurra x_1 veces; E_2 ocurra x_2 veces… y E_k ocurra x_k veces en n pruebas independientes, donde

$$x_1 + x_2 + \cdots + x_k = n.$$

Denotaremos esta distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n).$$

Claramente, $p_1 + p_2 + \cdots + p_k = 1$, pues el resultado de cada ensayo debe ser uno de los k resultados posibles.

Forma general para probabilidades multinomiales

Para derivar la fórmula general, procedemos como en el caso binomial. Como las pruebas son independientes, cualquier orden especificado que produzca x_1 resultados para E_1, x_2 para $E_2, \ldots,$ x_k para E_k ocurrirá con probabilidad $p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$. El número total de órdenes que den resultados similares para las n pruebas es igual al número de particiones de n artículos en k grupos con x_1 en el primer grupo; x_2 en el segundo grupo… y x_k en el k-ésimo grupo. Esto se realiza en

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1!\, x_2! \cdots x_k!}$$

formas. Como todas las particiones son mutuamente excluyentes y ocurren con igual probabilidad, obtenemos la distribución multinomial al multiplicar la probabilidad para un orden específico por el número total de particiones.

Distribución multinomial

Si una prueba dada puede conducir a los k resultados E_1, $E_2, \ldots,$ E_k con probabilidades p_1, $p_2, \ldots,$ p_k, entonces la distribución de probabilidad de las variables aleatorias X_1, $X_2, \ldots,$ X_k, que representa el número de ocurrencias para E_1, $E_2, \ldots, ,$ E_k en n pruebas independientes, es

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

con

$$\sum_{i=1}^{k} x_i = n, \quad \text{y} \qquad \sum_{i=1}^{k} p_i = 1.$$

La distribución multinomial deriva su nombre del hecho de que los términos de la expansión multinomial de $(p_1 + p_2 + \cdots + p_k)^n$ corresponden a todos los posibles valores de

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n).$$

Ejemplo 5.10: La complejidad de las llegadas y las salidas en un aeropuerto es tal que a menudo se utiliza la simulación computarizada para modelar las condiciones "ideales". Para un

aeropuerto específico que contiene tres pistas se sabe que, en el escenario ideal, las siguientes son las probabilidades de que las pistas individuales sean utilizadas por un avión comercial que llega aleatoriamente:

Pista 1: $p_1 = 2/9$,
Pista 2: $p_2 = 1/16$,
Pista 3: $p_3 = 11/18$.

¿Cuál es la probabilidad de que 6 aviones que llegan al azar se distribuyan de la siguiente manera?

Pista 1: 2 aviones,
Pista 2: 1 avión,
Pista 3: 3 aviones.

Solución: Usando la distribución multinomial, tenemos

$$f\left(2,1,3;\frac{2}{9},\frac{1}{6},\frac{11}{18},6\right) = \binom{6}{2,1,3}\left(\frac{2}{9}\right)^2\left(\frac{1}{6}\right)^1\left(\frac{11}{18}\right)^3$$
$$= \frac{6!}{2!\,1!\,3!}\cdot\frac{2^2}{9^2}\cdot\frac{1}{6}\cdot\frac{11^3}{18^3} = 0.1127.$$

Ejercicios

5.1 Se elige a un empleado de un equipo de 10 para supervisar cierto proyecto, mediante la selección de una etiqueta al azar de una caja que contiene 10 etiquetas numeradas del 1 al 10. Encuentre la fórmula para la distribución de probabilidad de X que represente el número en la etiqueta que se saca. ¿Cuál es la probabilidad de que el número que se extrae sea menor que 4?

5.2 Se dan dos altavoces idénticos a doce personas para que escuchen diferencias, si las hubiera. Suponga que estas personas responden sólo adivinando. Encuentre la probabilidad de que tres personas afirmen haber escuchado alguna diferencia entre los dos altavoces.

5.3 Encuentre la media y la varianza de la variable aleatoria X del ejercicio 5.1.

5.4 En cierto distrito de la ciudad la necesidad de dinero para comprar drogas se establece como la razón del 75% de todos los robos. Encuentre la probabilidad de que entre los siguientes cinco casos de robo que se reporten en este distrito,

a) exactamente 2 resulten de la necesidad de dinero para comprar drogas;

b) al menos 3 resulten de la necesidad de dinero para comprar drogas.

5.5 De acuerdo con *Chemical Engineering Progress* (noviembre de 1990), aproximadamente 30% de todas las fallas de operación en las tuberías de plantas químicas son ocasionadas por errores del operador.

a) ¿Cuál es la probabilidad de que de las siguientes 20 fallas en las tuberías al menos 10 se deban a un error del operador?

b) ¿Cuál es la probabilidad de que no más de 4 de 20 fallas se deban al error del operador?

c) Suponga, para una planta específica, que de la muestra aleatoria de 20 de tales fallas, exactamente 5 sean errores de operación. ¿Considera que la cifra de 30% anterior se aplique a esta planta? Comente.

5.6 De acuerdo con una investigación de la Administrative Management Society, la mitad de las compañías estadounidenses dan a sus empleados 4 semanas de vacaciones después de 15 años de servicio en la compañía. Encuentre la probabilidad de que entre 6 compañías encuestadas al azar, el número que da a sus empleados 4 semanas de vacaciones después de 15 años de servicio es

a) cualquiera entre 2 y 5;

b) menor que 3.

5.7 Un prominente médico afirma que 70% de las personas con cáncer pulmonar son fumadores empedernidos. Si su aseveración es correcta,

a) encuentre la probabilidad de que de 10 de tales pacientes con ingreso reciente en un hospital, menos de la mitad sean fumadores empedernidos;

b) encuentre la probabilidad de que de 20 de tales pacientes que recientemente hayan ingresado a un hospital, menos de la mitad sean fumadores empedernidos.

5.8 De acuerdo con un estudio publicado por un grupo de sociólogos de la Universidad de Massachusetts, aproximadamente 60% de los consumidores de Valium en el estado de Massachusetts tomaron Valium por pri-

mera vez a causa de problemas psicológicos. Encuentre la probabilidad de que entre los siguientes 8 consumidores entrevistados de este estado,

a) exactamente 3 comenzarán a tomar Valium por problemas psicológicos;

b) al menos 5 comenzarán a consumir Valium por problemas que no fueron psicológicos.

5.9 Al probar cierta clase de neumático para camión en un terreno accidentado, se encuentra que 25% de los camiones no completaban la prueba de recorrido sin ponchaduras. De los siguientes 15 camiones probados, encuentre la probabilidad de que

a) de 3 a 6 tengan ponchaduras;

b) menos de 4 tengan ponchaduras;

c) más de 5 tengan ponchaduras.

5.10 Según un reportaje publicado en la revista *Parade*, una encuesta a nivel nacional de la Universidad de Michigan a estudiantes universitarios de último año revela que casi 70% desaprueban el consumo de mariguana. Si se seleccionan 12 estudiantes al azar y se les pide su opinión, encuentre la probabilidad de que el número de los que desaprueban fumar mariguana sea

a) cualquier valor entre 7 y 9;

b) a lo más 5;

c) no menos de 8.

5.11 La probabilidad de que un paciente se recupere luego de una delicada operación de corazón es 0.9. ¿Cuál es la probabilidad de que exactamente 5 de los siguientes 7 pacientes intervenidos sobrevivan?

5.12 Un ingeniero de control de tráfico reporta que 75% de los vehículos que pasan por un punto de verificación son de residentes del estado. ¿Cuál es la probabilidad de que menos de 4 de los siguientes 9 vehículos sean de otro estado?

5.13 Un estudio examinó las actitudes nacionales acerca de los antidepresivos. El estudio reveló que aproximadamente 70% cree que "los antidepresivos en realidad no curan nada, sólo disfrazan el problema real". De acuerdo con este estudio, ¿cuál es la probabilidad de que al menos 3 de las siguientes 5 personas seleccionadas al azar tengan esta opinión?

5.14 Se sabe que el porcentaje de victorias para que el equipo de baloncesto Toros de Chicago pasara a las finales en la temporada 1996-1997 fue 87.7. Redondee 87.7 a 90 con la finalidad de utilizar la tabla A. l.

a) ¿Cuál es la probabilidad de que los Toros ganen los primeros 4 de los 7 de la serie final?

b) ¿Cuál es la probabilidad de que los Toros *ganen* toda la serie final?

c) ¿Qué suposición importante se realiza para contestar los incisos *a*) y *b*)?

5.15 Se sabe que 60% de los ratones inoculados con un suero quedan protegidos contra cierta enfermedad. Si se inoculan 5 ratones, encuentre la probabilidad de que

a) ninguno contraiga la enfermedad;

b) menos de 2 contraigan la enfermedad;

c) más de 3 contraigan la enfermedad.

5.16 Suponga que los motores de un avión operan de forma independiente y fallan con probabilidad igual a 0.4. Suponiendo que un avión tiene un vuelo seguro si funcionan al menos la mitad de sus motores, determine si un avión de 4 motores o uno de 2 tiene la probabilidad más alta de un vuelo exitoso.

5.17 Si X representa el número de personas del ejercicio 5.13 que creen que los antidepresivos no curan sino que sólo disfrazan el problema real, encuentre la media y la varianza de X cuando se seleccionan al azar 5 personas y después utilice el teorema de Chebyshev para interpretar el intervalo $\mu \pm 2\sigma$.

5.18 *a*) ¿En el ejercicio 5.9 cuántos de los 15 camiones esperaría que tuviera ponchaduras?

b) De acuerdo con el teorema de Chebyshev, ¿hay una probabilidad de al menos 3/4 de que el número de camiones entre los siguientes 15 que tengan ponchaduras caiga en un intervalo? ¿En cuál?

5.19 Un estudiante que maneja hacia su escuela encuentra un semáforo. Este semáforo permanece verde por 35 segundos, ámbar cinco segundos, y rojo 60 segundos. Suponga que el estudiante va a la escuela toda la semana entre 8:00 y 8:30. Sea X_1 el número de veces que encuentra una luz verde, X_2 el número de veces que encuentra una luz ámbar y X_3 el número de veces que encuentra una luz roja. Encuentre la distribución conjunta de X_1, X_2 y X_3.

5.20 Según el periódico *USA Today* (18 de marzo de 1997) de 4 millones de trabajadores en la fuerza laboral, 5.8% resultó positivo en una prueba de drogas. De quienes resultaron positivos, 22.5% fueron usuarios de cocaína y 54.4% de mariguana.

a) ¿Cuál es la probabilidad de que de 10 trabajadores que resultaron positivos, 2 sean usuarios de cocaína, 5 de mariguana y 3 de otras drogas?

b) ¿Cuál es la probabilidad de que de 10 trabajadores que resultaron positivos, todos sean usuarios de mariguana?

c) ¿Cuál es la probabilidad de que de 10 trabajadores que resultaron positivos, ninguno sea usuario de cocaína?

5.21 La superficie de un tablero circular para dardos tiene un pequeño círculo central llamado ojo de toro y 20 regiones en forma de rebanada de pastel numeradas del 1 a1 20. Asimismo, cada una de estas regiones está

dividida en tres partes, de manera que una persona que lanza un dardo que cae en un número específico obtiene una puntuación igual al valor del número, el doble del número o el triple de éste, según en cuál de las tres partes caiga el dardo. Si una persona atina al ojo de toro con probabilidad de 0.01, atina un doble con probabilidad de 0.10, un triple con probabilidad de 0.05 y no le atina al tablero con probabilidad de 0.02, ¿cuál es la probabilidad de que 7 lanzamientos tengan como resultado ningún ojo de toro, ningún triple, un doble dos veces y dar fuera del tablero?

5.22 De acuerdo con la teoría genética, cierta cruza de conejillos de Indias tendrá crías rojas, negras y blancas con la relación 8:4:4. Encuentre la probabilidad de que entre 8 crías 5 sean rojas, 2 negras y 1 blanca.

5.23 Las probabilidades de que un delegado a cierta convención llegue por avión, autobús, automóvil o tren son, respectivamente, 0.4, 0.2, 0.3 y 0.1. ¿Cuál es la probabilidad de que entre 9 delegados a esta convención seleccionados al azar, 3 lleguen por avión, 3 por autobús, 1 en automóvil y 2 en tren?

5.24 Un ingeniero de seguridad afirma que sólo 40% de todos los trabajadores utilizan cascos de seguridad cuando comen en el lugar de trabajo. Suponga que esta afirmación es cierta, y encuentre la probabilidad de que 4 de 6 trabajadores elegidos al azar utilicen sus cascos mientras comen en el lugar de trabajo.

5.25 Suponga que para un embarque muy grande de chips de circuitos integrados, la probabilidad de falla para cualquier chip es 0.10. Suponga que se cumplen las suposiciones en que se basan las distribuciones binomiales y encuentre la probabilidad de que a lo más 3 chips fallen en una muestra aleatoria de 20.

5.26 Suponga que 6 de 10 accidentes automovilísticos se deben principalmente a que no se respeta el límite de velocidad, y encuentre la probabilidad de que entre 8 accidentes automovilísticos 6 se deban principalmente a no respetar el límite de velocidad

a) mediante el uso de la fórmula para la distribución binomial;

b) usando la tabla binomial.

5.27 Si la probabilidad de que una luz fluorescente tenga una vida útil de al menos 800 horas es 0.9, encuentre las probabilidades de que entre 20 de tales luces

a) exactamente 18 tengan una vida útil de al menos 800 horas;

b) al menos 15 tengan una vida útil de al menos 800 horas;

c) al menos 2 *no* tengan una vida útil de al menos 800 horas.

5.28 Un fabricante sabe que, en promedio, 20% de los tostadores eléctricos que fabrica requerirán reparaciones dentro de 1 año después de su venta. Cuando se seleccionan al azar 20 tostadores, encuentre los números x y y adecuados tales que

a) la probabilidad de que al menos x de ellos requieran reparaciones sea menor que 0.5;

b) la probabilidad de que al menos y de ellos *no* requieran reparaciones sea mayor que 0.8.

5.4 Distribución hipergeométrica

La manera más simple de ver la diferencia entre la distribución binomial de la sección 5.3 y la distribución hipergeométrica está en la forma en que se realiza el muestreo. Los tipos de aplicaciones de la distribución hipergeométrica son muy similares a los de la distribución binomial. Nos interesamos en el cálculo de probabilidades para el número de observaciones que caen en una categoría específica. Sin embargo, en el caso de la binomial, se requiere la independencia entre las pruebas. Como resultado, si se aplica la binomial, digamos, al tomar muestras de un lote de artículos (barajas, lotes de artículos producidos), el muestreo se debe efectuar **con reemplazo** de cada artículo después de que se observe. Por otro lado, la distribución hipergeométrica no requiere independencia y se basa en el muestreo que se realiza **sin reemplazo**.

Las aplicaciones de la distribución hipergeométrica se encuentran en muchas áreas, con gran uso en muestreo de aceptación, pruebas electrónicas y garantía de calidad. Evidentemente, para muchos de estos campos el muestreo se realiza a expensas del artículo que se prueba. Es decir, el artículo se destruye y por ello no se puede reemplazar en la muestra. Así, es necesario un muestreo sin reemplazo. Utilizamos un ejemplo simple con barajas para ilustración.

Si deseamos encontrar la probabilidad de observar 3 cartas rojas en 5 extracciones de una baraja ordinaria de 52 cartas, la distribución binomial de la sección 5.3

no se aplica a menos que cada carta se reemplace y que el paquete se revuelva antes de que se extraiga la siguiente carta. Para resolver el problema de muestreo sin reemplazo, volvamos a plantearnos el problema. Si se sacan 5 cartas al azar, nos interesamos en la probabilidad de seleccionar 3 cartas rojas de las 26 disponibles y 2 negras de las 26 cartas negras de que dispone la baraja. Hay $\binom{26}{3}$ formas de seleccionar 3 cartas rojas, y para cada una de estas formas podemos elegir 2 cartas negras de $\binom{26}{2}$ maneras. Por lo tanto, el número total de formas de seleccionar 3 cartas rojas y 2 negras en cinco extracciones es el producto $\binom{26}{3}\binom{26}{2}$. El número total de formas de seleccionar cualesquiera 5 cartas de las 52 disponibles es $\binom{52}{5}$. Por ello, la probabilidad de seleccionar 5 cartas sin reemplazo de las cuales 3 sean rojas y 2 negras está dada por

$$\frac{\binom{26}{3}\binom{26}{2}}{\binom{52}{5}} = \frac{(26!/3!\,23!)(26!/2!\,24!)}{52!/5!\,47!} = 0.3251.$$

En general, nos interesa la probabilidad de seleccionar x éxitos de los k artículos considerados como éxito y $n - x$ fracasos de los $N - k$ artículos que se consideran fracasos cuando una muestra aleatoria de tamaño n se selecciona de N artículos. Esto se conoce como un **experimento hipergeométrico**; es decir, aquel que posee las siguientes dos propiedades:

1. Se selecciona una muestra aleatoria de tamaño n sin reemplazo de N artículos.
2. k de los N artículos se pueden clasificar como éxitos y $N - k$ se clasifican como fracasos.

El número X de éxitos de un experimento hipergeométrico se denomina **variable aleatoria hipergeométriea**. En consecuencia, la distribución de probabilidad de la variable hipergeométrica se llama **distribución hipergeométrica**, y sus valores se denotan como $h(x; N, n, k)$, debido a que dependen del número de éxitos k en el conjunto N del que seleccionamos n artículos.

Distribución hipergeométrica en el muestro de aceptación

Como en el caso de la distribución binomial, la distribución hipergeométrica encuentra aplicaciones en el muestreo de aceptación, donde lotes del material o las partes se muestrean con la finalidad de determinar si se acepta o no el lote completo.

Ejemplo 5.11: Una pieza específica que se utiliza como dispositivo de injección se vende en lotes de 10. El productor considera que el lote es aceptable si no tiene más de un artículo defectuoso. Algunos lotes se muestrean y el plan de muestreo implica muestreo aleatorio y probar 3 partes de cada 10. Si ninguna de las 3 está defectuosa, se acepta el lote. Comente acerca de la utilidad de este plan.

Solución: Supongamos que el lote es verdaderamente **inaceptable** (es decir, que 2 de cada 10 están defectuosos). La probabilidad de que nuestro plan de muestreo encuentre el lote aceptable es

$$P(X = 0) = \frac{\binom{2}{0}\binom{8}{3}}{\binom{10}{3}} = 0.467.$$

De esta manera, si el lote es en verdad inaceptable con 2 partes defectuosas, este plan de muestreo permitirá su aceptación aproximadamente 47% de las veces. Como resultado, este plan debería considerarse defectuoso.

Generalicemos el ejemplo 5.8 para encontrar una fórmula para $h(x; N, n, k)$. El número total de muestras de tamaño n que se eligen de N artículos es $\binom{N}{n}$. Se supone que estas muestras tienen igual probabilidad. Hay $\binom{k}{x}$ formas de seleccionar x éxitos de los k disponibles, y por cada una de estas formas podemos elegir $n - x$ fracasos en $\binom{N-k}{n-x}$ formas. De esta manera, el número total de muestras favorables entre las $\binom{N}{n}$ muestras posibles está dado por $\binom{k}{x}\binom{N-k}{n-x}$. De aquí, tenemos la siguiente definición.

Distribución hipergeométrica

La distribución de probabilidad de la variable aleatoria hipergeométrica X, el número de éxitos en una muestra aleatoria de tamaño n que se selecciona de N artículos, en los que k se denomina **éxito** y $N - k$ **fracaso**, es

$$h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad \max\{0, n-(N-k)\} \le x \le \min\{n, k\}.$$

El rango de x puede determinarse mediante los tres coefcicentes binomiales en la definición, donde x y $n - x$ no son más que k y $N - k$; respectivamente; y ambas no pueden ser menores que 0. Por lo general, cuando tanto k (el número de éxitos) como $N - k$ (el número de fracasos) son mayores que el tamaño de la muestra n, el rango de una variable aleatoria hipergeométrica será $x = 0, 1, \ldots, n$.

Ejemplo 5.12: Lotes de 40 componentes cada uno se denominan aceptables si no contienen más de tres defectuosos. El procedimiento para muestrear el lote consiste en seleccionar 5 componentes al azar y rechazar el lote si se encuentra un componente defectuoso. ¿Cuál es la probabilidad de que se encuentre exactamente 1 defectuoso en la muestra, si hay 3 defectuosos en todo el lote?

Solución: Si se utiliza la distribución hipergeométrica con $n = 5$, $N = 40$, $k = 3$ y $x = 1$, encontramos que la probabilidad de obtener un defectuoso es

$$h(1; 40, 5, 3) = \frac{\binom{3}{1}\binom{37}{4}}{\binom{40}{5}} = 0.3011.$$

De nueva cuenta, probablemente este plan no sea deseable porque detecta un lote malo (con 3 defectuosos) sólo 30% de las veces.

Teorema 5.3: La media y la varianza de la distribución hipergeométrica $h(x; N, n, k)$ son

$$\mu = \frac{nk}{N}, \quad \text{y} \quad \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right).$$

La demostración para la media se muestra en el apéndice A.25.

Ejemplo 5.13: Volvamos a investigar el ejemplo 3.9. La finalidad de este ejemplo fue ilustrar la noción de una variable aleatoria y el espacio muestral correspondiente. En el ejemplo, tenemos un lote de 100 artículos de los cuales 12 están defectuosos. ¿Cuál es la probabilidad de que haya 3 defectuosos en una muestra de 10?

Solución: Utilizando la función de probabilidad hipergeométrica, tenemos

$$h(3;100,10,12)=\frac{\binom{12}{3}\binom{88}{7}}{\binom{100}{10}}=0.08.$$

Ejemplo 5.14: Encuentre la media y la varianza de la variable aleatoria del ejemplo 5.12, y después utilice el teorema de Chebyshev para interpretar el intervalo $\mu \pm 2\sigma$.

Solución: Como el ejemplo 5.12 fue un experimento hipergeométrico con $N = 40$, $n = 5$ y $k = 3$, entonces por el teorema 5.3, tenemos

$$\mu=\frac{(5)(3)}{40}=\frac{3}{8}=0.375,$$

y

$$\sigma^2=\left(\frac{40-5}{39}\right)(5)\left(\frac{3}{40}\right)\left(1-\frac{3}{40}\right)=0.3113.$$

Al obtener la raíz cuadrada de 0.3113, encontramos que $\sigma = 0.558$. De aquí, el intervalo que se requiere es $0.375 \pm (2)(0.558)$, o de -0.741 a 1.491. El teorema de Chebyshev establece que el número de componentes defectuosos que se obtienen cuando 5 se seleccionan al azar de un lote de 40 componentes, de los que tres son defectuosos, tiene una probabilidad de al menos 3/4 de caer entre -0.741 y 1.491. Es decir, al menos tres cuartos de las veces los 5 componentes incluirán menos de 2 defectuosos.

Relación con la distribución binomial

En este capítulo examinamos varias distribuciones discretas importantes que tienen amplia aplicabilidad, muchas de las cuales se relacionan bien entre sí. El estudiante principiante debería tener una clara comprensión de tales relaciones. Hay una relación interesante entre las distribuciones hipergeométrica y binomial. Como se esperaría, si n es pequeña comparada con N, la naturaleza de los N artículos cambia muy poco en cada prueba. De manera que una distribución binomial puede utilizarse para aproximar la distribución hipergeométrica cuando n es pequeña en comparación con N. De hecho, por regla general la aproximación es buena cuando $\frac{n}{N} \leq 0.05$.

Así, la cantidad $\frac{k}{N}$ juega el papel del parámetro binomial p. Como consecuencia, la distribución binomial se puede ver como una versión de población grande de las distribuciones hipergeométricas. La media y la varianza entonces se obtienen de las fórmulas

$$\mu=np=\frac{nk}{N},\quad \sigma^2=npq=n\cdot\frac{k}{N}\left(1-\frac{k}{N}\right).$$

A1 comparar estas fórmulas con las del teorema 5.3, vemos que la media es la misma mientras que la varianza difiere por un factor de corrección de $(N - n)/(N - 1)$, que es insignificante cuando n es pequeña en relación con N.

Ejemplo 5.15: Un fabricante de neumáticos para automóvil reporta que entre un cargamento de 5000 que se mandan a un distribuidor local, 1000 están ligeramente manchados. Si se compran al azar 10 de estos neumáticos al distribuidor, ¿cuál es la probabilidad de que exactamente 3 estén manchados?

Solución: Como $N = 5000$ es grande en relación con la muestra de tamaño $n = 10$, aproximaremos la probabilidad que se desea usando la distribución binomial. La probabilidad de obtener un neumático manchado es 0.2. Por lo tanto, la probabilidad de obtener exactamente 3 manchados es

$$h(3; 5000, 10, 1000) \approx b(3; 10, 0.2) = \sum_{x=0}^{3} b(x; 10, 0.2) - \sum_{x=0}^{2} b(x; 10, 0.2)$$
$$= 0.8791 - 0.6778 = 0.2013.$$

Por otro lado, la probabilidad exacta es $h(3;\ 5000,\ 10,\ 1000) = 0.2015$.

La distribución hipergeométrica se puede extender para tratar el caso donde los N artículos se pueden dividir en k celdas $A_1,\ A_2, \ldots,\ A_k$ con a_1 elementos en la primera celda, a_2 en la segunda, $\ldots,\ a_k$ elementos en la k-ésima celda. Nos interesamos ahora en la probabilidad de que una muestra aleatoria de tamaño n dé x_1 elementos de A_1, x_2 elementos de $A_2, \ldots,$ y x_k de A_k. Representemos esta probabilidad por

$$f(x_1, x_2, \ldots, x_k; a_1, a_2, \ldots a_k, N, n).$$

Para obtener una fórmula general, notamos que el número total de muestras de tamaño n que se pueden elegir a partir de N artículos es aún $\binom{N}{n}$. Hay $\binom{a_1}{x_1}$ formas de seleccionar x_1 artículos de los que hay en A_1, y para cada uno de éstos podemos elegir x_2 de los de A_2 en $\binom{a_2}{x_2}$ formas. Por lo tanto, podemos seleccionar x_1 artículos de A_1, y x_2 de A_2 en $\binom{a_1}{x_1}\binom{a_2}{x_2}$ formas. Si continuamos de esta forma, podemos seleccionar todos los n artículos que consisten en x_1 de A_1, x_2 de $A_2, \ldots,$ y x_k de A_k en

$$\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k} \text{ formas.}$$

La distribución de probabilidad que se requiere se define ahora como sigue.

Distribución hipergeométrica multivariada

Si N artículos se pueden dividir en las k celdas $A_1,\ A_2, \ldots,\ A_k$ con $a_1,\ a_2, \ldots,$ a_k elementos, respectivamente, entonces la distribución de probabilidades de las variables aleatorias $X_1,\ X_2, \ldots,\ X_k$, que representan el número de elementos que se seleccionan de $A_1,\ A_2, \ldots,\ A_k$ en una muestra aleatoria de tamaño n, es

$$f(x_1, x_2, \ldots, x_k; a_1, a_2, \ldots, a_k, N, n) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}}{\binom{N}{n}},$$

con $\sum_{i=1}^{k} x_i = n$ y $\sum_{i=1}^{k} a_i = N$.

Ejemplo 5.16: Un grupo de 10 individuos se usa para un estudio de caso biológico. El grupo contiene 3 personas con sangre tipo O, 4 con sangre tipo A y 3 con tipo B. ¿Cuál es la probabilidad de que una muestra aleatoria de 5 contenga 1 persona con sangre tipo O, 2 personas con tipo A y 2 personas con tipo B?

Solución: Al usar la extensión de la distribución hipergeométrica con $x_1 = 1$, $x_2 = 2$, $x_3 = 2$, $a_1 = 3$, $a_2 = 4$, $a_3 = 3$, $N = 10$ y $n = 5$, encontramos que la probabilidad que se desea es

$$f(1, 2, 2; 3, 4, 3, 10, 5) = \frac{\binom{3}{1}\binom{4}{2}\binom{3}{2}}{\binom{10}{5}} = \frac{3}{14}.$$

Ejercicios

5.29 Si se reparten 7 cartas de una baraja ordinaria de 52 cartas, ¿cuál es la probabilidad de que

a) exactamente 2 de ellas sean mayores a 10?

b) al menos 1 de ellas sea una reina?

5.30 Para evitar la detección en la aduana, un viajero coloca seis comprimidos con narcóticos en una botella que contiene 9 píldoras de vitamina que son similares en apariencia. Si el oficial de la aduana selecciona 3 de las tabletas al azar para su análisis, ¿cuál es la probabilidad de que el viajero sea arrestado por posesión ilegal de narcóticos?

5.31 El dueño de una casa planta 6 bulbos seleccionados al azar de una caja que contiene 5 bulbos de tulipán y cuatro de narciso. ¿Cuál es la probabilidad de que plante 2 bulbos de narciso y 4 de tulipán?

5.32 De un lote de 10 proyectiles, se seleccionan 4 al azar y se lanzan. Si el lote contiene tres proyectiles defectuosos que no explotarán, ¿cuál es la probabilidad de que

a) los 4 exploten?

b) a lo más 2 fallen?

5.33 Se selecciona al azar un comité de 3 personas a partir de 4 doctores y 2 enfermeras. Escriba una fórmula para la distribución de probabilidad de la variable aleatoria X que representa el número de doctores en el comité. Encuentre $P(2 \leq X \leq 3)$.

5.34 ¿Cuál es la probabilidad de que una mesera se rehuse a servir bebidas alcohólicas a sólo dos menores si ella verifica al azar las identificaciones de 5 estudiantes de entre 9 estudiantes, de los cuales 4 no tienen la edad legal para beber?

5.35 Una compañía está interesada en evaluar su procedimiento de inspección actual en embarques de 50 artículos idénticos. El procedimiento consiste en tomar una muestra de 5 y pasar el embarque si no se encuentran más de 2 defectuosos. ¿Qué proporción de embarques con 20% defectuosos se aceptará?

5.36 Una compañía fabricante utiliza un esquema de aceptación de producción de artículos antes de que se embarquen. El plan tiene dos etapas. Se preparan cajas de 25 artículos para su embarque y se prueba una muestra de 3 en busca de defectuosos. Si se encuentra alguno defectuoso, toda la caja se regresa para verificar e1 100%. Si no se encuentran defectuosos, la caja se embarca.

a) ¿Cuál es la probabilidad de que se embarque una caja que contiene 3 defectuosos?

b) ¿Cuál es la probabilidad de que una caja que contenga sólo 1 artículo defectuoso se regrese para su revisión?

5.37 Suponga que la compañía fabricante del ejercicio 5.36 decide cambiar su esquema de aceptación. Con el nuevo esquema un inspector toma un artículo al azar, lo inspecciona y después lo reemplaza en la caja; un segundo inspector hace lo mismo. Finalmente, un tercer inspector lleva a cabo el mismo procedimiento. La caja no se embarca si cualquiera de los tres encuentra uno defectuoso. Responda el ejercicio 5.36 con este nuevo plan.

5.38 En el ejercicio 5.32, ¿cuántos proyectiles defectuosos se pueden incluir entre los 4 que se seleccionan? Utilice el teorema de Chebyshev para describir la variabilidad del número de proyectiles defectuosos que se incluyen cuando se seleccionan 4 de varios lotes, cada uno de tamaño 10 con 3 proyectiles defectuosos.

5.39 Si a una persona se le reparten varias veces 13 cartas de una baraja ordinaria de 52 cartas, ¿cuántas cartas de corazones por mano puede esperar? ¿Entre cuáles dos valores esperaría que cayera el número de corazones al menos 75% de las veces?

5.40 Se estima que 4000 de los 10,000 residentes con derecho al voto de una ciudad están en contra de un nuevo impuesto sobre ventas. Si se seleccionan al azar 15 votantes y se les pide su opinión, ¿cuál es la probabilidad de que a lo más 7 estén a favor del nuevo impuesto?

5.41 Una ciudad vecina considera una petición de anexión de 1200 residencias contra una subdivisión del condado. Si los ocupantes de la mitad de las residencias objetan la anexión, ¿cuál es la probabilidad de que en una muestra aleatoria de 10 al menos 3 estén a favor de la petición de anexión?

5.42 Entre 150 empleados de IRS en una ciudad grande, sólo 30 son mujeres. Si se eligen al azar 10 de los aspirantes para que proporcionen asistencia libre de impuestos a los residentes de esta ciudad, utilice la aproximación binomial a la hipergeométrica para encontrar la probabilidad de que al menos 3 mujeres se seleccionen.

5.43 Una encuesta a nivel nacional de la Universidad de Michigan a 17,000 estudiantes universitarios de último año revela que casi 70% desaprueba el consumo de mariguana. Si se seleccionan al azar 18 de tales estudiantes y se les pide su opinión, ¿cuál es la probabilidad de que más de 9 pero menos de 14 desaprueben el consumo de mariguana?

5.44 Encuentre la probabilidad de que cuando se le reparta una mano de bridge de 13 cartas tenga 5 de espadas, 2 de corazones, 3 de diamantes y 3 de tréboles.

5.45 Un club de estudiantes extranjeros tiene como miembros a 2 canadienses, 3 japoneses, 5 italianos y 2 alemanes. Si se selecciona al azar un comité de 4, encuentre la probabilidad de que

a) todas las nacionalidades estén representadas;

b) todas las nacionalidades estén representadas excepto los italianos.

5.46 Una urna contiene 3 bolas verdes, 2 azules y 4 rojas. En una muestra aleatoria de 5 bolas, encuentre la probabilidad de que se seleccionen bolas azules y al menos una roja.

5.47 Estudios de población de biología y el ambiente a menudo etiquetan y sueltan a sujetos con la finalidad de estimar el tamaño y el grado de ciertas características en la población. Se capturan 10 animales de una población que se piensa extinta (o cerca de la extinción), se etiquetan y se liberan en cierta región. Después de un periodo se selecciona en la región una muestra aleatoria de 15 animales del tipo. ¿Cuál es la probabilidad de que 5 de estos seleccionados sean animales etiquetados si hay 25 animales de este tipo en la región?

5.48 Una compañía grande tiene un sistema de inspección para los lotes de compresores pequeños que se compran a los vendedores. Un lote típico contiene 15 compresores. En el sistema de inspección se selecciona una muestra aleatoria de 5 y todos se prueban. Suponga que en el lote de 15 hay 2 compresores defectuosos.

a) ¿Cuál es la probabilidad de que para una muestra dada haya 1 compresor defectuoso?

b) ¿Cuál es la probabilidad de que la inspección descubra ambos compresores defectuosos?

5.49 Una fuerza de tarea gubernamental sospecha que algunas fábricas infringen los reglamentos federales contra la contaminación ambiental en cuanto a la descarga de cierto tipo de producto. Veinte empresas están bajo sospecha pero no todas se pueden inspeccionar. Suponga que 3 de las empresas infringen los reglamentos.

a) ¿Cuál es la probabilidad de que la inspección de 5 empresas no encuentre ninguna infracción?

b) ¿Cuál es la probabilidad de que el plan anterior encuentre a dos que infringen el reglamento?

5.50 Cada hora, una máquina llena 10,000 latas de bebida gaseosa, entre las cuales se producen 300 con un llenado insuficiente. Cada hora se elige al azar una muestra de 30 latas y se verifica el número de onzas de gaseosa. Denote con X el número de latas seleccionadas que tiene llenado insuficiente. Encuentre la probabilidad de que habrá al menos una con llenado insuficiente entre las muestreadas.

5.5 Distribuciones binomial negativa y geométrica

Consideremos un experimento donde las propiedades son las mismas que las que se indican para un experimento binomial, con la excepción de que las pruebas se repetirán hasta que ocurra un número *fijo* de éxitos. Por lo tanto, en vez de encontrar la probabilidad de x éxitos en n pruebas, donde n es fija, ahora nos interesa la probabilidad de que ocurra el k-ésimo éxito en la x-ésima prueba. Los experimentos de este tipo se llaman **experimentos binomiales negativos**.

Como ejemplo, considere el uso de un medicamento que se sabe que es efectivo en 60% de los casos en que se utiliza. El uso del medicamento se considerará un éxito si es efectivo al proporcionar algún grado de alivio al paciente. Nos interesa encontrar la probabilidad de que el quinto paciente que experimente alivio sea el séptimo paciente en recibir el medicamento en una semana dada. Designamos éxito con S y fracaso con F, un orden posible para alcanzar el resultado que se desea es $SFSSSFS$, que ocurre con probabilidad

$$(0.6)(0.4)(0.6)(0.6)(0.6)(0.4)(0.6) = (0.6)^5(0.4)^2.$$

Podríamos listar todos los posibles órdenes mediante el reacomodo de las F y las S excepto para el último resultado, que debe ser el quinto éxito. El número total de órdenes posibles es igual al número de particiones de las primeras seis pruebas en dos grupos con 2 fracasos asignados a un grupo y los 4 éxitos asignados al otro grupo. Esto se puede realizar de $\binom{6}{4} = 15$ formas mutuamente excluyentes. De aquí, si X representa el resultado en el que ocurre el quinto éxito, entonces

$$P(X = 7) = \binom{6}{4}(0.6)^5(0.4)^2 = 0.1866.$$

¿Cuál es la variable aleatoria binomial negativa?

El número X de pruebas que genera k éxitos en un experimento binomial negativo se llama **variable aleatoria binomial negativa** y su distribución de probabilidad se llama **distribución binomial negativa**. Como sus probabilidades dependen del número de éxitos que se desean y la probabilidad de un éxito en una prueba dada, las denotaremos con el símbolo $b^*(x;\, k,\, p)$. Para obtener la fórmula general para $b^*(x;\, k,\, p)$, considere la probabilidad de un éxito en la x-ésima prueba precedido por $k - 1$ éxitos y $x - k$ fracasos en un orden específico. Como las pruebas son independientes, podemos multiplicar todas las probabilidades que corresponden a cada resultado que se desea. Cada éxito ocurre con probabilidad p y cada fracaso con probabilidad $q = 1 - p$. Por lo tanto, la probabilidad para el orden específico, que termina en un éxito, es

$$p^{k-1}q^{x-k}p = p^k q^{x-k}.$$

El número total de puntos muestrales en el experimento que termina en un éxito, después de la ocurrencia de $k - 1$ éxitos y $x - k$ fracasos en cualquier orden, es igual al número de particiones de $x - 1$ pruebas en dos grupos con $k - 1$ éxitos que corresponden a un grupo y $x - k$ fracasos que corresponden al otro grupo. Este número se especifica con el término $\binom{x-1}{k-1}$, cada uno es mutuamente excluyente y ocurre con igual probabilidad $p^k q^{x-k}$. Obtenemos la fórmula general al multiplicar $p^k q^{x-k}$ por $\binom{x-1}{k-1}$.

Distribución binomial negativa

Si pruebas independientes repetidas pueden tener como resultado un éxito con probabilidad p y un fracaso con probabilidad $q = 1 - p$, entonces la distribución de probabilidad de la variable aleatoria X, el número de la prueba en la que ocurre el k-ésimo éxito, es

$$b^*(x; k, p) = \binom{x-1}{k-1}p^k q^{x-k}, \quad x = k, k+1, k+2, \ldots.$$

Ejemplo 5.17: En la serie de campeonato de la NBA (Asociación Nacional de Basquetbol), el equipo que gane cuatro juegos de siete será el gandor. Suponga que el equipo A tiene una probabilidad de 0.55 de ganarle al equipo B, y que ambos equipos, A y B, se enfrentarán entre sí en los juegos de campeonato.

a) ¿Cuál es la probabilidad de que el equipo A ganará la serie en seis juegos?

b) ¿Cuál es la probabilidad de que el equipo A ganará la serie?

c) Si ambos equipos se enfrentan entre sí en una serie regional de play-off y el ganador es quien gana tres de cinco juegos, ¿cuál es la probabilidad de que el equipo A ganará un juego de play-off?

Solución: *a)* $b^*\ (6;\ 4,\ 0.55) = \binom{5}{3}0.55^4((1 - 0.55)^{6-4} = 0.1853.$
b) P(el equipo A gana la serie de campeonato) es

$$b^*(4;4,0.55) + b^*(5;4,0.55) + b^*(6;4,0.55) + b^*(7;4,0.55)$$
$$= 0.0915 + 0.1647 + 0.1853 + 0.1668 = 0.6083.$$

c) P(el equipo A gana el juego de playoff) es

$$b^*(3;3,0.55) + b^*(4;3,0.55) + b^*(5;3,0.55)$$
$$= 0.1664 + 0.2246 + 0.2021 = 0.5931.$$

La distribución binomial negativa deriva su nombre del hecho de que cada término de la expansión de $p^k(1 - q)^{-k}$ corresponde a los valores de $b^*(x,\ k,\ p)$ para $x = k,\ k + \text{l},\ k + 2, \ldots$. Si consideramos el caso especial de la distribución binomial negativa donde $k = 1$, tenemos una distribución de probabilidad para el número de pruebas que se requieren para un solo éxito. Un ejemplo sería lanzar una moneda hasta que salga una cara. Nos podemos interesar en la probabilidad de que ocurra la primera cara en el cuarto lanzamiento. La distribución binomial negativa se reduce a la forma

$$b^*(x;1,p) = pq^{x-1}, \quad x = 1, 2, 3, \ldots.$$

Como los términos sucesivos constituyen una progresión geométrica, se acostumbra referirse a este caso especial como la **distribución geométrica** y denotar sus valores con $g(x;\ p)$.

Distribución geométrica

Si pruebas independientes repetidas pueden tener como resultado un éxito con probabilidad p y un fracaso con probabilidad $q = 1 - p$, entonces la distribución de probabilidad de la variable aleatoria X, el número de la prueba en el que ocurre el primer éxito, es

$$g(x;p) = pq^{x-1}, \quad x = 1, 2, 3, \ldots.$$

Ejemplo 5.18: Se sabe que en cierto proceso de fabricación, en promedio, uno de cada 100 artículos está defectuoso. ¿Cuál es la probabilidad de que el quinto artículo que se inspecciona sea el primer defectuoso que se encuentra?

Solución: Utilizando la distribución geométrica con $x = 5$ y $p = 0.01$, tenemos

$$g(5;0.01) = (0.01)(0.99)^4 = 0.0096.$$

Ejemplo 5.19: En "tiempo ocupado" un conmutador telefónico está muy cerca de su capacidad, por lo que los usuarios tienen dificultad al hacer sus llamadas. Puede ser de interés conocer el número de intentos necesario para conseguir un enlace telefónico. Suponga que $p = 0.05$ es la probabilidad de conseguir un enlace durante el tiempo ocupado. Nos interesa conocer la probabilidad de que se necesiten 5 intentos para una llamada exitosa.

Solución: El uso de la distribución geométrica con $x = 5$ y $p = 0.05$ da

$$P(X = x) = g(5;0.05) = (0.05)(0.95)^4 = 0.041.$$

Muy a menudo, en aplicaciones que tienen que ver con la distribución hipergeométrica, la media y la varianza son importantes. Es así que, en el ejemplo 5.19 el número *esperado* de llamadas necesario para lograr un enlace es muy importante.

A continuación se establecen, sin demostración, la media y la varianza de la distribución geométrica.

Teorema 5.4: La media y la varianza de una variable aleatoria que sigue la distribución geométrica son

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}.$$

Aplicaciones de distribuciones binomial negativa y geométrica

Las áreas de aplicación para las distribuciones binomial negativa y geométrica son evidentes cuando nos enfocamos en los ejemplos de esta sección y en los ejercicios que se dedican a tales distribuciones al final de la sección 5.6. En el caso de la distribución geométrica, el ejemplo 5.19 describe una situación donde los ingenieros o administradores intentan determinar cuán ineficiente es un sistema de conmutación telefónica durante periodos ocupados. En este caso, claramente las pruebas ocurren antes de que un éxito represente un costo. Si hay una alta probabilidad de hacer varios intentos antes del enlace, entonces se deberían hacer planes para rediseñar el sistema.

Las aplicaciones de la binomial negativa son similares por naturaleza. Los intentos son costosos en algún sentido y *ocurren en sucesión*. Una alta probabilidad de que se requiera un número "grande" de intentos para experimentar un número fijo de éxitos no es benéfica para el científico ni para el ingeniero. Considere los escenarios de los ejercicios de repaso 5.94 y 5.95. En el ejercicio 5.95 el perforador define cierto nivel de éxitos a partir de los sitios de perforación secuenciales que se hacen en busca de petróleo. Si sólo se llevan seis intentos al momento en que se experimenta el segundo éxito, las utilidades parecen dominar de forma considerable la inversión requerida por la perforación.

5.6 Distribución de Poisson y proceso de Poisson

Los experimentos que dan valores numéricos de una variable aleatoria X, el número de resultados que ocurren durante un intervalo dado o en una región específica, se llaman **experimentos de Poisson**. El intervalo dado puede ser de cualquier longitud, como un minuto, un día, una semana, un mes o incluso un año. Por ello, un experimento de Poisson puede generar observaciones para la variable aleatoria X que representa el número de llamadas telefónicas por hora que recibe una oficina, el número de días que la escuela permanece cerrada debido a la nieve durante el invierno o el número de juegos suspendidos debido a la lluvia durante la temporada de béisbol. La región específica podría ser un segmento de línea, un área, un volumen o quizá una pieza de material. En tales casos X puede representar el número de ratas de campo por acre, el número de bacterias en un cultivo dado o el número de errores mecanográficos por página. Un experimento de Poisson se deriva del **proceso de Poisson** y tiene las siguientes propiedades:

Propiedades del proceso de Poisson

1. El número de resultados que ocurren en un intervalo o región específica es independiente del número que ocurre en cualquier otro intervalo o región del espacio disjunto. De esta forma vemos que el proceso de Poisson no tiene memoria.

2. La probabilidad de que ocurra un solo resultado durante un intervalo muy corto o en una región pequeña es proporcional a la longitud del intervalo o al tamaño de la región, y no depende del número de resultados que ocurren fuera de este intervalo o región.
3. La probabilidad de que ocurra más de un resultado en tal intervalo corto o que caiga en tal región pequeña es insignificante.

El número X de resultados que ocurren durante un experimento de Poisson se llama **variable aleatoria de Poisson** y su distribución de probabilidad se llama **distribución de Poisson**. El número medio de resultados se calcula de $\mu = \lambda t$, donde t es el "tiempo", la "distancia", el "área" o el "volumen" específicos de interés. Como sus probabilidades dependen de λ, la tasa de ocurrencia de los resultados, las denotaremos con el símbolo $P(x; \lambda t)$. La derivación de la fórmula para $p(x; \lambda t)$, que se basa en las tres propiedades de un proceso de Poisson que se listan arriba, está fuera del alcance de este texto. El siguiente concepto se utiliza para calcular probabilidades de Poisson.

Distribución de Poisson

La distribución de probabilidad de la variable aleatoria de Poisson X, que representa el número de resultados que ocurren en un intervalo dado o región específicos se denota con t, es

$$p(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

donde λ es el número promedio de resultados por unidad de tiempo, distancia, área o volumen, y $e = 2.71828\cdots$.

La tabla A.2 contiene la suma de la probabilidad de Poisson

$$P(r; \lambda t) = \sum_{x=0}^{r} p(x; \lambda t),$$

para algunos valores selectos de λt que van de 0.1 a 18. Ilustramos el uso de esta tabla con los siguientes dos ejemplos.

Ejemplo 5.20: Durante un experimento de laboratorio el número promedio de partículas radiactivas que pasan a través de un contador en un milisegundo es 4. ¿Cuál es la probabilidad de que 6 partículas entren al contador en un milisegundo dado?

Solución: Al usar la distribución de Poisson con $x = 6$ y $\lambda t = 4$, y la tabla A.2, tenemos que

$$p(6; 4) = \frac{e^{-4}4^6}{6!} = \sum_{x=0}^{6} p(x; 4) - \sum_{x=0}^{5} p(x; 4) = 0.8893 - 0.7851 = 0.1042.$$

Ejemplo 5.21: El número promedio de camiones-tanque que llega cada día a cierta ciudad portuaria es 10. Las instalaciones en el puerto pueden manejar a lo más 15 camiones-tanque por día. ¿Cuál es la probabilidad de que en un día dado los camiones se tengan que regresar?

Solución: Sea X el número de camiones-tanque que llegan cada día. Entonces, usando la tabla A.2, tenemos

$$P(X > 15) = 1 - P(X \leq 15) = 1 - \sum_{x=0}^{15} p(x; 10) = 1 - 0.9513 = 0.0487.$$

Como la distribución binomial, la distribución de Poisson se utiliza para control de calidad, aseguramiento de calidad y muestreo de aceptación. Además, ciertas distribuciones continuas importantes que se usan en la teoría de confiabilidad y en la teoría de colas dependen del proceso de Poisson. Algunas de estas distribuciones se estudian y se desarrollan en el capítulo 6.

Teorema 5.5: Tanto la media como la varianza de la distribución de Poisson $p(x;\, \lambda t)$ tienen el valor λt.

La demostración de este teorma se encuentra en el Apéndice A.26.

En el ejemplo 5.20, donde $\lambda t = 4$, también tenemos $\sigma^2 = 4$ y, por ello, $\sigma = 2$. Utilizando el teorema de Chebyshev, establecemos que nuestra variable aleatoria tiene una probabilidad de, al menos, 3/4 de caer en el intervalo $\mu \pm 2\sigma = 4 \pm (2)(2)$, o de 0 a 8. Por lo tanto, concluimos que, al menos, tres cuartos de las veces el número de partículas radiactivas que entran al contador estará en cualquier valor entre 0 y 8 durante un milisegundo dado.

Distribución de Poisson como forma limitante de la binomial

Considerando los tres principios del proceso de Poisson debería ser evidente de que la distribución de Poisson se relaciona con la distribución binomial. Aunque la de Poisson, por lo general, encuentra aplicaciones en problemas de espacio y tiempo como se ilustra con los ejemplos 5.20 y 5.21, se puede ver como una forma limitante de la distribución binomial. En el caso de la binomial, si n es bastante grande y p es pequeña, las condiciones comienzan a simular las implicaciones *de espacio continuo o región temporal* del proceso de Poisson. La independencia entre las pruebas de Bernoulli en el caso binomial es consistente con la propiedad 2 del proceso de Poisson. Si se hace al parámetro p cercano a cero se relaciona con la propiedad 3 del proceso de Poisson. De hecho, si n es grande y p es cercana a 0, se puede usar la distribución de Poisson, con $\mu = np$, para aproximar probabilidades binomiales. Si p es cercana a 1, aún podemos utilizar la distribución de Poisson para aproximar probabilidades binomiales mediante el intercambio de lo que definimos como éxito y fracaso, y así cambiamos p a un valor cercano a 0.

Teorema 5.6: Sea X una variable aleatoria binomial con distribución de probabilidad $b(x;\, n,\, p)$. Cuando $n \to \infty$, $p \to 0$, y $np \overset{n\to\infty}{\longrightarrow} \mu$ permanece constante,

$$b(x; n, p) \overset{n\to\infty}{\longrightarrow} p(x; \mu).$$

La demostración de este teorema se encuentra en el Apéndice A.27.

Naturaleza de la función de probabilidad de Poisson

Al igual que muchas distribuciones discretas y continuas, la forma de la distribución se vuelve cada vez más simétrica, incluso con forma de campana, conforme la media se hace más grande. La figura 5.2 ilustra lo anterior. Tenemos gráficas de la función de probabilidad para $\mu = 0.1$, $\mu = 2$ y finalmente $\mu = 5$. Observe la cercanía con la simetría conforme μ se vuelve tan grande como 5. Una condición similar existe para la distribución binomial como se ilustrará en el lugar adecuado más adelante en este texto.

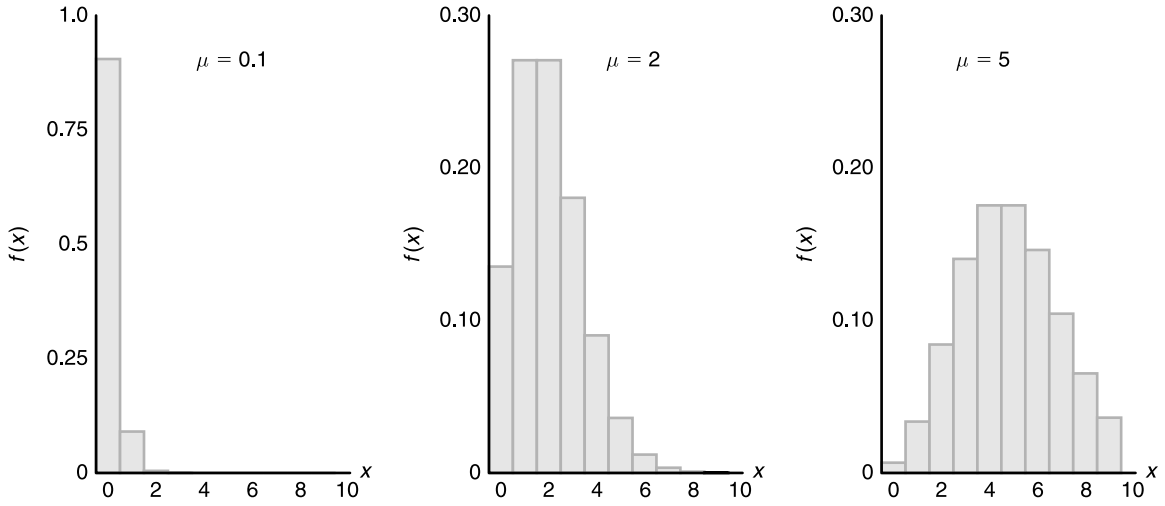


Figura 5.2: Funciones de densidad de Poisson para medias diferentes.

Ejemplo 5.22: En ciertas instalaciones industriales los accidentes ocurren con muy poca frecuencia. Se sabe que la probabilidad de un accidente en cualquier día dado es 0.005 y los accidentes son independientes entre sí.

a) ¿Cuál es la probabilidad de que en cualquier periodo dado de 400 días habrá un accidente en un día?

b) ¿Cuál es la probabilidad de que haya a lo más tres días con un accidente?

Solución: Sea X una variable aleatoria binomial con $n = 400$ y $p = 0.005$. Así, $np = 2$. Con la aproximación de Poisson,

a) $P(X = 1) = e^{-2}2^1 = 0.271$ y

b) $P(X \leq 3) = \sum_{x=0}^{3} e^{-2}2^x/x! = 0.857.$

Ejemplo 5.23: En un proceso de fabricación donde se manufacturan productos de vidrio ocurren defectos o burbujas, lo cual ocasionalmente deja a la pieza indeseable para su venta. Se sabe que, en promedio, 1 de cada 1000 de estos artículos que se producen tiene una o más burbujas. ¿Cuál es la probabilidad de que una muestra aleatoria de 8000 tenga menos de 7 artículos con burbujas?

Solución: Éste es en esencia un experimento binomial con $n = 8000$ y $p = 0.001$. Como p es muy cercana a cero y n es bastante grande, haremos la aproximación con la distribución de Poisson utilizando

$$\mu = (8000)\ (0.001) = 8.$$

De aquí, si X representa el número de burbujas, tenemos

$$P(X < 7) = \sum_{x=0}^{6} b(x; 8000, 0.001) \approx p(x; 8) = 0.3134.$$

Ejercicios

5.51 La probabilidad de que una persona, que vive en cierta ciudad, tenga un perro se estima en 0.3. Encuentre la probabilidad de que la décima persona entrevistada al azar en esta ciudad sea la quinta que tiene un perro.

5.52 Un científico inocula a varios ratones, uno a la vez, con el germen de una enfermedad hasta que encuentra a 2 que contraen la enfermedad. Si la probabilidad de contraer la enfermedad es 1/6, ¿cuál es la probabilidad de que se requieran 8 ratones?

5.53 El estudio de un inventario determina que, en promedio, las demandas de un artículo particular en un almacén se realizan 5 veces al día. ¿Cuál es la probabilidad de que en un día dado se pida este artículo

a) más de 5 veces?

b) ninguna vez?

5.54 Encuentre la probabilidad de que una persona que lanza una moneda obtenga

a) la tercera cara en el séptimo lanzamiento;

b) la primera cara en el cuarto lanzamiento.

5.55 Tres personas lanzan una moneda legal y el disparejo paga los cafés. Si todas las monedas tienen el mismo resultado, se lanzan de nuevo. Encuentre la probabilidad de que se necesiten menos de 4 lanzamientos.

5.56 De acuerdo con un estudio publicado por un grupo de sociólogos de la Universidad de Massachusetts, en Estados Unidos cerca de dos tercios de los 20 millones de personas que consumen Valium son mujeres. Suponga que esta cifra es una estimación válida, y encuentre la probabilidad de que en un día dado la quinta prescripción de Valium que da un médico sea

a) la primera que prescribe Valium para una mujer;

b) la tercera que prescribe Valium para una mujer.

5.57 La probabilidad de que un estudiante para piloto apruebe el examen escrito para obtener una licencia de piloto privado es 0.7. Encuentre la probabilidad de que el estudiante aprobará el examen

a) en el tercer intento;

b) antes del cuarto intento.

5.58 En promedio en cierto crucero ocurren tres accidentes de tránsito por mes. ¿Cuál es la probabilidad de que para cualquier mes dado en este crucero

a) ocurran exactamente 5 accidentes?

b) ocurran menos de 3 accidentes?

c) ocurran al menos 2 accidentes?

5.59 Una secretaria comete dos errores por página, en promedio. ¿Cuál es la probabilidad de que en la siguiente página cometa

a) 4 o más errores?

b) ningún error.

5.60 Cierta área del este de Estados Unidos resulta, en promedio, afectada por 6 huracanes al año. Encuentre la probabilidad de que para cierto año esta área resulte afectada por

a) menos de 4 huracanes;

b) cualquier cantidad entre 6 a 8 huracanes.

5.61 Suponga que la probabilidad de que una persona dada crea un rumor acerca de las transgresiones de cierta actriz famosa es 0.8. ¿Cuál es la probabilidad de que

a) la sexta persona en escuchar este rumor sea la cuarta en creerlo?

b) la tercera persona en escuchar este rumor sea la primera en creerlo?

5.62 El número promedio de ratas de campo por acre en un campo de 5 acres de trigo se estima en 12. Encuentre la probabilidad de que se encuentren menos de 7 ratas de campo

a) en un acre dado;

b) en 2 de los siguientes 3 acres que se inspeccionen.

5.63 El chef de un restaurante prepara una ensalada revuelta que contiene, en promedio, 5 vegetales. Encuentre la probabilidad de que la ensalada contenga más de 5 vegetales

a) en un día dado;

b) en 3 de los siguientes 4 días;

c) por primera vez en abril el día 5.

5.64 La probabilidad de que una persona muera de cierta infección respiratoria es 0.002. Encuentre la probabilidad de que mueran menos de 5 de los siguientes 2000 infectados de esta forma.

5.65 Suponga que, en promedio, 1 persona en 1000 comete un error numérico al preparar su declaración de impuestos. Si se seleccionan 10,000 formas al azar y se examinan, encuentre la probabilidad de que 6, 7 u 8 de las formas contengan un error.

5.66 Se sabe que la probabilidad de que un estudiante de una preparatoria local presente escoliosis (curvatura de la espina dorsal) es 0.004. De los siguientes 1875 estudiantes que se revisen en búsqueda de escoliosis, encuentre la probabilidad de que

a) menos de 5 presenten el problema;

b) 8, 9 o 10 presenten el problema.

5.67 *a*) Encuentre la media y la varianza de la variable aleatoria X, que representa el número de personas entre 2000 que mueren de la infección respiratoria del ejercicio 5.64.

b) De acuerdo con el teorema de Chebyshev, ¿hay una probabilidad de al menos 3/4 de que el número de personas que morirán entre las 2000 infectadas caiga dentro de un intervalo? ¿De cuál?

5.68 *a*) Encuentre la media y la varianza de la variable aleatoria X, que representa el número de personas entre 10,000 que cometen un error al preparar su declaración de impuestos del ejercicio 5.55.

b) De acuerdo con el teorema de Chebyshev, ¿hay una probabilidad de al menos 8/9 de que el número de personas que cometerán errores al preparar sus declaraciones de impuestos entre 10,000 esté dentro de un intervalo? ¿De cuál?

5.69 Un fabricante de automóviles se preocupa por una falla en el mecanismo de freno de un modelo específico. La falla puede causar en raras ocasiones una catástrofe a alta velocidad. Suponga que la distribución del número de automóviles por año que experimentará la falla es una variable aleatoria de Poisson con $\lambda = 5$.

a) ¿Cuál es la probabilidad de que, a lo más, 3 automóviles por año sufran una catástrofe?

b) ¿Cuál es la probabilidad de que más de 1 automóvil por año experimente una catástrofe?

5.70 Los cambios en los procedimientos de los aeropuertos requieren una planeación considerable. Los índices de llegadas de los aviones son factores importantes que deben tomarse en cuenta. Suponga que los aviones pequeños llegan a cierto aeropuerto, de acuerdo con un proceso de Poisson, con un índice de 6 por hora. De esta manera, el parámetro de Poisson para las llegadas en un periodo de horas es $\mu = 6t$.

a) ¿Cuál es la probabilidad de que exactamente 4 aviones pequeños lleguen durante un periodo de 1 hora?

b) ¿Cuál es la probabilidad de que al menos 4 lleguen durante un periodo de 1 hora?

c) Si definimos un día laboral como 12 horas, ¿cuál es la probabilidad de que al menos 75 aviones pequeños lleguen durante un día?

5.71 El número de clientes que llegan por hora a ciertas instalaciones de servicio automotriz se supone que sigue una distribución de Poisson con media $\lambda = 7$.

a) Calcule la probabilidad de que más de 10 clientes lleguen en un periodo de 2 horas.

b) ¿Cuál es el número medio de llegadas durante un periodo de 2 horas?

5.72 Considere el ejercicio 5.66. ¿Cuál es el número medio de estudiantes que fallan en el examen?

5.73 La probabilidad de que una persona muera cuando contrae una infección por virus es 0.001. De los siguientes 4000 infectados con virus, ¿cuál es el número medio que morirá?

5.74 Una compañía compra lotes grandes de cierta clase de dispositivo electrónico. Se utiliza un método que rechaza un lote si se encuentran 2 o más unidades defectuosas en una muestra aleatoria de 100 unidades.

a) ¿Cuál es el número medio de unidades defectuosas que se encuentran en una muestra de 100 unidades si el lote tiene 1% de defectuosas?

b) ¿Cuál es la varianza?

5.75 En el caso de cierto tipo de alambre de cobre, se sabe que, en promedio, ocurren 1.5 fallas por milímetro. Suponiendo que el número de fallas es una variable aleatoria de Poisson, ¿cuál es la probabilidad de que no ocurran fallas en cierta proción de alambre con longitud de 5 milímetros? ¿Cuál es el número medio de fallas en una porción de 5 milímetros de longitud?

5.76 Los baches en ciertas carreteras pueden ser un problema grave y tener la necesidad constante de repararse. Con un tipo específico de terreno y mezcla de concreto, la experiencia sugiere que hay, en promedio, 2 baches por milla después de cierta cantidad de uso. Se supone que el proceso de Poisson se aplica a la variable aleatoria "número de baches".

a) ¿Cuál es la probabilidad de que no más de un bache aparezca en un tramo de una milla?

b) ¿Cuál es la probabilidad de que no más de 4 baches ocurrirán en un tramo dado de 5 millas?

5.77 En ciudades grandes los administradores de los hospitales se preocupan por la cuestión del tráfico de personas en las salas de urgencias de los nosocomios. Para un hospital específico en una ciudad grande, el personal disponible no puede alojar el tráfico de pacientes cuando hay más de 10 casos de emergencia en una hora dada. Se supone que la llegada del paciente sigue un proceso de Poisson y los datos históricos sugieren que, en promedio, llegan 5 emergencias cada hora.

a) ¿Cuál es la probabilidad de que en una hora dada el personal no pueda alojar más al tráfico?

b) ¿Cuál es la probabilidad de que más de 20 emergencias lleguen durante un turno de 3 horas del personal?

5.78 En las revisiones de equipaje en el aeropuerto se sabe que 3% de la gente isnpeccionada lleva objetos cuestionables en su equipaje. ¿Cuál es la probabilidad de que una serie de 15 personas cruce sin problemas antes de que se atrape a un individuo con un objeto cuestionable? ¿Cuál es el número esperado en una fila que pasa antes de que se detenga a un invividuo?

5.79 La tecnología cibernética generó un ambiente donde los "robots" funcionan con el uso de microproce-

sadores. La probabilidad de que un robot falle durante cualquier turno de 6 horas es 0.10. ¿Cuál es la probabilidad de que un robot funcionará durante al menos 5 turnos antes de fallar?

5.80 Se sabe que la tasa de rechazo en las encuestas telefónicas es de aproximadamente 20%. Un reportaje del periódico indica que se encuestaron a 50 personas antes de que la primera rechazara.

a) Comente acerca de la validez del reportaje. Utilice una probabilidad en su argumnento.

b) ¿Cuál es el número esperado de personas encuestadas antes de un rechazo?

Ejercicios de repaso

5.81 Durante un proceso de producción se seleccionan al azar 15 unidades cada día de la línea de ensamble para verificar el porcentaje de defectuosos. A partir de información histórica se sabe que la probabilidad de tener una unidad defectuosa es 0.05. En cualquier momento en que se encuentran dos o más unidades defectuosas en la muestra de 15, el proceso se detiene. Este procedimiento se utiliza para proporcionar una señal en caso de que aumente la probabilidad de unidades defectuosas.

a) ¿Cuál es la probabilidad de que en un día dado el proceso de producción se detenga? (Suponga 5% de unidades defectuosas.)

b) Suponga que la probabilidad de una unidad defectuosa aumenta a 0.07. ¿Cuál es la probabilidad de que en algún día dado el proceso de producción no se detenga?

5.82 Una máquina automática de soldar se considera para la producción. Se considerará para su compra si es exitosa en 99% de sus soldaduras. De otra manera, no se considerará eficiente. Se lleva a cabo la prueba de un prototipo que realizará 100 soldaduras. La máquina se aceptará para la producción si no falla en más de 3 soldaduras.

a) ¿Cuál es la probabilidad de que se rechace una buena máquina?

b) ¿Cuál es la probabilidad de que se acepte una máquina ineficiente con 95% de soldaduras exitosas?

5.83 Una agencia de renta de automóviles en un aeropuerto local tiene disponibles 5 Ford, 7 Chevrolet, 4 Dodge, 3 Honda y 4 Toyota. Si la agencia selecciona al azar 9 de estos automóviles para transportar delegados desde el aeropuerto hasta el centro de convenciones del centro de la ciudad, encuentre la probabilidad de que se utilicen 2 Ford, 3 Chevrolet, 1 Dodge, 1 Honda y 2 Toyota.

5.84 Las llamadas de servicio llegan a un centro de mantenimiento de acuerdo con un proceso de Poisson con un promedio de 2.7 llamadas por minuto. Encuentre la probabilidad de que

a) no más de 4 llamadas lleguen en cualquier minuto;

b) lleguen menos de 2 llamadas en cualquier minuto;

c) lleguen más de 10 llamadas en un periodo de 5 minutos.

5.85 Una empresa de electrónica afirma que la proporción de unidades defectuosas de cierto proceso es 5%. Un comprador tiene un procedimiento estándar para inspeccionar 15 unidades que selecciona al azar de un lote grande. En una ocasión específica, el comprador encuentra 5 artículos defectuosos.

a) ¿Cuál es la probabilidad de esta ocurrencia, dado que la afirmación de 5% de defectuosos es correcta?

b) ¿Cuál sería su reacción si fuera el comprador?

5.86 Un dispositivo electrónico de conmutación ocasionalmente falla y podría ser necesario su reemplazo. Se sabe que el dispositivo es satisfactorio si, en promedio, no comete más de 0.20 errores por hora. Se elige un periodo particular de cinco horas como "prueba" del dispositivo. Si no ocurre más de 1 error, el dispositivo se considera satisfactorio.

a) ¿Cuál es la probabilidad de que un dispositivo satisfactorio se considere que no lo es sobre la base de la prueba? Suponga que existe un proceso de Poisson.

b) ¿Cuál es la probabilidad de que un dispositivo se acepte como satisfactorio cuando, de hecho, el número medio de errores es 0.25? De nuevo, suponga que existe un proceso de Poisson.

5.87 Una compañía, por lo general, compra lotes grandes de cierta clase de dispositivo electrónico. Se utiliza un método que rechaza un lote, si se encuentran dos o más unidades defectuosas en una muestra aleatoria de 100 unidades.

a) ¿Cuál es la probabilidad de rechazar un lote que tiene 1% de unidades defectuosas?

b) ¿Cuál es la probabilidad de aceptar un lote que tiene 5% de unidades defectuosas?

5.88 El propietario de una farmacia local sabe que, en promedio, llegan a su farmacia 100 personas cada hora.

a) Encuentre la probabilidad de que en un periodo dado de 3 minutos nadie entre a la farmacia.

b) Encuentre la probabilidad de que en un periodo dado de 3 minutos entren más de 5 personas a la farmacia.

5.89 a) Suponga que lanza 4 dados. Encuentre la probabilidad de que obtenga al menos un 1.

b) Suponga que lanza 24 veces 2 dados. Encuentre la probabilidad de que obtenga al menos uno (1, l), es decir, que lanza "ojos de serpiente" [NOTA: La probabilidad del inciso *a*) es mayor que la del inciso *b*).]

5.90 Suponga que se venden 500 billetes de lotería. Entre ellos, 200 billetes pagan al menos el costo del billete. Suponga ahora que compra 5 billetes. Encuentre la probabilidad de que gane al menos el costo de 3 billetes.

5.91 Las imperfecciones en las tarjetas de circuitos y los chips para computadora se prestan por sí mismos a tratamiento estadístico. Para un tipo particular de tarjeta la probabilidad de falla de un diodo es 0.03. Suponga que una tarjeta de circuitos contiene 200 diodos.

a) ¿Cuál es el número medio de fallas entre los diodos?

b) ¿Cuál es la varianza?

c) La tarjeta funcionará si no hay diodos defectuosos. ¿Cuál es la probabilidad de que una tarjeta funcione?

5.92 El comprador potencial de un motor particular requiere (entre otras cuestiones) que el motor encienda exitosamente 10 veces consecutivas. Suponga que la probabilidad de un encendido exitoso es 0.990. Supongamos que los resultados de intentos de encendido son independientes.

a) ¿Cuál es la probabilidad de que el motor sea aceptado después de sólo 10 encendidos?

b) ¿Cuál es la probabilidad de que se realicen 12 intentos de encendido durante el proceso de aceptación?

5.93 El esquema de aceptación para comprar lotes que contienen un número grande de baterías consiste en probar no más de 75 baterías seleccionadas al azar, y rechazar un lote si falla una sola batería. Suponga que la probabilidad de una falla es 0.001.

a) ¿Cuál es la probabilidad de que se acepte un lote?

b) ¿Cuál es la probabilidad de que se rechace un lote en la 20a. prueba?

c) ¿Cuál es la probabilidad de que se rechace en 10 o menos pruebas?

5.94 Una compañía perforadora de pozos petroleros se arriesga en varios sitios, y su éxito o fracaso es independiente de un sitio a otro. Suponga que la probabilidad de éxito en cualquier sitio específico es 0.25.

a) ¿Cuál es la probabilidad de que un perforador barrene 10 sitios y tenga 1 éxito?

b) El perforador cree que irá a la bancarrota si perfora 10 veces antes de que ocurra el primer éxito. ¿Cuáles son las perspectivas del perforador para la bancarrota?

5.95 Considere la información del ejercicio de repaso 5.94. El perforador cree que "dará en el clavo" si el segundo éxito ocurre en o antes del sexto intento. ¿Cuál es la probabilidad de que el perforador "dé en el clavo"?

5.96 Una pareja de esposos decide que continuarán teniendo hijos hasta que tengan dos hombres. Suponiendo que $P(\text{hombre}) = 0.5$, ¿cuál es la probabilidad de que su segundo hombre sea su cuarto hijo?

5.97 Por los investigadores se sabe que 1 de cada 100 personas es portadora del gen que lleva a la herencia de cierta enfermedad crónica. A partir de una muestra aleatoria de 1000 individuos, ¿cuál es la probabilidad de que menos de 7 individuos porten el gen? Utilizando la aproximación de Poisson, ¿cuál es el número medio aproximado de personas de cada 1000 que portan el gen?

5.98 Un proceso de manufactura produce piezas para componentes electrónicos. Se supone con fundamento que la probabilidad de una pieza defectuosa es 0.01. Durante una prueba de esta suposición, se muestrearon al azar 500 artículos y se observaron 15 defectuosos de cada 500.

a) ¿Cuál es su repuesta ante la suposición de que el proceso es 1% defectuosos? Asegúrese de que una probabilidad calculada acompaña su comentario.

b) Con la suposición de un proceso 1% defectuoso, ¿cuál es la probabilidad de que sólo se encontrarían 3 defectuosos?

c) Resuelva de nueva cuenta los incisos *a*) y *b*) utilizando la aproximación de Poisson.

5.99 Un proceso de manufactura produce artículos en lotes de 50. Se dispone de planes de muestreo en los cuales los lotes se apartan periódicamente y se exponen a cierto tipo de inspección. Por lo genral, se supone que la proporción de defectuosos en el proceso es muy pequeña. También es importante para la compañía que los lotes que contienen defectuosos sean un evento raro. En la actualidad el plan de inspección para la compañía consiste en periódicamente muestrear al azar 10 de cada 50 artículos en un lote y, si no hay defectuosos, no se hace ninguna intervención al proceso.

a) Suponga que se elige un lote al azar, y 2 de cada 50 están defectuosos. ¿Cuál es la probabilidad de que al menos 1 en la muestra de 10 del lote esté defectuoso?

b) A partir de su respuesta en el inciso *a*), comente sobre la calidad de este plan de muestreo.

c) ¿Cuál es el número medio de defectuosos encontrados en cada 10?

5.100 Considere la situación del ejercicio de repaso 5.99. Se ha determinado que el plan de muestreo debería ser lo sifucientemente amplio como para que haya una probabilidad alta de, digamos, 0.9, de que si hay tantos como 2 defectuosos en el lote de 50 que se muestrean, al menos 1 se encontrará en el muestreo. Con

tales restricciones, ¿cuántos de los 50 deberían muestrearse?

5.101 Homeland Security y la tecnología de defensa de misiles hacen que seamos capaces de detectar proyectiles o misiles ofensivos. Para que la defensa sea exitosa se requieren múltiples pantallas de radar. Suponga que se determina que hay tres pantallas independientes para operar y que la probabilidad de cualquiera detectará un misil ofensivo es 0.8. En efecto, si ninguna pantalla detecta un misil ofensivo, el sistema no será confiable y deberá reemplazarsre.

a) ¿Cuál es la probabilidad de que un misil ofensivo no será detectado por cualesquiera de las tres pantallas?

b) ¿Cuál es la probabilidad de que el misil no será detectado por solo una pantalla?

c) ¿Cuál es la probabilidad de que será detectado por al menos dos de las tres pantallas.

5.102 Considere el ejercicio de repaso 5.101. Suponga que es importante que el sistema general sea tan perfecto como sea posible. Suponiendo que la calidad de las pantallas es la que se indica en el ejercicio de repaso 5.101,

a) ¿cuántas se requieren para asegurarse de que la probabilidad de que el misil pase sin ser detectado sea 0.0001?

b) Suponga que se decide quedarse con sólo 3 pantallas e intentar mejorar la capacidad de detección de las mismas. ¿Cuál debe ser la eficacia individual de las pantallas (es decir, la probabilidad de detección), para alcanzar la eficacia que se requiere en el inciso *a*)?

5.103 Regrese al ejercicio de repaso 5.99*a*). Vuelva a calcular la probabilidad de usar la distribución binomial. Comente.

5.104 En cierto departamento de estadística en el país hay dos vacantes. Cinco individuos las solicitan. Dos de ellos tienen habilidad en modelos lineales y uno tiene habilidad en probabilidad aplicada. Al comite de selección se le indicó elegir a los dos miembros aleatoriamente.

a) ¿Cuál es la probabilidad de que los dos seleccionados sean quienes tienen habilidad en modelos lineales?

b) ¿Cuál es la probabilidad de que de los dos elegidos, uno tenga habilidad en modelos lineales y el otro en probabilidad aplicada?

5.105 El fabricante de un triciclo para niños ha recibido quejas por los frenos defectuosos en el producto. De acuerdo con el diseño del producto y bastantes pruebas preliminares, se determinó que la probabilidad de que el tipo de defecto en la queja era 1 en 10,000 (es decir, .0001). Después de una minuciosa investigación de las quejas, se determinó que durante cierto periodo se eligieran aleatoriamente 200 productos de la producción, de los cuales 5 tuvieron defecto en los frenos.

a) Comente sobre la reclamación "1 en 10,000" del fabricante. Utilice un argumento probabilístico. Use la distribución binomial para sus cálculos.

b) Haga el trabajo utilizando la aproximación de Poisson.

5.7 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

Las distribuciones discretas estudiadas en este capítulo ocurren con mucha frecuencia en los escenarios de la ingeniería y las ciencias biológica y física, como, evidentemente, lo sugieren los ejemplos y los ejercicios. En el caso de las distribuciones binomial y de Poisson, los planes de muestreo industrial y muchos de los criterios de ingeniería se determinan con base en ambas distribuciones. Esto también es el caso para la distribución hipergeométrica. Mientras que las distribuciones binomial negativa y geométrica se utilizan en menor grado, también tienen aplicaciones. En específico, una variable aleatoria binomial negativa puede verse como una mezcla de variables aleatorias gamma y de Poisson. (La distribución gamma se estudiará en el siguiente capítulo.)

A pesar de la vasta utilidad que estas distribuciones tienen en aplicaciones de la vida real, pueden utilizarse de manera incorrecta, a menos que el científico sea prudente y tome las debidas precauciones. Desde luego, cualquier cálculo de probabilidad para las distribuciones que se estudiaron en este capítulo se realiza bajo el supuesto de que se conoce el valor del parámatro. Las aplicaciones del mundo real a menudo resultan en un valor del parámetro que "se desplaza" debido a factores que son difíciles de controlar en el proceso, o debido a las intervenciones en el proceso

que no se toman en cuenta. Por ejemplo, en el ejercicio de repaso 5.81 se utiliza "información histórica". No obstante, ¿el proceso actual es el mismo que aquel en que se recabaron los datos históricos? El uso de la distribución de Poisson puede sufrir incluso más por esta dificultad. Por ejemplo, considere el ejercicio de repaso 5.84. Las preguntas de los incisos *a*), *b*) y *c*) se basan en el uso de $\mu = 2.7$ llamadas por minuto. Con base en los registros históricos, éste es el número de llamadas que se realizan "en promedio". Pero en ésta y muchas otras aplicaciones de la distribución de Poisson, hay "tiempos de baja actividad" y "tiempos ocupados", de manera que se espera que haya momentos en que las condiciones para el proceso de Poisson quizá parezcan cumplirse, cuando en realidad no se cumplen. Así, los cálculos de probabilidades pueden ser incorrectos. En el caso de la binomial la suposición que podría fallar en ciertas aplicaciones (además de la falta de constancia de p) es la suposición de independencia, estipulando que los experimentos de Bernoulli deben ser independientes.

Una de las aplicaciones incorrectas más célebres de la distribución binomial ocurrió en la temporada de béisbol de 1961, cuando Mickey Mantle y Roger Maris se enfrascaron en una batalla amistosa por romper el récord de todos los tiempos Babe Ruth de 60 home-runs. En el artículo de una revista famosa se hizo una predicción con base en la teoría de la probabilidad y se predijo que Mantle rompería el record de acuerdo con un cálculo de una probabilidad mayor con el uso de la distribución binomial. El error clásico cometido fue la elección de las estimaciones del parámetro p (uno para cada jugador) con base en la frecuencia histórica relativa de home-runs a lo largo de sus carreras. Maris, a diferencia de Mantle, no había sido un jonronero prodigio antes de 1961, de manera que su "estimado" de p fue bastante bajo. Como resultado, la probabilidad calculada para romper el récord fue bastante alta para Mantle y baja para Maris. El resultado final: Mantle fracasó al intentar romper el récord y Maris sí lo logró.

Capítulo 6

Algunas distribuciones continuas de probabilidad

6.1 Distribución uniforme continua

En estadística una de las distribuciones continuas más simples es la **distribución uniforme continua**. Esta distribución se caracteriza por una función de densidad que es "plana" y, por ello, la probabilidad es uniforme en un intervalo cerrado, digamos $[A, B]$. Aunque las aplicaciones de la distribución uniforme continua no son tan abundantes como lo son para otras distribuciones que se presentan en este capítulo, resulta apropiado para el principiante comenzar esta introducción a las distribuciones continuas con la distribución uniforme.

Distribución uniforme

La función de densidad de la variable aleatoria uniforme continua X en el intervalo $[A, B]$ es

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Se debe destacar al lector que la función de densidad forma un rectángulo con base $B - A$ y **altura constante** $\frac{1}{B-A}$. Como resultado, la distribución uniforme a menudo se llama **distribución rectangular**. En la figura 6.1 se muestra la función de densidad para una variable aleatoria uniforme en el intervalo [1, 3].

Resulta sencillo calcular las probabilidades para la distribución uniforme debido a la naturaleza simple de la función de densidad. Sin embargo, note que la aplicación de esta distribución se basa en la suposición de que es constante la probabilidad de caer en un intervalo de longitud fija dentro de $[A, B]$.

Ejemplo 6.1: Suponga que una sala de conferencias grande se puede reservar para cierta compañía por no más de cuatro horas. Sin embargo, el uso de la sala de conferencias es tal que muy a menudo tienen conferencias largas y cortas. De hecho, se puede suponer que la duración X de una conferencia tiene una distribución uniforme en el intervalo [0, 4].

a) ¿Cuál es la función de densidad de la probabilidad?

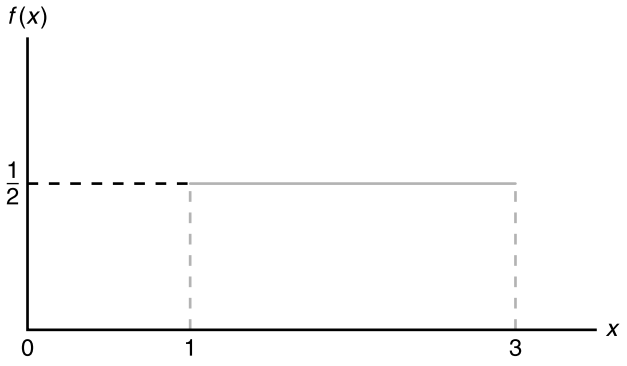


Figura 6.1: Función de densidad para una variable aleatoria en el intervalo [1, 3].

b) ¿Cuál es la probabilidad de que cualquier conferencia dada dure al menos 3 horas?

Solución: *a*) La función de densidad apropiada para la variable aleatoria distribuida uniformemente X en esta situación es

$$f(x) = \begin{cases} \frac{1}{4}, & 0 \leq x \leq 4, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

b) $P[X \geq 3] = \int_3^4 \frac{1}{4}\, dx = \frac{1}{4}$.

Teorema 6.1: La media y la varianza de la distribución uniforme son

$$\mu = \frac{A+B}{2}, \quad \text{y} \quad \sigma^2 = \frac{(B-A)^2}{12}.$$

Las demostraciones de los teoremas se dejan al lector. Véase el ejercicio 6.20 de la página 187.

6.2 Distribución normal

La distribución continua de probabilidad más importante en todo el campo de la estadística es la **distribución normal**. Su gráfica, que se denomina **curva normal**, es la curva con forma de campana de la figura 6.2, la cual describe aproximadamente muchos fenómenos que ocurren en la naturaleza, la industria y la investigación. Las mediciones físicas en áreas como los experimentos meteorológicos, estudios de lluvia y mediciones de partes fabricadas a menudo se explican más que adecuadamente con una distribución normal. Además, los errores en las mediciones científicas se aproximan extremadamente bien mediante una distribución normal. En 1733, Abraham DeMoivre desarrolló la ecuación matemática de la curva normal. Ésta ofrece una base sobre la que se fundamenta gran parte de la teoría de la estadística inductiva. La distribución normal a menudo se denomina **distribución gaussiana**, en honor de Karl Friedrich Gauss (1777-1855), quien también derivó su ecuación a partir de un estudio de errores en mediciones repetidas de la misma cantidad.

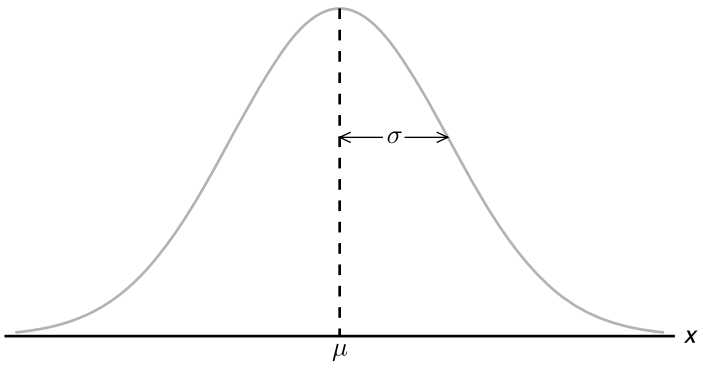


Figura 6.2: La curva normal.

Una variable aleatoria continua X que tiene la distribución con forma de campana de la figura 6.2 se denomina **variable aleatoria normal**. La ecuación matemática para la distribución de probabilidad de la variable normal depende de los dos parámetros μ y σ, su media y su desviación estándar. De aquí, denotamos los valores de la densidad de X con $n(x;\, \mu,\, \sigma)$.

Distribución normal

La densidad de la variable aleatoria normal X, con media μ y varianza σ^2, es

$$n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty,$$

donde $\pi = 3.14159\ldots$, y $e = 2.71828\ldots$.

Una vez que se especifican μ y σ, la curva normal queda determinada por completo. Por ejemplo, si $\mu = 50$ y $\sigma = 5$, entonces se pueden calcular las ordenadas $n(x;\, 50,\, 5)$ para diferentes valores de x y dibujar la curva. En la figura 6.3 dibujamos dos curvas normales que tienen la misma desviación estándar pero diferentes medias. Las dos curvas son idénticas en forma; pero están centradas en diferentes posiciones a lo largo del eje horizontal.

En la figura 6.4 trazamos dos curvas normales con la misma media pero con diferentes desviaciones estándar. Esta vez observamos que las dos curvas están centradas exactamente en la misma posición sobre el eje horizontal; pero la curva con la mayor desviación estándar es más baja y se extiende más lejos. Recuerde que el área bajo una curva de probabilidad debe ser igual a 1 y, por lo tanto, cuanto más variable sea el conjunto de observaciones más baja y más ancha será la curva correspondiente.

La figura 6.5 muestra el resultado de trazar dos curvas normales que tienen diferentes medias y diferentes desviaciones estándar. Evidentemente, están centradas en posiciones diferentes sobre el eje horizontal y sus formas reflejan los dos valores diferentes de σ.

De una inspección de las figuras 6.2 a 6.5, y al examinar la primera y la segunda derivadas de $n\ (x;\, \mu,\, \sigma)$, listamos las siguientes propiedades de la curva normal:

1. La moda, que es el punto sobre el eje horizontal donde la curva es un máximo, ocurre en $x = \mu$.

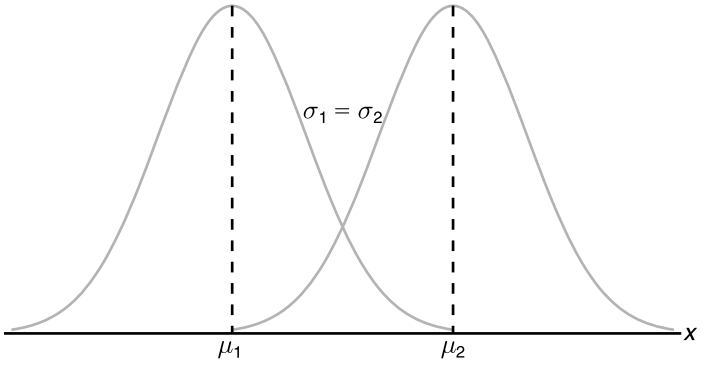


Figura 6.3: Curvas normales con $\mu_1 < \mu_2$ y $\sigma_1 = \sigma_2$.

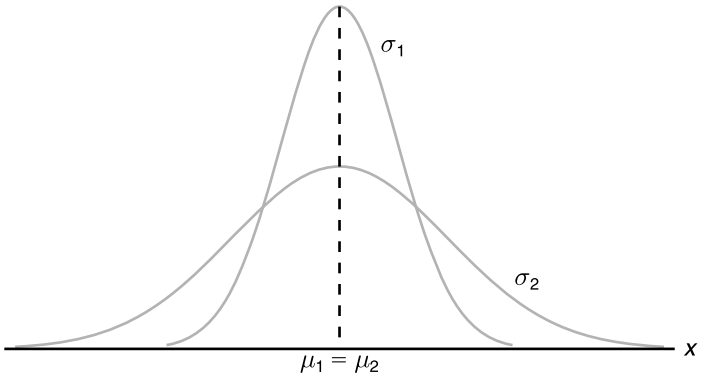


Figura 6.4: Curvas normales con $\mu_1 = \mu_2$ y $\sigma_1 < \sigma_2$.

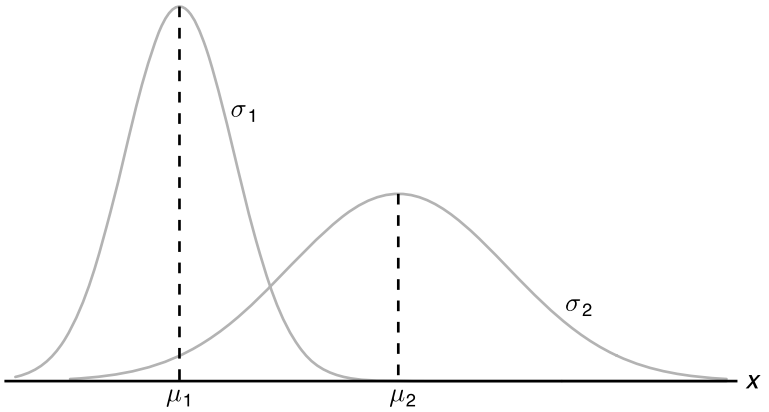


Figura 6.5: Curvas normales con $\mu_1 < \mu_2$ y $\sigma_1 < \sigma_2$.

2. La curva es simétrica alrededor de un eje vertical a través de la media μ.
3. La curva tiene sus puntos de inflexión en $x = \mu \pm \sigma$, es cóncava hacia abajo si $\mu - \sigma < X < \mu + \sigma$, y es cóncava hacia arriba en cualquier otro caso.
4. La curva normal se aproxima al eje horizontal de manera asintótica, conforme nos alejamos de la media en cualquier dirección.
5. El área total bajo la curva y sobre el eje horizontal es igual a 1.

Mostraremos ahora que los parámetros μ y σ^2 son realmente la media y la varianza de la distribución normal. Para evaluar la media, escribimos

$$E(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} xe^{-\frac{1}{2}[(x-\mu)/\sigma]^2}\, dx.$$

Al hacer $z = (x - \mu)/\sigma$ y $dx = \sigma\, dz$, obtenemos

$$\begin{aligned} E(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z)e^{-\frac{z^2}{2}}\, dz \\ &= \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-\frac{z^2}{2}}\, dz. \end{aligned}$$

El primer término de la derecha es μ veces el área bajo una curva normal con media cero y varianza 1 y, por ello, igual a μ. Por integración directa, el segundo término es igual a 0. De aquí que

$$E(X) = \mu.$$

La varianza de la distribución normal está dada por

$$E[(X - \mu)^2] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-\frac{1}{2}[(x-\mu)/\sigma]^2}\, dx.$$

Nuevamente, al hacer $z = (x - \mu)/\sigma$ y $dx = \sigma\, dz$, obtenemos

$$E[(X - \mu)^2] = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}}\, dz.$$

Al integrar por partes con $u = z$ y $dv = ze^{-z^2/2}\, dz$, de manera que $du = dz$ y $v = -e^{-z^2/2}$, encontramos que

$$E[(X - \mu)^2] = \frac{\sigma^2}{\sqrt{2\pi}} \left(-ze^{-z^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-z^2/2}\, dz \right) = \sigma^2(0 + 1) = \sigma^2.$$

Muchas variables aleatorias tienen distribuciones de probabilidad que pueden describirse de forma adecuada mediante la curva normal, una vez que se especifiquen μ y σ^2. En este capítulo supondremos que se conocen estos dos parámetros, quizás a partir de investigaciones anteriores. Más tarde haremos inferencias estadísticas cuando se desconozcan μ y σ^2 y se estimen a partir de los datos experimentales disponibles.

En un principio señalamos el papel que juega la distribución normal como una aproximación razonable de variables científicas en experimentos de la vida real. Hay otras aplicaciones de la distribución normal que el lector apreciará conforme avance en el estudio de este libro. La distribución normal tiene una gran aplicación como *distribución limitante*. Bajo ciertas condiciones, la distribución normal ofrece una buena aproximación continua a las distribuciones binomial e hipergeométrica. El caso de la aproximación a la binomial se examina en la sección 6.5. En el capítulo 8 el lector aprenderá acerca de las **distribuciones muestrales**. Resulta que la distri-

bución limitante de promedios muestrales es normal, lo cual brinda una base amplia para la inferencia estadística, que es muy valiosa para el analista de datos interesado en la estimación y la prueba de hipótesis. Las importantes áreas del análisis de varianza (capítulos 13, 14 y 15) y del control de calidad (capítulo 17) tienen su teoría basada en suposiciones que utilizan la distribución normal.

En mucho de lo que sigue en la sección 6.3, se ofrecen ejemplos para demostrar el uso de las tablas de la distribución normal. En la sección 6.4 continúan los ejemplos de aplicaciones de la distribución normal.

6.3 Áreas bajo la curva normal

La curva de cualquier distribución continua de probabilidad o función de densidad se construye de manera que el área bajo la curva limitada por las dos ordenadas $x = x_1$ y $x = x_2$ sea igual a la probabilidad de que la variable aleatoria X tome un valor entre $x = x_1$ y $x = x_2$. Así, para la curva normal de la figura 6.6,

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x; \mu, \sigma)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx,$$

está representada por el área de la región sombreada.

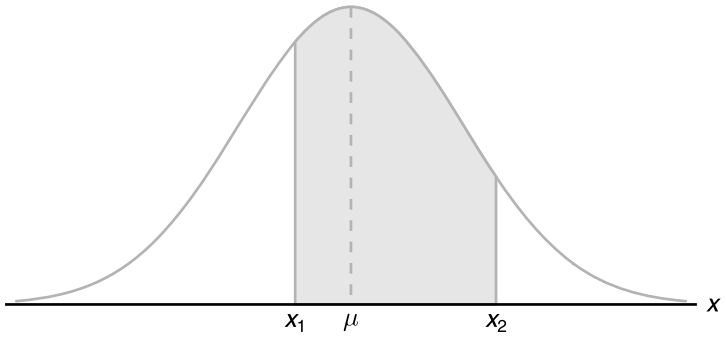


Figura 6.6: $P(x_1 < X < x_2)$ = área de la región sombreada.

En las figuras 6.3, 6.4 y 6.5 vimos cómo la curva normal depende de la media y de la desviación estándar de la distribución bajo investigación. El área bajo la curva entre cualesquiera dos ordenadas también debe depender de los valores μ y σ. Esto es evidente en la figura 6.7, donde sombreamos las regiones que corresponden a $P(x_1 < X < x_2)$ para dos curvas con medias y varianzas diferentes. La $P(x_1 < X < x_2)$, donde X es la variable aleatoria que describe la distribución A, se indica por el área sombreada más oscura. Si X es la variable aleatoria que describe la distribución B, entonces $P(x_1 < X < x_2)$ está dada por toda la región sombreada. Evidentemente, las dos regiones sombreadas tienen tamaños diferentes; por lo tanto, la probabilidad que se asocia con cada distribución será diferente para los dos valores dados de X.

La dificultad que se encuentra al resolver las integrales de funciones de densidad normal necesita de la tabulación de las áreas de la curva normal para una referencia

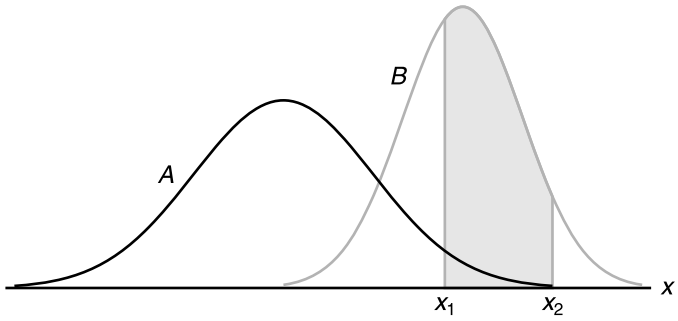


Figura 6.7: $P(x_1 < X < x_2)$ para diferentes curvas normales.

rápida. Sin embargo, sería una tarea desesparada intentar establecer tablas separadas para cada valor concebible de μ y σ. Afortunadamente, somos capaces de transformar todas las observaciones de cualquier variable aleatoria normal X a un nuevo conjunto de observaciones de una variable aleatoria normal Z con media 0 y varianza 1. Esto se puede realizar mediante la transformación

$$Z = \frac{X - \mu}{\sigma}.$$

Siempre que X tome un valor x, el valor correspondiente de Z está dado por $z = (x - \mu)/\sigma$. Por lo tanto, si X cae entre los valores $x = x_1$ y $x = x_2$, la variable aleatoria Z caerá entre los valores correspondientes $z_1 = (x_1 - \mu)/\sigma$ y $z_2 = (x_2 - \mu)/\sigma$. En consecuencia, podemos escribir

$$\begin{aligned} P(x_1 < X < x_2) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2}\, dz \\ &= \int_{z_1}^{z_2} n(z; 0, 1)\, dz = P(z_1 < Z < z_2), \end{aligned}$$

donde Z se ve como una variable aleatoria normal con media 0 y varianza 1.

Definición 6.1: La distribución de una variable aleatoria normal con media 0 y varianza 1 se llama **distribución normal estándar**.

Las distribuciones original y transformada se ilustran en la figura 6.8. Como todos los valores de X caen entre x_1 y x_2 tienen valores z correspondientes entre z_1 y z_2, el área bajo la curva X entre las ordenadas $x = x_1$ y $x = x_2$ de la figura 6.8 es igual al área bajo la curva Z entre las ordenadas transformadas $z = z_1$ y $z = z_2$.

Ahora hemos reducido el número requerido de tablas de áreas de curva normal a una, la de la distribución normal estándar. La tabla A.3 indica el área bajo la curva normal estándar que corresponde a $P(Z < z)$ para valores de z que van de -3.49 a 3.49. Para ilustrar el uso de esta tabla, encontremos la probabilidad de que Z sea menor que 1.74. Primero, localizamos un valor de z igual a 1.7 en la columna izquierda, después nos movemos a lo largo del renglón a la columna bajo 0.04, donde leemos 0.9591. Por lo tanto, $P(Z < 1.74) = 0.9591$. Para encontrar un valor z que corresponda a una probabilidad dada, se invierte el proceso. Por ejemplo, el valor z que deja un área de 0.2148 bajo la curva a la izquierda de z se observa que es -0.79.

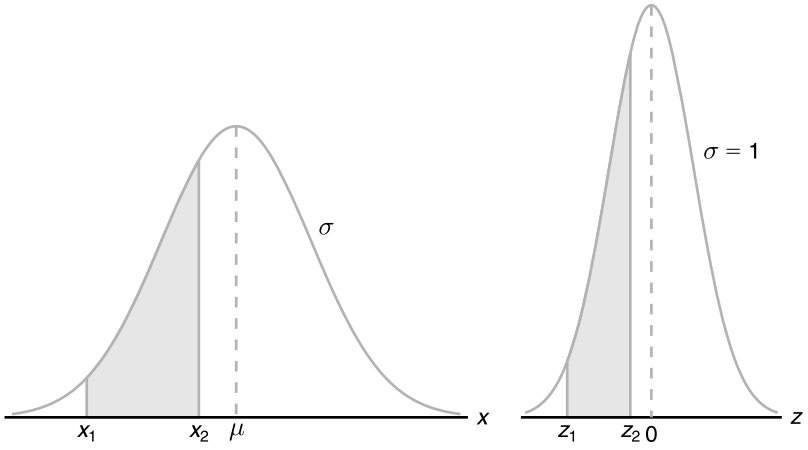


Figura 6.8: Distribuciones normales original y transformada.

Ejemplo 6.2: Dada una distribución normal estándar, encuentre el área bajo la curva que yace

a) a la derecha de $z = 1.84$ y

b) entre $z = -1.97$ y $z = 0.86$.

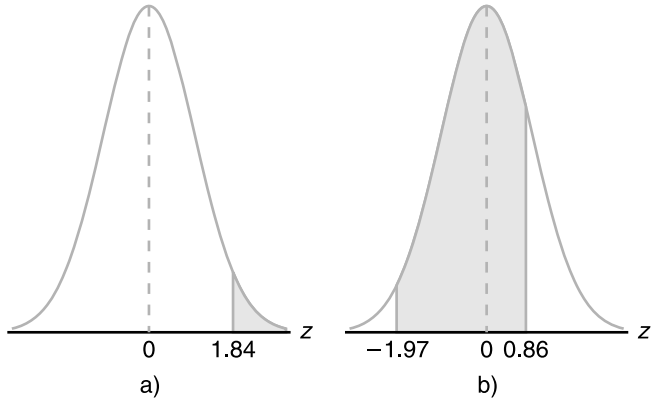


Figura 6.9: Áreas para el ejemplo 6.2.

Solución: *a*) El área en la figura 6.9*a*) a la derecha de $z = 1.84$ es igual a 1 menos el área en la tabla A.3 a la izquierda de $z = 1.84$; a saber, $1 - 0.9671 = 0.0329$.

b) El área en la figura 6.9*b*) entre $z = -1.97$ y $z = 0.86$ es igual al área a la izquierda de $z = 0.86$ menos el área a la izquierda de $z = -1.97$. De la tabla A.3 encontramos que el área que se desea es $0.8051 - 0.0244 = 0.7807$.

Ejemplo 6.3: Dada una distribución normal estándar, encuentre el valor de k tal que

a) $P(Z > k) = 0.3015$ y

b) $P(k < Z < -0.18) = 0.4197$.

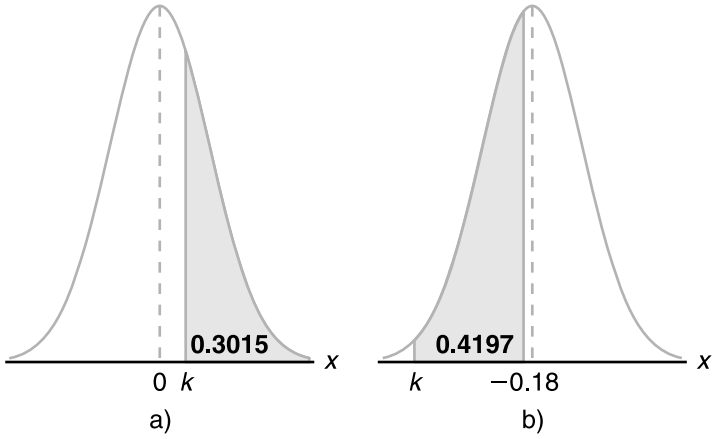


Figura 6.10: Áreas para el ejemplo 6.3.

Solución: *a*) En la figura 6.10*a*) vemos que el valor k que deja un área de 0.3015 a la derecha debe dejar entonces un área de 0.6985 a la izquierda. De la tabla A.3 se sigue que $k = 0.52$.

b) De la tabla A.3 notamos que el área total a la izquierda de -0.18 es igual a 0.4286. En la figura 6.10*b*) vemos que el área entre k y -0.18 es 0.4197, de manera que el área a la izquierda de k debe ser $0.4286 - 0.4197 = 0.0089$. Por lo tanto, de la tabla A.3, tenemos $k = -2.37$.

Ejemplo 6.4: Dada una variable aleatoria X que tiene una distribución normal con $\mu = 50$ y $\sigma = 10$, encuentre la probabilidad de que X tome un valor entre 45 y 62.

Solución: Los valores z que corresponden a $x_1 = 45$ y $x_2 = 62$ son

$$z_1 = \frac{45 - 50}{10} = -0.5, \quad \text{y} \quad z_2 = \frac{62 - 50}{10} = 1.2.$$

Por lo tanto,

$$P(45 < X < 62) = P(-0.5 < Z < 1.2).$$

La $P(-0.5 < Z < 1.2)$ se muestra por el área de la región sombreada de la figura 6.11. Esta área se puede encontrar al restar el área a la izquierda de la ordenada $z = -0.5$ de toda el área a la izquierda de $z = 1.2$. Usando la tabla A.3, tenemos

$$\begin{aligned} P(45 < X < 62) &= P(-0.5 < Z < 1.2) = P(Z < 1.2) - P(Z < -0.5) \\ &= 0.8849 - 0.3085 = 0.5764. \end{aligned}$$

Ejemplo 6.5: Dado que X tiene una distribución normal con $\mu = 300$ y $\sigma = 50$, encuentre la probabilidad de que X tome un valor mayor que 362.

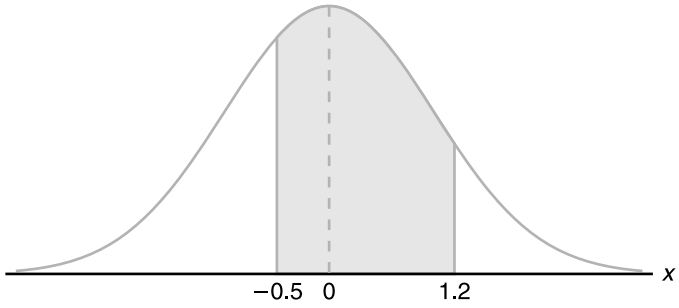


Figura 6.11: Área para el ejemplo 6.4.

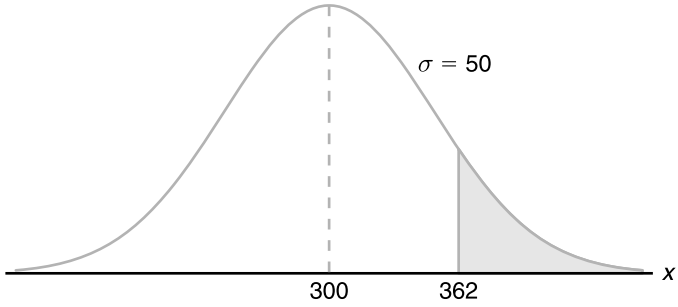


Figura 6.12: Área para el ejemplo 6.5.

Solución: La distribución de probabilidad normal que muestra el área que se desea se representa en la figura 6.12. Para encontrar la $P(X > 362)$, necesitamos evaluar el área bajo la curva normal a la derecha de $x = 362$. Esto se puede realizar al transformar $x = 362$ al valor z correspondiente, al obtener el área a la izquierda de z de la tabla A.3 y después restar esta área de 1. Encontramos que

$$z = \frac{362 - 300}{50} = 1.24.$$

De aquí,

$$P(X > 362) = P(Z > 1.24) = 1 - P(Z < 1.24) = 1 - 0.8925 = 0.1075.$$

De acuerdo con el teorema de Chebyshev, la probabilidad de que una variable aleatoria tome un valor dentro de 2 desviaciones estándar de la media es al menos 3/4. Si la variable aleatoria tiene una distribución normal, los valores z que corresponden a $x_1 = \mu - 2\sigma$ y $x_2 = \mu + 2\sigma$ se calculan fácilmente y son

$$z_1 = \frac{(\mu - 2\sigma) - \mu}{\sigma} = -2, \quad \text{y} \quad z_2 = \frac{(\mu + 2\sigma) - \mu}{\sigma} = 2.$$

De aquí,

$$\begin{aligned} P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2) = P(Z < 2) - P(Z < -2) \\ &= 0.9772 - 0.0228 = 0.9544, \end{aligned}$$

que es una afirmación mucho más fuerte que la que se establece mediante el teorema de Chebyshev.

Uso de la curva normal a la inversa

En ocasiones se nos pide encontrar el valor de z que corresponde a una probabilidad específica que cae entre los valores que se listan en la tabla A.3 (véase el ejemplo 6.6). Por conveniencia, siempre elegiremos el valor z que corresponde a la probabilidad tabular que está más cerca de la probabilidad que se especifica.

Los dos ejemplos anteriores se resolvieron al ir primero de un valor de x a un valor z y después calcular el área que se desea. En el ejemplo 6.6 invertimos el proceso y comenzamos con un área o probabilidad conocida, encontramos el valor z y después determinamos x reacomodando la fórmula

$$z = \frac{x - \mu}{\sigma} \quad \text{para obtener} \quad x = \sigma z + \mu.$$

Ejemplo 6.6: Dada una distribución normal con $\mu = 40$ y $\sigma = 6$, encuentre el valor de x que tiene

a) 45% del área a la izquierda y

b) 14% del área a la derecha.

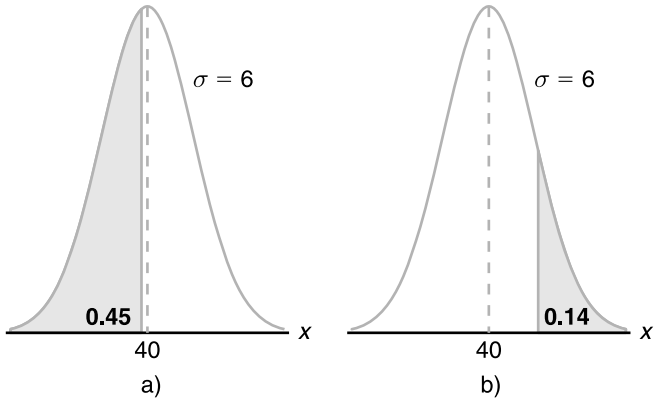


Figura 6.13: Áreas para el ejemplo 6.6.

Solución: *a*) En la figura 6.13*a*) se sombrea un área de 0.45 a la izquierda del valor x que se desea. Requerimos un valor z que deje un área de 0.45 a la izquierda. De la tabla A.3 encontramos $P(Z < -0.13) = 0.45$, por lo que el valor z que se desea es -0.13. De aquí,

$$x = (6)(-0.13) + 40 = 39.22.$$

b) En la figura 6.13*b*) sombreamos un área igual a 0.14 a la derecha del valor x que se desea. Esta vez requerimos un valor z que deje 0.14 del área a la derecha y, por ello, un área de 0.86 a la izquierda. De nuevo, de la tabla A.3, encontramos $P(Z < 1.08) = 0.86$, por lo que el valor z que se desea es 1.08 y

$$x = (6)(1.08) + 40 = 46.48.$$

6.4 Aplicaciones de la distribución normal

Algunos de los muchos problemas para los que es aplicable la distribución normal se tratan en los siguientes ejemplos. El uso de la curva normal para aproximar probabilidades binomiales se considera en la sección 6.5.

Ejemplo 6.7: Cierto tipo de batería de almacenamiento dura, en promedio, 3.0 años, con una desviación estándar de 0.5 años. Suponiendo que las duraciones de la batería se distribuyen normalmente, encuentre la probabilidad de que una batería dada dure menos de 2.3 años.

Solución: Primero construya un diagrama como el de la figura 6.14, que muestra la distribución dada de duraciones de las baterías y el área que se desea. Para encontrar la $P(X < 2.3)$, necesitamos evaluar el área bajo la curva normal a la izquierda de 2.3. Esto se logra al encontrar el área a la izquierda del valor z correspondiente. De aquí encontramos que

$$z = \frac{2.3 - 3}{0.5} = -1.4,$$

y entonces con la tabla A.3 tenemos

$$P(X < 2.3) = P(Z < -1.4) = 0.0808.$$

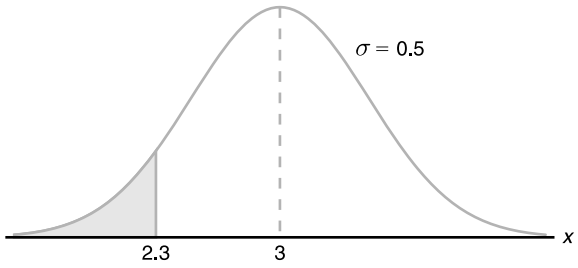


Figura 6.14: Área para el ejemplo 6.7.

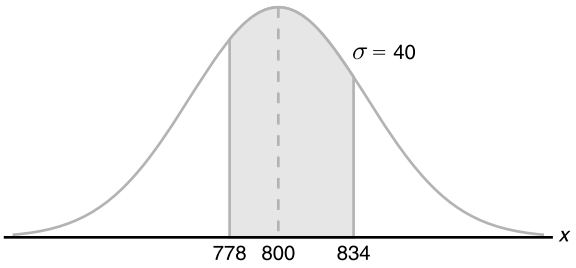


Figura 6.15: Área para el ejemplo 6.8.

Ejemplo 6.8: Una empresa de material eléctrico fabrica bombillas de luz que tienen una duración, antes de quemarse (fundirse), que se distribuye normalmente con media igual a 800 horas y una desviación estándar de 40 horas. Encuentre la probabilidad de que una bombilla se queme entre 778 y 834 horas.

Solución: La distribución de las bombillas se ilustra en la figura 6.15. Los valores z que corresponden a $x_1 = 778$ y $x_2 = 834$ son

$$z_1 = \frac{778 - 800}{40} = -0.55, \quad \text{y} \quad z_2 = \frac{834 - 800}{40} = 0.85.$$

De aquí,

$$\begin{aligned} P(778 < X < 834) &= P(-0.55 < Z < 0.85) = P(Z < 0.85) - P(Z < -0.55) \\ &= 0.8023 - 0.2912 = 0.5111. \end{aligned}$$

Ejemplo 6.9: En un proceso industrial el diámetro de un cojinete de bolas es una parte componente importante. El comprador establece que las especificaciones en el diámetro sean

3.0 ± 0.01 cm. La implicación es que no se aceptará ninguna parte que quede fuera de estas especificaciones. Se sabe que en el proceso el diámetro de un cojinete tiene una distribución normal con media 3.0 y desviación estándar $\sigma = 0.005$. En promedio, ¿cuántos cojinetes fabricados se descartarán?

Solución: La distribución de diámetros se ilustra en la figura 6.16. Los valores que corresponden a los límites especificados son $x_1 = 2.99$ y $x_2 = 3.01$. Los valores z correspondientes son

$$z_1 = \frac{2.99 - 3.0}{0.005} = -2.0, \quad \text{y} \quad z_2 = \frac{3.01 - 3.0}{0.005} = +2.0.$$

De aquí,

$$P(2.99 < X < 3.01) = P(-2.0 < Z < 2.0).$$

De la tabla A.3, $P(Z < -2.0) = 0.0228$. Debido a la simetría de la distribución normal, encontramos que

$$P(Z < -2.0) + P(Z > 2.0) = 2(0.0228) = 0.0456.$$

Como resultado se anticipa que, en promedio, se descartarán 4.56% de los cojinetes fabricados.

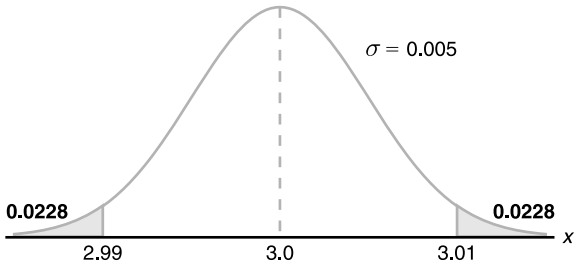


Figura 6.16: Área para el ejemplo 6.9.

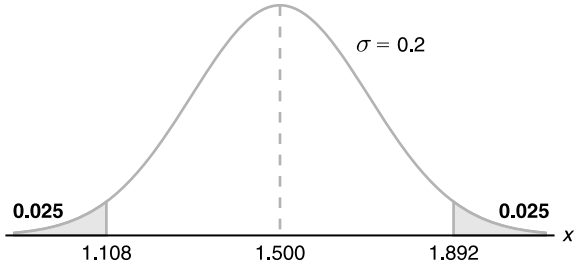


Figura 6.17: Especificaciones para el ejemplo 6.10.

Ejemplo 6.10: Se utilizan medidores para rechazar todos los componentes donde cierta dimensión no esté dentro de la especificación $1.50 \pm d$. Se sabe que esta medición se distribuye normalmente con media 1.50 y desviación estándar 0.2. Determine el valor d tal que las especificaciones "cubran" 95% de las mediciones.

Solución: De la tabla A.3 sabemos que

$$P(-1.96 < Z < 1.96) = 0.95.$$

Por lo tanto,

$$1.96 = \frac{(1.50 + d) - 1.50}{0.2},$$

de la que obtenemos

$$d = (0.2)(1.96) = 0.392.$$

Una ilustración de las especificaciones se muestra en la figura 6.17.

Ejemplo 6.11: Cierta máquina fabrica resistencias eléctricas que tienen una resistencia media de 40 ohms y una desviación estándar de 2 ohms. Suponiendo que la resistencia sigue una distribución normal y se puede medir con cualquier grado de precisión, ¿qué porcentaje de resistencias tendrán una resistencia que exceda 43 ohms?

Solución: Se encuentra un porcentaje al multiplicar la frecuencia relativa por 100%. Como la frecuencia relativa para un intervalo es igual a la probabilidad de caer en el intervalo, debemos encontrar el área a la derecha de $x = 43$ en la figura 6.18. Esto se realiza al transformar $x = 43$ al valor z correspondiente, con lo cual se obtiene el área a la izquierda de z de la tabla A.3, y después se resta esta área de 1. Encontramos que

$$z = \frac{43 - 40}{2} = 1.5.$$

Por lo tanto,

$$P(X > 43) = P(Z > 1.5) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668.$$

Así, 6.68% de las resistencias tendrán una resistencia que exceda 43 ohms.

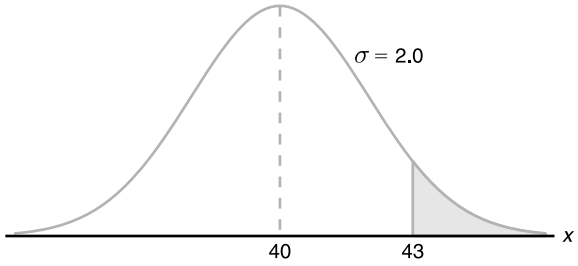


Figura 6.18: Área para el ejemplo 6.11.

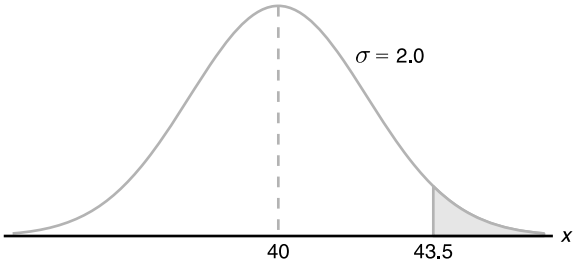


Figura 6.19: Área para el ejemplo 6.12.

Ejemplo 6.12: Encuentre el porcentaje de resistencias que excedan 43 ohms para el ejemplo 6.11 si la resistencia se mide al ohm más cercano.

Solución: Este problema difiere del ejemplo 6.11, pues ahora asignamos una medida de 43 ohms a todas las resistencias cuyas resistencias sean mayores que 42.5 y menores que 43.5. Realmente aproximamos una distribución discreta por medio de una distribución continua normal. El área que se requiere es la región sombreada a la derecha de 43.5 en la figura 6.19. Encontramos ahora que

$$z = \frac{43.5 - 40}{2} = 1.75.$$

De aquí,

$$P(X > 43.5) = P(Z > 1.75) = 1 - P(Z < 1.75) = 1 - 0.9599 = 0.0401.$$

Por lo tanto, 4.01% de las resistencias exceden 43 ohms cuando se miden al ohm más cercano. La diferencia 6.68% − 4.01% = 2.67% entre esta respuesta y la del ejemplo 6.11 representa todas las resistencias que tienen una resistencia mayor que 43 y menor que 43.5, que ahora se registran como de 43 ohms.

Ejemplo 6.13: La calificación promedio para un examen es 74 y la desviación estándar es 7. Si 12% de la clase obtiene A y las calificaciones siguen una curva que tiene una distribución normal, ¿cuál es la A más baja posible y la B más alta posible?

Solución: En este ejemplo comenzamos con un área de probabilidad conocida, encontramos el valor z y después determinamos x de la fórmula $x = \sigma z + \mu$. Un área de 0.12, que corresponde a la fracción de estudiantes que reciben A, se sombrea en la figura, 6.20. Requerimos un valor z que deje 0.12 del área a la derecha y, por ello, un área de 0.88 a la izquierda. De la tabla A.3, $P(Z < 1.18)$ tiene el valor más cercano a 0.88, de manera que el valor z que se desea es 1.18. De aquí,

$$x = (7)(1.18) + 74 = 82.26.$$

Por lo tanto, la A más baja es 83 y la B más alta es 82.

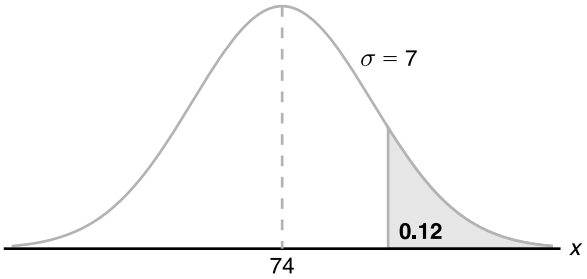


Figura 6.20: Área para el ejemplo 6.13.

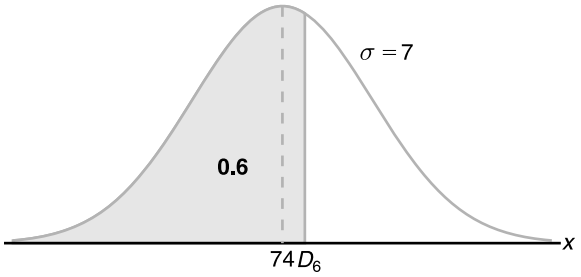


Figura 6.21: Área para el ejemplo 6.14.

Ejemplo 6.14: Refiérase al ejemplo 6.13 y encuentre el sexto decil.

Solución: El sexto decil, escrito como D_6, es el valor x que deja 60% del área a la izquierda, como se muestra en la figura 6.21. De la tabla A.3 encontramos $P(Z < 0.25) \approx 0.6$, de manera que el valor z que se desea es 0.25. Ahora, $x = (7)\ (0.25) + 74 = 75.75$. De aquí, $D_6 = 75.75$. Es decir, 60% de las calificaciones son de 75 o menos.

Ejercicios

6.1 Dada una distribución normal estándar, encuentre el área bajo la curva que está

a) a la izquierda de $z = 1.43$;

b) a la derecha de $z = -0.89$;

c) entre $z = -2.16$ y $z = -0.65$;

d) a la izquierda de $z = -1.39$;

e) a la derecha de $z = 1.96$;

f) entre $z = -0.48$ y $z = 1.74$.

6.2 Encuentre el valor de z si el área bajo una curva normal estándar

a) a la derecha de z es 0.3622;

b) a la izquierda de z es 0.1131;

c) entre 0 y z, con $z > 0$, es 0.4838;

d) entre $-z$ y z, con $z > 0$, es 0.9500.

6.3 Dada una distribución normal estándar, encuentre el valor de k tal que

a) $P(Z < k) = 0.0427$;

b) $P(Z > k) = 0.2946$;

c) $P(-0.93 < Z < k) = 0.7235$.

6.4 Dada una distribución normal con $\mu = 30$ y $\sigma = 6$, encuentre

a) el área de la curva normal a la derecha de $x = 17$;

b) el área de la curva normal a la izquierda de $x = 22$;

c) el área de la curva normal entre $x = 32$ y $x = 41$;

d) el valor de x que tiene 80% del área de la curva normal a la izquierda;

e) los dos valores de x que contienen el 75% central del área de la curva normal.

6.5 Dada la variable X normalmente distribuida con media 18 y desviación estándar 2.5, encuentre

a) $P(X < 15)$;

b) el valor de k tal que $P(X < k) = 0.2236$;

c) el valor de k tal que $P(X > k) = 0.1814$;

d) $P(17 < X < 21)$.

6.6 De acuerdo con el teorema de Chebyshev, la probabilidad de que cualquier variable aleatoria tome un valor dentro de tres desviaciones estándar de la media es al menos 8/9. Si se sabe que la distribución de probabilidad de una variable aleatoria X es normal con media μ y varianza σ^2, ¿cuál es el valor exacto de $P(\mu - 3\sigma < X < \mu + 3\sigma)$?

6.7 Un investigador científico informa que unos ratones vivirán un promedio de 40 meses cuando sus dietas se restringen drásticamente y después se enriquecen con vitaminas y proteínas. Suponiendo que la vidas de tales ratones se distribuyen normalmente con una desviación estándar de 6.3 meses, encuentre la probabilidad de que un ratón dado vivirá

a) más de 32 meses;

b) menos de 28 meses;

c) entre 37 y 49 meses.

6.8 Las barras de pan de centeno que cierta panadería distribuye a las tiendas locales tienen una longitud promedio de 30 centímetros y una desviación estándar de 2 centímetros. Suponiendo que las longitudes están distribuidas normalmente, ¿qué porcentaje de las barras son

a) más largas que 31.7 centímetros?

b) de entre 29.3 y 33.5 centímetros de longitud?

c) más cortas que 25.5 centímetros?

6.9 Una máquina expendedora de bebidas gaseosas se regula para que sirva un promedio de 200 mililitros por vaso. Si la cantidad de bebida se distribuye normalmente con una desviación estándar igual a 15 mililitros,

a) ¿qué fracción de los vasos contendrá más de 224 mililitros?

b) ¿cuál es la probabilidad de que un vaso contenga entre 191 y 209 mililitros?

c) ¿cuántos vasos probablemente se derramarán si se utilizan vasos de 230 mililitros para las siguientes 1000 bebidas?

d) ¿por debajo de qué valor obtendremos el 25% más pequeño de las bebidas?

6.10 El diámetro interior del anillo de un pistón terminado se distribuye normalmente con una media de 10 centímetros y una desviación estándar de 0.03 centímetros.

a) ¿Qué proporción de anillos tendrán diámetros interiores que excedan 10.075 centímetros?

b) ¿Cuál es la probabilidad de que el anillo de un pistón tenga un diámetro interior entre 9.97 y 10.03 centímetros?

c) ¿Por debajo de qué valor del diámetro interior caerá 15% de los anillos de pistón?

6.11 Un abogado viaja todos los días de su casa en los suburbios a su oficina en el centro de la ciudad. El tiempo promedio para un viaje sólo de ida es 24 minutos, con una desviación estándar de 3.8 minutos. Suponga que la distribución de los tiempos de viaje está distribuida normalmente.

a) ¿Cuál es la probabilidad de que un viaje tome al menos 1/2 hora?

b) Si la oficina abre a las 9:00 A.M. y él sale diario de su casa a las 8:45 A.M., ¿qué porcentaje de las veces llegará tarde al trabajo?

c) Si sale de su casa a las 8:35 A.M. y el café se sirve en la oficina de 8:50 A.M. a 9:00 A.M., cuál es la probabilidad de que se pierda el café?

d) Encuentre la longitud de tiempo por arriba de la cual encontramos e1 15% de los viajes más lentos.

e) Encuentre la probabilidad de que 2 de los siguientes 3 viajes tomen al menos 1/2 hora.

6.12 En el ejemplar de noviembre de 1990 de *Chemical Engineering Progress*, un estudio analiza el porcentaje de pureza del oxígeno de cierto proveedor. Suponga que la media fue 99.61 con una desviación estándar de 0.08. Suponga que la distribución del porcentaje de pureza fue aproximadamente normal.

a) ¿Qué porcentaje de los valores de pureza esperaría que estuvieran entre 99.5 y 99.7?

b) ¿Qué valor de pureza esperaría que excediera exactamente 5% de la población?

6.13 La vida promedio de cierto tipo de motor pequeño es de 10 años con una desviación estándar de 2 años. El fabricante reemplaza gratis todos los motores que fallen dentro del periodo de garantía. Si él está dispuesto a reemplazar sólo 3% de los motores que fallan, ¿cuánto tiempo de garantía debería ofrecer? Suponga que la duración de un motor sigue una distribución normal.

6.14 Las alturas de 1000 estudiantes se distribuyen normalmente con una media de 174.5 centímetros y

una desviación estándar de 6.9 centímetros. Suponiendo que las alturas se registran al medio centímetro más cercano, ¿cuántos de estos estudiantes esperaría que tuvieran alturas

a) menores que 160.0 centímetros?

b) de entre 171.5 y 182.0 centímetros inclusive?

c) iguales a 175.0 centímetros?

d) mayores que o iguales a 188.0 centímetros?

6.15 Una compañía paga a sus empleados un salario promedio de \$15.90 por hora con una desviación estándar de \$1.50. Si los salarios se distribuyen aproximadamente de forma normal y se pagan al centavo más cercano,

a) ¿qué porcentaje de los trabajadores reciben salarios entre \$13.75 y \$16.22 inclusive por hora?

b) ¿el 5% más alto de los salarios por hora de los empleados es mayor a qué cantidad?

6.16 Los pesos de un número grande de *poodle* (caniche) miniatura se distribuyen aproximadamente de forma normal con una media de 8 kilogramos y una desviación estándar de 0.9 kilogramos. Si las mediciones se registran al décimo de kilogramo más cercano, encuentre la fracción de estos *poodle* con pesos

a) por arriba de 9.5 kilogramos;

b) a lo más 8.6 kilogramos;

c) entre 7.3 y 9.1 kilogramos inclusive.

6.17 La resistencia a la tensión de cierto componente de metal se distribuye normalmente con una media de 10,000 kilogramos por centímetro cuadrado y una desviación estándar de 100 kilogramos por centímetro cuadrado. Las mediciones se registran a los 50 kilogramos por centímetro cuadrado más cercanos.

a) ¿Qué proporción de estos componentes excede 10,150 kilogramos por centímetro cuadrado de resistencia a la tensión?

b) Si las especificaciones requieren que todos los componentes tengan resistencia a la tensión entre 9800 y 10,200 kilogramos por centímetro cuadrado inclusive, ¿qué proporción de piezas esperaría que se descartara?

6.18 Si un conjunto de observaciones se distribuye de manera normal, ¿qué porcentaje de éstas difieren de la media en

a) más de 1.3σ?

b) menos de 0.52σ?

6.19 Los CI de 600 aspirantes de cierta universidad se distribuyen aproximadamente de forma normal con una media de 115 y una desviación estándar de 12. Si la universidad requiere un CI de al menos 95, ¿cuántos de estos estudiantes serán rechazados sobre esta base sin importar sus otras calificaciones?

6.20 Dada una distribución continua uniforme, demuestre que

a) $\mu = \frac{A+B}{2}$, y

b) $\sigma^2 = \frac{(B-A)^2}{12}$.

6.21 La cantidad de café diaria, en litros, que sirve una máquina que se localiza en el vestíbulo de un aeropuerto es una variable aleatoria X que tiene una distribución continua uniforme con $A = 7$ y $B = 10$. Encuentre la probabilidad de que en un día dado la cantidad de café que sirve esta máquina sea

a) a lo más 8.8 litros;

b) más de 7.4 litros, pero menos de 9.5 litros;

c) al menos 8.5 litros.

6.22 Un autobús llega cada 10 minutos a una parada. Se supone que el tiempo de espera para un individuo en particular es una variable aleatoria con distribución continua uniforme.

a) ¿Cuál es la probabilidad de que el individuo espere más de 7 minutos?

b) ¿Cuál es la probabilidad de que el individuo espere entre 2 y 7 minutos?

6.5 Aproximación normal a la binomial

Las probabilidades asociadas con experimentos binomiales se obtienen fácilmente a partir de la fórmula $b(x; n, p)$ de la distribución binomial o de la tabla A.1 cuando n es pequeña. Además, las probabilidades binomiales están fácilmente disponibles en muchos paquetes de software. Sin embargo, resulta instructivo aprender la relación entre la distribución binomial y la normal. En la sección 5.6 ilustramos como la distribución de Poisson se puede utilizar para aproximar probabilidades binomiales cuando n es bastante grande y p está muy cercana a 0 o a 1. Las distribuciones binomial y de Poisson son ambas discretas. La primera aplicación de

una distribución continua de probabilidad para aproximar probabilidades sobre un espacio muestral discreto se demuestra con el ejemplo 6.12, donde se utiliza la curva normal. La distribución normal a menudo es una buena aproximación a una distribución discreta cuando la última adquiere una forma de campana simétrica. Desde un punto de vista teórico, algunas distribuciones convergen a la normal conforme sus parámetros se aproximan a ciertos límites. La distribución normal es una distribución de aproximación conveniente, ya que la función de distribución acumulada se tabula con mucha facilidad. La distribución binomial se aproxima bien por la normal en problemas prácticos cuando se trabaja con la función de distribución acumulada. Establecemos ahora un teorema que nos permitirá utilizar áreas bajo la curva normal para aproximar propiedades binomiales cuando n es suficientemente grande.

Teorema 6.2: Si X es una variable aleatoria binomial con media $\mu = np$ y varianza $\sigma^2 = npq$, entonces la forma limitante de la distribución de

$$Z = \frac{X - np}{\sqrt{npq}},$$

conforme $n \to \infty$, es la distribución normal estándar $n(z; 0, 1)$.

Resulta que la distribución normal con $\mu = np$ y $\sigma^2 = np(1 - p)$ no sólo ofrece una aproximación muy precisa a la distribución binomial cuando n es grande y p no está extremadamente cercana a 0 o a 1, sino que también brinda una aproximación bastante buena aun cuando n sea pequeña y p esté razonablemente cercana a $1/2$.

Para ilustrar la aproximación normal a la distribución binomial, primero dibujamos el histograma para $b(x; 15, 0.4)$ y después superponemos la curva normal particular que tenga las mismas media y varianza que la variable binomial X. De aquí dibujamos una curva normal con

$$\mu = np = (15)\ (0.4) = 6 \text{ y } \sigma^2 = npq = (15)\ (0.4)\ (0.6) = 3.6.$$

El histograma de $b(x; 15, 0.4)$ y la curva normal superpuesta correspondiente, que está determinada por completo por su media y su varianza, se ilustran en la figura 6.22.

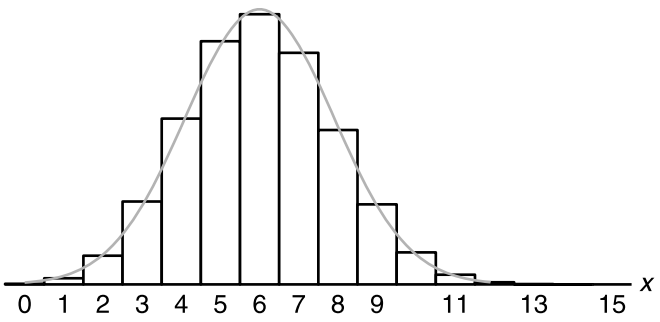


Figura 6.22: Aproximación normal de $b(x; 15, 0.4)$.

La probabilidad exacta de que la variable aleatoria binomial X tome un valor dado x es igual al área de la barra cuya base se centra en x. Por ejemplo, la probabilidad exacta de que X tome el valor 4 es igual al área del rectángulo con base centrada en $x = 4$. Con la tabla A.1, encontramos que esta área es

$$P(X = 4) = b(4;15, 0.4) = 0.1268,$$

que es aproximadamente igual al área de la región sombreada bajo la curva normal entre las dos ordenadas $x_1 = 3.5$ y $x_2 = 4.5$ en la figura 6.23. Al convertir a valores z, tenemos

$$z_1 = \frac{3.5 - 6}{1.897} = -1.32 \qquad \text{y} \qquad z_2 = \frac{4.5 - 6}{1.897} = -0.79.$$

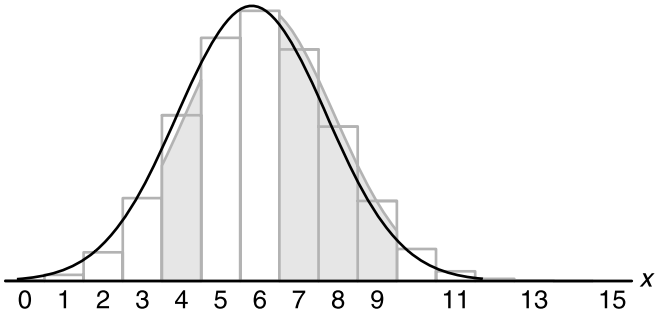


Figura 6.23: Aproximación normal de $b(x; 15, 0.4)$ y $\sum_{x=7}^{9} b(x; 15, 0.4)$.

Si X es una variable aleatoria binomial y Z una variable normal estándar, entonces,

$$\begin{aligned} P(X = 4) = b(4; 15, 0.4) &\approx P(-1.32 < Z < -0.79) \\ &= P(Z < -0.79) - P(Z < -1.32) = 0.2148 - 0.0934 = 0.1214. \end{aligned}$$

Esto coincidie bastante con el valor exacto de 0.1268.

La aproximación normal es más útil al calcular sumas binomiales para valores grandes de n. Con referencia a la figura 6.23, nos podemos interesar en la probabilidad de que X tome un valor de 7 a 9 inclusive. La probabilidad exacta está dada por

$$\begin{aligned} P(7 \le X \le 9) &= \sum_{x=0}^{9} b(x; 15, 0.4) - \sum_{x=0}^{6} b(x; 15, 0.4) \\ &= 0.9662 - 0.6098 = 0.3564, \end{aligned}$$

que es igual a la suma de las áreas de los rectángulos cuyas bases están centradas en $x = 7$, 8 y 9. Para la aproximación normal encontramos el área de la región sombreada bajo la curva entre las ordenadas $x_1 = 6.5$ y $x_2 = 9.5$ en la figura 6.23. Los valores z correspondientes son

$$z_1 = \frac{6.5 - 6}{1.897} = 0.26 \qquad \text{y} \qquad z_2 = \frac{9.5 - 6}{1.897} = 1.85.$$

Ahora,

$$P(7 \leq X \leq 9) \approx P(0.26 < Z < 1.85) = P(Z < 1.85) - P(Z < 0.26)$$
$$= 0.9678 - 0.6026 = 0.3652.$$

Una vez más, la aproximación de la curva normal ofrece un valor que coincide mucho con el valor exacto de 0.3564. El grado de precisión, que depende de qué tan bien se ajuste la curva al histograma, aumentará conforme n aumente. Esto es particularmente cierto cuando p no está muy cercana a 1/2 y el histograma ya no es simétrico. Las figuras 6.24 y 6.25 muestran los histogramas para $b(x; 6, 0.2)$ y $b(x; 15, 0.2)$, respectivamente. Es evidente que una curva normal se ajustará considerablemente mejor al histograma cuando $n = 15$ que cuando $n = 6$.

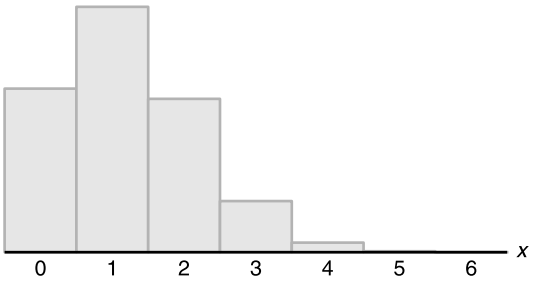


Figura 6.24: Histograma para $b(x; 6, 0.2)$.

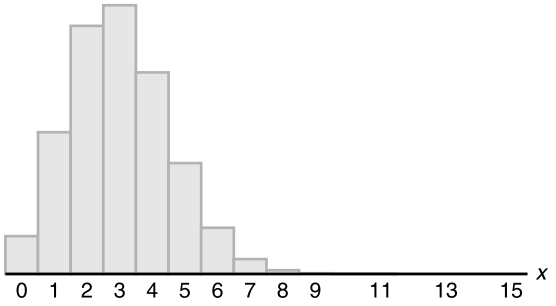


Figura 6.25: Histograma para $b(x; 15, 0.2)$.

En nuestras ilustraciones de la aproximación normal a la binomial, se hizo evidente que si buscamos el área bajo la curva normal hacia la izquierda de, digamos x, es más preciso utilizar $x + 0.5$. Esto es una corrección para dar cabida al hecho de que una distribución discreta se aproxima mediante una distribución continua. La corrección $+\mathbf{0.5}$ se llama **corrección de continuidad**. A partir de la explicación anterior, damos la siguiente aproximación normal formal a la binomial.

Aproximación normal a la distribución normal

Sea X una variable aleatoria binomial con parámetros n y p. Entonces, X tiene aproximadamente una distribución normal con $\mu = np$ y $\sigma^2 = npq = np(1 - p)$ y

$$P(X \leq x) = \sum_{k=0}^{x} b(k; n, p)$$

$$\approx \text{área bajo la curva normal a la izquierda de } x + 0.5$$

$$= P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{npq}}\right),$$

y la aproximación será buena si np y $n(1 - p)$ son mayores que o iguales a 5.

Como indicamos antes, la calidad de la aproximación es bastante buena para n grande. Si p es cercana a 1/2, un tamaño de la muestra moderado o pequeño será suficiente para una aproximación razonable. Ofrecemos la tabla 6.1 como una

indicación de la calidad de la aproximación. Se dan tanto la aproximación normal como las probabilidades binomiales acumuladas reales. Observe que en $p = 0.05$ y $p = 0.10$, la aproximación es bastante gruesa para $n = 10$. Sin embargo, aun para $n = 10$, note la mejoría para $p = 0.50$. Por otro lado, cuando p es fija en $p = 0.05$, observe la mejoría de la aproximación conforme vamos de $n = 20$ a $n = 100$.

Tabla 6.1: Aproximación normal y probabilidades binomiales acumuladas reales

	$p = 0.05,\ n = 10$		$p = 0.10,\ n = 10$		$p = 0.50,\ n = 10$	
r	**Binomial**	**Normal**	**Binomial**	**Normal**	**Binomial**	**Normal**
0	0.5987	0.5000	0.3487	0.2981	0.0010	0.0022
1	0.9139	0.9265	0.7361	0.7019	0.0107	0.0136
2	0.9885	0.9981	0.9298	0.9429	0.0547	0.0571
3	0.9990	1.0000	0.9872	0.9959	0.1719	0.1711
4	1.0000	1.0000	0.9984	0.9999	0.3770	0.3745
5			1.0000	1.0000	0.6230	0.6255
6					0.8281	0.8289
7					0.9453	0.9429
8					0.9893	0.9864
9					0.9990	0.9978
10					1.0000	0.9997
	$p = 0.05$					
	$n = 20$		$n = 50$		$n = 100$	
r	**Binomial**	**Normal**	**Binomial**	**Normal**	**Binomial**	**Normal**
0	0.3585	0.3015	0.0769	0.0968	0.0059	0.0197
1	0.7358	0.6985	0.2794	0.2578	0.0371	0.0537
2	0.9245	0.9382	0.5405	0.5000	0.1183	0.1251
3	0.9841	0.9948	0.7604	0.7422	0.2578	0.2451
4	0.9974	0.9998	0.8964	0.9032	0.4360	0.4090
5	0.9997	1.0000	0.9622	0.9744	0.6160	0.5910
6	1.0000	1.0000	0.9882	0.9953	0.7660	0.7549
7			0.9968	0.9994	0.8720	0.8749
8			0.9992	0.9999	0.9369	0.9463
9			0.9998	1.0000	0.9718	0.9803
10			1.0000	1.0000	0.9885	0.9941

Ejemplo 6.15: La probabilidad de que un paciente se recupere de una rara enfermedad de la sangre es 0.4. Si se sabe que 100 personas contrajeron esta enfermedad, ¿cuál es la probabilidad de que menos de 30 sobrevivan?

Solución: Representemos con la variable binomial X el número de pacientes que sobreviven. Como $n = 100$, deberíamos obtener resultados bastante precisos usando la aproximación de la curva normal con

$$\mu = np = (100)(0.4) = 40,$$

y

$$\sigma = \sqrt{npq} = \sqrt{(100)(0.4)(0.6)} = 4.899.$$

Para obtener la probabilidad que se desea, tenemos que encontrar el área a la izquierda de $x = 29.5$. El valor z que corresponde a 29.5 es

$$z = \frac{29.5 - 40}{4.899} = -2.14,$$

y la probabilidad de que menos de 30 de los 100 pacientes sobrevivan está dada por la región sombreada en la figura 6.26. De aquí,

$$P(X < 30) \approx P(Z < -2.14) = 0.0162.$$

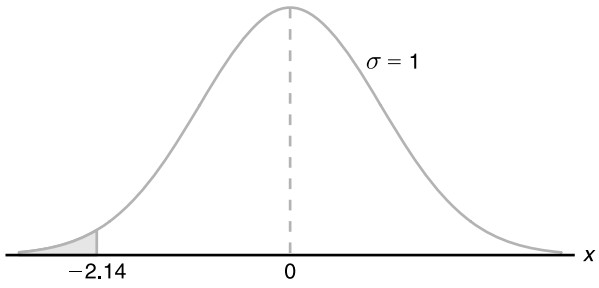


Figura 6.26: Área para el ejemplo 6.15.

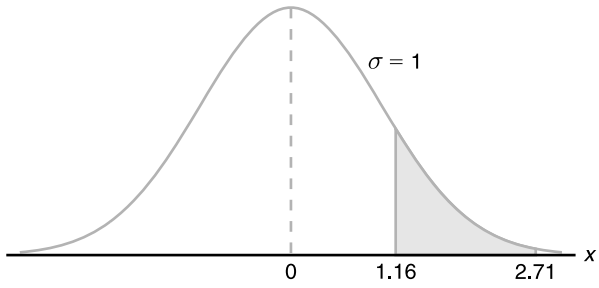


Figura 6.27: Área para el ejemplo 6.16.

Ejemplo 6.16: Una prueba de opción múltiple tiene 200 preguntas, cada una de las cuales con 4 respuestas posibles de las que sólo 1 es la correcta. ¿Cuál es la probabilidad de que solamente adivinando se obtengan de 25 a 30 respuestas correctas para 80 de los 200 problemas, sobre los que el estudiante no tiene conocimientos?

Solución: La probabilidad de una respuesta correcta para cada una de las 80 preguntas es $p = 1/4$. Si X representa el número de respuestas correctas por la mera adivinación, entonces,

$$P(25 \leq X \leq 30) = \sum_{x=25}^{30} b(x; 80, 1/4).$$

Al usar la aproximación de la curva normal con

$$\mu = np = (80)\left(\frac{1}{4}\right) = 20,$$

y

$$\sigma = \sqrt{npq} = \sqrt{(80)(1/4)(3/4)} = 3.873,$$

necesitamos el área entre $x_1 = 24.5$ y $x_2 = 30.5$. Los valores z correspondientes son

$$z_1 = \frac{24.5 - 20}{3.873} = 1.16, \quad \text{y} \quad z_2 = \frac{30.5 - 20}{3.873} = 2.71.$$

La probabilidad de adivinar correctamente de 25 a 30 preguntas está dada por la región sombreada de la figura 6.27. De la tabla A.3 encontramos que

$$\begin{aligned} P(25 \leq X \leq 30) &= \sum_{x=25}^{30} b(x; 80, 0.25) \approx P(1.16 < Z < 2.71) \\ &= P(Z < 2.71) - P(Z < 1.16) = 0.9966 - 0.8770 = 0.1196. \end{aligned}$$

Ejercicios

6.23 Evalúe $P(1 \leq X \leq 4)$ para una variable binomial con $n = 15$ y $p = 0.2$ utilizando
a) la tabla A.1 del apéndice;
b) la aproximación de la curva normal.

6.24 Se lanza una moneda 400 veces. Utilice la aproximación de la curva normal para encontrar la probabilidad de obtener
a) entre 185 y 210 caras inclusive;
b) exactamente 205 caras;
c) menos de 176 o más de 227 caras.

6.25 Un proceso para fabricar un componente electrónico tiene 1% de defectuosos. Un plan de control de calidad consiste en seleccionar 100 artículos del proceso, y si ninguno está defectuoso, el proceso continúa. Use la aproximación normal a la binomial para encontrar
a) la probabilidad de que el proceso continúe con el plan de muestreo que se describe;
b) la probabilidad de que el proceso continúe aun si éste va mal (es decir, si la frecuencia de componentes defectuosos cambió a 5.0% de defectuosos).

6.26 Un proceso produce 10% de artículos defectuosos. Si se seleccionan al azar 100 artículos del proceso, ¿cuál es la probabilidad de que el número de defectuosos
a) exceda los 13?
b) sea menor que 8?

6.27 La probabilidad de que un paciente se recupere de una operación de corazón delicada es 0.9. De los siguientes 100 pacientes que se someten a esta operación, ¿cuál es la probabilidad de que
a) sobrevivan entre 84 y 95 inclusive?
b) sobrevivan menos de 86?

6.28 Investigadores de la Universidad George Washington y del Instituto Nacional de Salud informan que aproximadamente 75% de las personas creen que "los tranquilizantes funcionan muy bien para lograr que una persona esté más tranquila y relajada". De las siguientes 80 personas entrevistadas, ¿cuál es la probabilidad de que
a) al menos 50 tengan esta opinión?
b) a lo más 56 tengan esta opinión?

6.29 Si 20% de los residentes de una ciudad estadounidense prefieren un teléfono blanco sobre cualquier otro color disponible, ¿cuál es la probabilidad de que entre los siguientes 1000 teléfonos que se instalen en esa ciudad
a) entre 170 y 185 inclusive sean blancos?
b) al menos 210 pero no más de 225 sean blancos?

6.30 Un fabricante de medicamentos sostiene que cierto medicamento cura una enfermedad de la sangre, en promedio, en 80% de las veces. Para verificar la aseveración, inspectores gubernamentales utilizan el medicamento en una muestra de 100 individuos y deciden aceptar la afirmación si se curan 75 o más.
a) ¿Cuál es la probabilidad de que la aseveración se rechace cuando la probabilidad de curación es, de hecho, 0.8?
b) ¿Cuál es la probabilidad de que el gobierno acepte la afirmación cuando la probabilidad de curación sea tan baja como 0.7?

6.31 Un sexto de los estudiantes hombres de primer año que entran a una escuela estatal grande provienen de otros estados. Si los estudiantes se asignan a los dormitorios al azar, 180 en un edificio, ¿cuál es la probabilidad de que en un dormitorio dado al menos un quinto de los estudiantes provenga de otro estado?

6.32 Una compañía farmacéutica sabe que aproximadamente 5% de sus píldoras anticonceptivas tienen un ingrediente que está por debajo de la dosis mínima, lo que vuelve ineficaz a la píldora. ¿Cuál es la probabilidad de que menos de 10 en una muestra de 200 píldoras sean ineficaces?

6.33 Estadísticas publicadas por la Administración Nacional de Seguridad de Tránsito en Carreteras y el Consejo de Seguridad Nacional muestran que en una noche promedio de fin de semana, 1 de cada 10 conductores está ebrio. Si se verifican 400 conductores al azar la siguiente noche de sábado, ¿cuál es la probabilidad de que el número de conductores ebrios sea
a) menor que 32?
b) mayor que 49?
c) al menos 35 pero menos que 47?

6.34 Un par de dados se lanza 180 veces. ¿Cuál es la probabilidad de que ocurra un total de 7
a) al menos 25 veces?
b) entre 33 y 41 veces inclusive?
c) exactamente 30 veces?

6.35 Una compañía produce componentes para un motor. Las especificaciones de las partes sugieren que 95% de los artículos cumplen con las especificaciones. Las partes se embarcan en lotes de 100 para los clientes.
a) ¿Cuál es la probabilidad de que más de 2 artículos estén defectuosos en un lote dado?
b) ¿Cuál es la probabilidad de que más de 10 artículos estén defectuosos en un lote?

6.36 Una práctica común por parte de las compañías de aviación consiste en vender más boletos que el número real de asientos para un vuelo específico, porque los clientes que compran boletos no siempre se presentan a tomar el vuelo. Suponga que el porcentaje de pasajeros que no se presentan a la hora del vuelo es 2%. Para un vuelo particular con 197 asientos, se vendieron un total de 200 boletos. ¿Cuál es la probabilidad de que la aerolínea tenga una sobresaturación del vuelo?

6.37 El nivel de colesterol X en chicos de 14 años tiene aproximadamente una distribución normal, con una media de 170 y desviación estándar de 30.

a) Determine la probabilidad de que el nivel de colesterol de un chico de 14 años, elegido al azar, exceda 230.

b) En una escuela secundaria hay 300 chicos de 14 años. Determine la probabilidad de que por lo menos 8 niños tengan un nivel de colesterol que exceda 230.

6.38 Una compañía de telemarketing tiene una máquina especial para abrir cartas, que abre y extrae el contenido de los sobres. Si un sobre se coloca de forma incorrecta en la máquina, su contenido no puede extraerse o incluso podría dañarse. En este caso, se dice que la máquina "falló".

a) Si la máquina tiene una probabilidad de fallar de 0.01, ¿cuál es la probabilidad de que ocurra más de 1 falla en un lote de 20 sobres?

b) Si la probabilidad de falla de la máquina es 0.01 y se va a abrir un lote de 500 sobres, ¿cuál es la probabilidad de que ocurran más de 8 fallas?

6.6 Distribuciones gamma y exponencial

Aunque la distribución normal se puede utilizar para resolver muchos problemas en ingeniería y en la ciencia, hay aún numerosas situaciones que requieren diferentes tipos de funciones de densidad. Dos de estas funciones de densidad, las **distribuciones gamma** y **exponencial**, se estudiarán en esta sección.

Resulta que la distribución exponencial es un caso especial de la distribución gamma. Ambas encuentran un gran número de aplicaciones. Las distribuciones exponencial y gamma juegan un papel importante en la teoría de colas y en problemas de confiabilidad. Los tiempos entre llegadas en instalaciones de servicio, y los tiempos de operación antes del fallo de partes componentes y sistemas eléctricos, a menudo quedan bien modelados mediante la distribución exponencial. La relación entre la distribución gamma y la exponencial permite que la gamma se involucre en tipos de problemas similares. En la sección 6.7 se ofrecerán más detalles e ilustraciones.

La distribución gamma deriva su nombre de la bien conocida **función gamma**, que se estudia en muchas áreas de las matemáticas. Antes de que procedamos con la distribución gamma, repasemos esta función y algunas de sus propiedades importantes.

Definición 6.2: La **función gamma** se define como

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx, \qquad \text{para } \alpha > 0.$$

Al integrar por partes con $u = x^{\alpha-1}$ y $dv = e^{-x}\,dx$, obtenemos

$$\Gamma(\alpha) = -e^{-x}\,x^{\alpha-1}\Big|_0^\infty + \int_0^\infty e^{-x}(\alpha-1)x^{\alpha-2}\,dx = (\alpha-1)\int_0^\infty x^{\alpha-2}e^{-x}\,dx,$$

para $\alpha > 1$, que produce la fórmula recursiva

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1).$$

La aplicación repetida de la fórmula recursiva da

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\Gamma(\alpha-2) = (\alpha-1)(\alpha-2)(\alpha-3)\Gamma(\alpha-3),$$

y así sucesivamente. Observe que cuando $\alpha = n$, donde n es un entero positivo,

$$\Gamma(n) = (n-1)(n-2)\cdots(1)\Gamma(1).$$

Sin embargo, por la definición 6.2,

$$\Gamma(1) = \int_0^\infty e^{-x}\, dx = 1,$$

y de aquí,

$$\Gamma(n) = (n-1)!.$$

Una propiedad importante de la función gamma, que se deja al lector para su verificación (véase el ejercicio 6.41 de la página 205), es que $\Gamma(1/2) = \sqrt{\pi}$.

Incluiremos ahora la función gamma en nuestra definición de la distribución gamma.

Distribución Gamma

La variable aleatoria continua X tiene una **distribución gamma**, con parámetros α y β, si su función de densidad está dada por

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & x > 0, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

donde $\alpha > 0$ y $\beta > 0$.

En la figura 6.28 se muestran gráficas de varias distribuciones gamma para ciertos valores específicos de los parámetros α y β. La distribución gamma especial para la que $\alpha = 1$ se llama **distribución exponencial**.

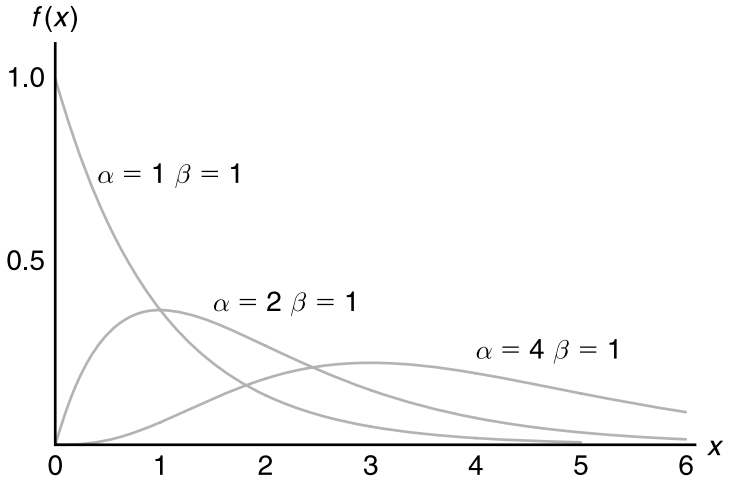


Figura 6.28: Distribuciones gamma.

Distribución exponencial

La variable aleatoria continua X tiene una **distribución exponencial**, con parámetro β, si su función de densidad está dada por

$$f(x;\beta) = \begin{cases} \frac{1}{\beta}e^{-x/\beta}, & x > 0, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

donde $\beta > 0$.

El siguiente teorema y corolario dan la media y la varianza de las distribuciones gamma y exponencial.

Teorema 6.3: La media y la varianza de la distribución gamma son

$$\mu = \alpha\beta, \quad \text{y} \quad \sigma^2 = \alpha\beta^2.$$

La demostración de este teorema se encuentra en el Apéndice A.28.

Corolario 6.1: La media y la varianza de la distribución exponencial son

$$\mu = \beta, \quad \text{y} \quad \sigma^2 = \beta^2.$$

Relación con el proceso de Poisson

Continuaremos con las aplicaciones de la distribución exponencial y después regresaremos a la distribución gamma. Las aplicaciones más importantes de la distribución exponencial son situaciones donde se aplica el proceso de Poisson (véase el capítulo 5). El lector debería recordar que el proceso de Poisson permite el uso de la distribución discreta llamada distribución de Poisson. Recuerde que la distribución de Poisson se utiliza para calcular la probabilidad de números específicos de "eventos", durante un *periodo o espacio* particulares. En muchas aplicaciones, el tiempo o la cantidad de espacio es la variable aleatoria. Por ejemplo, un ingeniero industrial se puede interesar en modelar el tiempo T entre llegadas a una intersección congestionada durante las horas de mayor afluencia en una ciudad grande. Una llegada representa el evento de Poisson.

La relación entre la distribución exponencial (a menudo denominada exponencial negativa) y el proceso de Poisson es bastante simple. En el capítulo 5 la distribución de Poisson se desarrolló como una distribución de un solo parámetro con parámetro λ, donde λ se interpreta como el número medio de eventos por *unidad de "tiempo"*. Considere ahora la variable aleatoria descrita por el tiempo que se requiere para que ocurra el primer evento. Usando la distribución de Poisson, encontramos que la probabilidad de que no ocurra algún evento, en el periodo hasta el tiempo t, está dada por

$$p(0;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Ahora podemos utilizar lo anterior y hacer que X sea el tiempo para el primer evento de Poisson. La probabilidad de que la duración del tiempo hasta el primer evento

exceda x es la misma que la probabilidad de que no ocurra algún evento de Poisson en x. Esto último, por supuesto, está dado por $e^{-\lambda x}$. Como resultado,

$$P(X > x) = e^{-\lambda x}.$$

Así, la función de distribución acumulada para X está dada por

$$P(0 \leq X \leq x) = 1 - e^{-\lambda x}.$$

Entonces, con la finalidad de que reconozcamos la presencia de la distribución exponencial, podemos diferenciar la función de distribución acumulada anterior para obtener la función de densidad

$$f(x) = \lambda e^{-\lambda x},$$

que es la función de densidad de la distribución exponencial con $\lambda = 1/\beta$.

6.7 Aplicaciones de las distribuciones exponencial y gamma

En lo anteriormente expuesto estudiamos las bases para la aplicación de la distribución exponencial en el "tiempo de llegada" o tiempo para problemas con eventos de Poisson. Ilustraremos aquí y después procederemos a discutir el papel de la distribución gamma en estas aplicaciones de modelado. Observe que la media de la distribución exponencial es el parámetro β, el recíproco del parámetro en la distribución de Poisson. El lector debería recordar que con frecuencia se dice que la distribución de Poisson no tiene memoria, lo cual implica que las ocurrencias en periodos sucesivos son independientes. El parámetro β importante es el tiempo medio entre eventos. En teoría de confiabilidad, donde la falla de equipo con frecuencia se ajusta a este proceso de Poisson, β se llama **tiempo medio entre fallas**. Muchas descomposturas de equipo siguen el proceso de Poisson y, por ello, se aplica la distribución exponencial. Otras aplicaciones incluyen tiempos de supervivencia en experimentos biomédicos y tiempo de respuesta de computadoras.

En el siguiente ejemplo mostramos una aplicación simple de la distribución exponencial a un problema de confiabilidad. La distribución binomial también juega un papel en la solución.

Ejemplo 6.17: Suponga que un sistema contiene cierto tipo de componente cuyo tiempo de operación antes del fallo, en años, está dado por T. La variable aleatoria T se modela bien mediante la distribución exponencial con tiempo medio de operación antes del fallo $\beta = 5$. Si se instalan 5 de estos componentes en diferentes sistemas, ¿cuál es la probabilidad de que al menos dos aún funcionen al final de 8 años?

Solución: La probabilidad de que un componente dado aún funcione después de 8 años está dada por

$$P(T > 8) = \frac{1}{5}\int_8^\infty e^{-t/5}\, dt = e^{-8/5} \approx 0.2.$$

Representemos con X el número de componentes que funcionan después de 8 años. Entonces, utilizando la distribución binomial

$$P(X \geq 2) = \sum_{x=2}^{5} b(x; 5, 0.2) = 1 - \sum_{x=0}^{1} b(x; 5, 0.2) = 1 - 0.7373 = 0.2627.$$

En el capítulo 3 hay ejercicios y ejemplos en que el lector ya trabajó con la distribución exponencial. Otros que implican problemas de tiempo de espera y de confiabilidad pueden encontrarse en el ejemplo 6.24 y en los ejercicios y ejercicios de repaso al final de este capítulo.

La propiedad de falta (o fallo) de memoria y su efecto en la distribución exponencial

Los tipos de aplicación de la distribución exponencial en la confiabilidad y en los problemas de tiempo de vida de una máquina o de un componente están influidos por la propiedad de **falta de memoria** (*memoryless*) de la distribución exponencial. Por ejemplo, en el caso de un componente electrónico donde la distribución del tiempo de vida sigue una distribución exponencial, la probabilidad de que el componente dure, por ejemplo, t horas, esto es, $P(X \geq t)$, es la misma que la probabilidad condicional

$$P(X \geq t_0 + t \mid X \geq t_0).$$

De manera que si el componente "alcanza" las t_0 horas, la probabilidad de durar t horas adicionales es la misma que la probabilidad de durar t horas. Así que no hay "castigo" a través del desgaste como resultado de durar las primeras t_0 horas. Por lo tanto, la distribución exponencial es más adecuada cuando se justifica la propiedad de falta de memoria. No obstante, si la falla del componente es resultado del desgaste gradual o lento (como en el caso del desgaste mecánico), entonces no se aplica la distribución exponencial, y la distribución gamma o de Weibull (sección 6.10) serían más adecuadas.

La importancia de la distribución gamma radica en el hecho de que define una familia de la que otras distribuciones son casos especiales. Pero la gamma misma tiene aplicaciones importantes en tiempo de espera y teoría de confiabilidad. Mientras que la distribución exponencial describe el tiempo hasta la ocurrencia de un evento de Poisson (o el tiempo entre eventos de Poisson), el tiempo (o espacio) que transcurre hasta que *ocurre un número específico de eventos de Poisson* es una variable aleatoria, cuya función de densidad está descrita por la de la distribución gamma. Este número específico de eventos es el parámetro α en la función de densidad gamma. Así se vuelve fácil comprender que cuando $\alpha = 1$, ocurre el caso especial de la distribución exponencial. La densidad gamma se puede desarrollar de su relación con el proceso de Poisson de la misma manera en que lo hicimos con la densidad exponencial. Los detalles se dejan al lector. El siguiente es un ejemplo numérico del uso de la distribución gamma en una aplicación de tiempo de espera.

Ejemplo 6.18: Suponga que las llamadas telefónicas que llegan a un conmutador particular siguen un proceso de Poisson con un promedio de 5 llamadas entrantes por minuto. ¿Cuál es la probabilidad de que transcurra a lo más un minuto hasta que lleguen 2 llamadas al conmutador?

Solución: El proceso de Poisson se aplica al tiempo que pasa hasta la ocurrencia de 2 eventos de Poisson que siguen una distribución gamma con $\beta = 1/5$ y $\alpha = 2$. Denote con X el tiempo en minutos que transcurre antes de que lleguen 2 llamadas. La probabilidad que se requiere está dada por

$$P(X \leq 1) = \int_0^1 \frac{1}{\beta^2} x e^{-x/\beta}\, dx = 25 \int_0^1 x e^{-5x}\, dx = 1 - e^{-5}(1 + 5) = 0.96.$$

Mientras el origen de la distribución gamma trata con el tiempo (o espacio) hasta la ocurrencia de α eventos de Poisson, hay muchos ejemplos donde una distribución

gamma trabaja muy bien aunque no exista una estructura de Poisson clara. Esto es particularmente cierto para problemas de **tiempo de supervivencia** en aplicaciones de ingeniería y biomédicas.

Ejemplo 6.19: En un estudio biomédico con ratas se utiliza una investigación de respuesta a la dosis para determinar el efecto de la dosis de un tóxico en su tiempo de supervivencia. El tóxico es uno que se descarga con frecuencia en la atmósfera desde el combustible de los aviones. Para cierta dosis del tóxico el estudio determina que el tiempo de supervivencia, en semanas, tiene una distribución gamma con $\alpha = 5$ y $\beta = 10$. ¿Cuál es la probabilidad de que una rata no sobreviva más de 60 semanas?

Solución: Sea la variable aleatoria X el tiempo de supervivencia (tiempo para morir). La probabilidad que se requiere es

$$P(X \leq 60) = \frac{1}{\beta^5} \int_0^{60} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(5)}\, dx.$$

La integral anterior se puede resolver mediante la **función gamma incompleta**, que resulta ser la función de distribución acumulada para la distribución gamma. Esta función se escribe como

$$F(x;\alpha) = \int_0^x \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy.$$

Si hacemos, $y = x/\beta$, so $x = \beta y$, tenemos

$$P(X \leq 60) = \int_0^6 \frac{y^4 e^{-y}}{\Gamma(5)}\, dy,$$

que se denota como $F(6; 5)$ en la tabla de la función gamma incompleta del Apéndice A.24. Observe que esto permite un cálculo rápido de las probabilidades para la distribución gamma. De hecho, para este problema la probabilidad de que la rata no sobreviva más de 60 días está dada por

$$P(X \leq 60) = F(6;5) = 0.715.$$

Ejemplo 6.20: A partir de los datos disponibles se sabe que la longitud de tiempo en meses entre las quejas de los clientes sobre cierto producto es una distribución gamma con $\alpha = 2$ y $\beta = 4$. Se realizaron cambios que implican intensificar los requerimientos del control de calidad. De acuerdo con tales cambios, pasan 20 meses antes de la primera queja. ¿Parecería que resultó eficaz la intensificación del control de calidad?

Solución: Sea X el tiempo para la primera queja, la cual, bajo las condiciones anteriores a los cambios, sigue una distribución gamma con $\alpha = 2$ y $\beta = 4$. La pregunta se centra alrededor de qué tan raro es $X \geq 20$ dado que α y β permanecen con los valores 2 y 4, repectivamente. En otras palabras, bajo las condiciones anteriores es razonable un "tiempo para la queja" tan grande como 20 meses? Siguiendo la solución del ejemplo 6.19, por lo tanto, necesitamos

$$P(X \geq 20) = 1 - \frac{1}{\beta^\alpha} \int_0^{20} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)} dx.$$

De nuevo, usando $y = x/\beta$, tenemos

$$P(X \geq 20) = 1 - \int_0^5 \frac{y e^{-y}}{\Gamma(2)}\, dy = 1 - F(5;2) = 1 - 0.96 = 0.04,$$

donde se encuentra $F(5; 2) = 0.96$ a partir de la tabla A.24.

Como resultado, concluimos que las condiciones de la distribución gamma con $\alpha = 2$ y $\beta = 4$ no están apoyadas por los datos de que un tiempo observado para la queja sea tan grande como 20 meses. Entonces, es razonable concluir que el trabajo de control de calidad resultó eficaz.

Ejemplo 6.21: Considere el ejercicio 3.31 de la página 90. Con base en pruebas de gran alcance se determinó que el tiempo Y en años antes de que se requiera una reparación mayor para cierta lavadora se caracteriza por la función de densidad

$$f(x) = \begin{cases} \frac{1}{4}e^{-y/4}, & y \geq 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Observe que es una exponencial con $\mu = 4$ años. La maquina se considera una ganga si es improbable que requiera una reparación mayor antes del sexto año. Entonces, ¿cuál es la probabilidad de $P(Y > 6)$? Además, ¿cuál es la probabilidad de ocurra una reparación mayor durante el primer año?

Solución: Considere la función de distribución acumulada $F(y)$ para la distribución exponencial

$$F(y) = \frac{1}{\beta}\int_0^y e^{-t/\beta}\, dt = 1 - e^{-y/\beta}.$$

De manera que

$$P(Y > 6) = 1 - F(6) = e^{-3/2} = 0.2231.$$

Por lo tanto, la probabilidad de que requiera una reparación mayor después de seis años es 0.223. Desde luego, requerirá la reparación antes del año seis con probabilidad de 0.777. Así, se podría concluir que la máquina no es realmente una ganga. La probabilidad de que ocurra una reparación mayor durante el primer año es

$$P(Y < 1) = 1 - e^{-1/4} = 1 - 0.779 = 0.221.$$

6.8 Distribución chi cuadrada

Otro caso especial muy importante de la distribución gamma se obtiene al hacer $\alpha = v/2$ y $\beta = 2$, donde v es un entero positivo. Este resultado se llama distribución **chi cuadrada**. La distribución tiene un solo parámetro, v, llamado **grados de libertad**.

Distribución chi cuadrada

La variable aleatoria continua X tiene una distribución chi cuadrada, con v grados de libertad, si su función de densidad está dada por

$$f(x; v) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)}x^{v/2-1}e^{-x/2}, & x > 0, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

donde v es un entero positivo.

La distribución chi cuadrada juega un papel fundamental en la inferencia estadística. Tiene una aplicación considerable tanto en la metodología como en la teoría. Aunque no estudiaremos con detalle sus aplicaciones en este capítulo, es importante tener en cuenta que los capítulos 8, 9 y 16 contienen aplicaciones importantes. La distribución chi cuadrada es un componente importante de la prueba de hipótesis y la estimación estadísticas.

Los temas que tratan con distribuciones de muestreo, análisis de varianza y estadística no paramétrica implican el uso extenso de la distribución chi cuadrada.

Teorema 6.4: La media y la varianza de la distribución chi cuadrada son

$$\mu = v \quad \text{y} \quad \sigma^2 = 2v.$$

6.9 Distribución logarítmica normal

La distribución logarítmica normal se utiliza en una amplia variedad de aplicaciones. La distribución se aplica en casos donde una transformación logarítmica natural tiene como resultado una distribución normal.

Distribución logarítmica normal

La variable aleatoria continua X tiene una distribución logarítmica normal si la variable aleatoria $Y = \ln(X)$ tiene una distribución normal con media μ y desviación estándar σ. La función de densidad de X que resulta es

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Las gráficas de las distribuciones logarítmicas normales se ilustran en la figura 6.29.

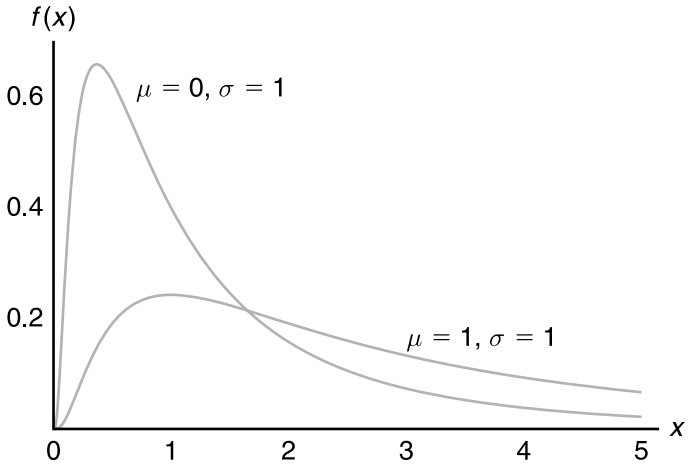


Figura 6.29: Distribuciones logarítmicas normales.

Teorema 6.5: La media y la varianza de la distribución logarítmica normal son

$$\mu = e^{\mu+\sigma^2/2} \quad \text{y} \quad \sigma^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

La función de distribución acumulada es bastante simple debido a su relación con la distribución normal. El uso de la función de distribución se ilustra con el siguiente ejemplo.

Ejemplo 6.22: Se sabe que históricamente la concentración de contaminantes producidos por plantas químicas exhiben un comportamiento que se parece a una distribución logarítmica normal. Esto es importante cuando se consideran problemas respecto de la obediencia de las regulaciones gubernamentales. Suponga que la concentración de cierto contaminante, en partes por millón, tiene una distribución logarítmica normal con parámetros $\mu = 3.2$ y $\sigma = 1$. ¿Cuál es la probabilidad de que la concentración exceda 8 partes por millón?

Solución: Sea la variable aleatoria X la concentración de contaminantes

$$P(X > 8) = 1 - P(X \leq 8).$$

Como $\ln(X)$ tiene una distribución normal con media $\mu = 3.2$ y desviación estándar $\sigma = 1$,

$$P(X \leq 8) = \Phi\left[\frac{\ln(8) - 3.2}{1}\right] = \Phi(-1.12) = 0.1314.$$

Aquí, utilizamos la notación Φ para denotar la función de distribución acumulada de la distribución normal estándar. Como resultado, la probabilidad de que la concentración del contaminante exceda 8 partes por millón es 0.1314.

Ejemplo 6.23: La vida, en miles de millas, de un cierto tipo de control electrónico para locomotoras tiene una distribución logarítmica normal aproximada con $\mu = 5.149$ y $\sigma = 0.737$. Encuentre el quinto percentil de la vida de esa locomotora.

Solución: A partir de la tabla A.3, sabemos que $P(Z < -1.645) = 0.05$. Denote como X la vida de la locomotora. Puesto que $\ln(X)$ tiene una distribución normal con media $\mu = 5.149$ y $\sigma = 0.737$, el quinto percentil de X se calcula como

$$\ln(x) = 5.149 + (0.737)(-1.645) = 3.937.$$

Entonces, $x = 51.265$. Esto significa que sólo el 5% de las locomotoras tendrán un tiempo de vida menor que 51,265 miles de millas.

6.10 Distribución de Weibull (opcional)

La tecnología actual nos permite diseñar muchos sistemas complicados cuya operación, o quizá seguridad, depende de la confiabilidad de los diversos componentes que conforman los sistemas. Por ejemplo, un fusible puede quemarse, una columna de acero puede torcerse o un dispositivo sensor de calor puede fallar. Componentes idénticos sujetos a idénticas condiciones ambientales fallarán en momentos diferentes

e impredecibles. Ya examinamos el papel que las distribuciones gamma y exponencial juegan en estos tipos de problemas. Otra distribución que se ha utilizado con amplitud en años recientes para tratar con tales problemas es la **distribución de Weibuill**, que presentó el físico sueco Waloddi Weibull en 1939.

Distribución de Weibull

La variable aleatoria continua X tiene una **distribución de Weibull**, con parámetros α y β si su función de densidad está dada por

$$f(x;\alpha,\beta)=\begin{cases}\alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, & x>0,\\ 0, & \text{en cualquier otro caso,}\end{cases}$$

donde $\alpha>0$ y $\beta>0$.

En la figura 6.30 se ilustran las gráficas de la distribución de Weibull para $\alpha=1$ y diversos valores del parámetro β. Vemos que las curvas cambian de forma de manera considerable para diferentes valores del parámetro β. Si hacemos $\beta=1$, la distribución de Weibull se reduce a la distribución exponencial. Para valores de $\beta>1$, las curvas se vuelven un poco en forma de campana y se asemejan a las curvas normales, pero muestran algo de asimetría.

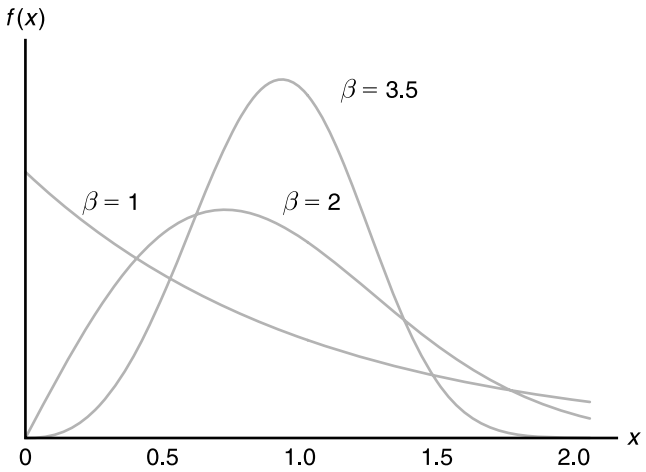


Figura 6.30: Distribuciones de Weibull ($\alpha=1$).

La media y varianza de la distribución de Weibull se establecen en el siguiente teorema. Se solicita al lector que dé la demostración en el ejercicio 6.48 de la página 205.

Teorema 6.6: La media y la varianza de la distribución de Weibull son

$$\mu=\alpha^{-1/\beta}\Gamma\left(1+\frac{1}{\beta}\right) \quad \text{y} \quad \sigma^2=\alpha^{-2/\beta}\left\{\Gamma\left(1+\frac{2}{\beta}\right)-\left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right\}.$$

Al igual que la distribución gamma y la exponencial, la distribución de Weibull también se aplica a problemas de confiabilidad y de prueba de vida como los de **tiempo de operación antes del fallo** o la **duración de la vida** de un componente,

que se miden desde algún tiempo específico hasta que falla. Representemos este tiempo de operación antes del fallo mediante la variable aleatoria continua T, con función de densidad de probabilidad $f(t)$, donde $f(t)$ es la distribución de Weibull. Ésta tiene la flexibilidad inherente de no requerir la propiedad de falta de memoria de la distribución exponencial. La función de distribución acumulada (fda) para la distribución de Weibull se puede escribir en forma cerrada y, en efecto, es muy útil para calcular probabilidades.

fda para la distribución de Weibull

La función de distribución acumulada para la distribución de Weibull está dada por

$$F(x) = 1 - e^{-\alpha x^{\beta}}, \qquad \text{para } x \geq 0,$$

para $\alpha > 0$ y $\beta > 0$.

Ejemplo 6.24: El tiempo de vida X, en horas, de un artículo en el taller mecánico tiene una distribución de Weibull con $\alpha = 0.01$ y $\beta = 2$. ¿Cuál es la probabilidad de que falle antes de ocho horas de uso?

Solución: $P(X < 8) = F(8) = 1 - e^{-(0.01)8^2} = 1 - 0.527 = 0.473.$

La tasa de falla para la distribución de Weibull

Cuando se aplica la distribución de Weibull, con frecuencia es útil determinar la **tasa de falla** (algunas veces denominada tasa de riesgo) para tener conocimiento del desgaste o deterioro del componente. Definamos primero la confiabilidad de un componente o producto como la *probabilidad de que funcione adecuadamente por al menos un tiempo específico bajo condiciones experimentales específicas*. Por lo tanto, si $R(t)$ se define como la confiabilidad del componente dado en el tiempo t, escribimos

$$R(t) = P(T > t) = \int_t^{\infty} f(t)\, dt = 1 - F(t),$$

donde $F(t)$ es la función de distribución acumulada de T. La probabilidad condicional de que un componente caiga en el intervalo de $T = t$ a $T = t + \Delta t$, dado que sobrevive al tiempo t, es

$$\frac{F(t + \Delta t) - F(t)}{R(t)}.$$

Al dividir esta proporción entre Δt y tomar el límite cuando $\Delta t \to 0$, obtenemos la **tasa de falla**, denotada con $Z(t)$. De aquí,

$$Z(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{R(t)} = \frac{F'(t)}{R(t)} = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)},$$

que expresa la tasa de falla en términos de la distribución del tiempo de operación antes del fallo.

Como $Z(t) = f(t)/[1 - F(t)]$, entonces la tasa de falla está dada como sigue:

Tasa de falla para la distribución de Weibull

La tasa de falla en el tiempo t para la distribución de Weibull está dada por

$$Z(t) = \alpha\beta t^{\beta - 1}, \qquad t > 0.$$

Interpretación de la tasa de falla

La cantidad $Z(t)$ es bien llamada tasa de falla porque en verdad cuantifica la tasa de cambio con el tiempo de la probabilidad condicional de que el componente dure una Δt adicional *dado que ha durado el tiempo t*. Es importante la tasa de disminución (o crecimiento) con el tiempo. Los siguientes puntos son fundamentales.

a) Si $\beta = 1$, la tasa de falla $= \alpha$, una constante. Esto, como se indicó anteriormente, es el caso especial de la distribución exponencial en que predomina la falta de memoria.

b) Si $\beta > 1$, $Z(t)$ es una función creciente de t que indica que el componente se desgasta con el tiempo.

c) Si $\beta < 1$, $Z(t)$ es una función decreciente de tiempo y, por lo tanto, el componente se fortalece o endurece con el paso del tiempo.

Por ejemplo, el artículo en el taller mecánico del ejemplo 6.24 tiene $\beta = 2$ y por consiguiente se desgasta con el tiempo. De hecho, la función de la tasa de falla está dada por $Z(t) = .02t$. Por otro lado, suponga un parámtero donde $\beta = 3/4$ y $\alpha = 2$. $Z(t) = 1.5/t^4$ y, por lo tanto, el componente se hace más fuerte con el tiempo.

Ejercicios

6.39 Si una variable aleatoria X tiene una distribución gamma con $\alpha = 2$ y $\beta = 1$, encuentre $P(1.8 < X < 2.4)$.

6.40 En cierta ciudad, el consumo diario de agua (en millones de litros) sigue aproximadamente una distribución gamma con $\alpha = 2$ y $\beta = 3$. Si la capacidad diaria de dicha ciudad es 9 millones de litros de agua, ¿cuál es la probabilidad de que en cualquier día dado el suministro de agua sea inadecuado?

6.41 Utilice la función gamma con $y = \sqrt{2x}$ para demostrar que $\Gamma(1/2) = \sqrt{\pi}$.

6.42 Suponga que el tiempo, en horas, que toma reparar una bomba de calor es una variable aleatoria X que tiene una distribución gamma con parámetros $\alpha = 2$ y $\beta = 1/2$. ¿Cuál es la probabilidad de que la siguiente llamada de servicio requiera

a) a lo más 1 hora para reparar la bomba de calor?

b) al menos 2 horas para reparar la bomba de calor?

6.43 *a*) Encuentre la media y la varianza del consumo diario de agua del ejercicio 6.40.

b) De acuerdo con el teorema de Chebyshev, ¿hay una probabilidad de al menos 3/4 de que el consumo de agua en cualquier día dado caiga dentro de un intervalo? ¿De cuál?

6.44 En cierta ciudad, el consumo diario de energía eléctrica, en millones de kilowatts-hora, es una variable aleatoria X que tiene una distribución gamma con media $\mu = 6$ y varianza $\sigma^2 = 12$.

a) Encuentre los valores de α y β.

b) Encuentre la probabilidad de que en cualquier día dado el consumo de energía diario exceda los 12 millones de kilowatts-hora.

6.45 La longitud de tiempo para que un individuo sea atendido en una cafetería es una variable aleatoria que tiene una distribución exponencial con una media de 4 minutos. ¿Cuál es la probabilidad de que una persona sea atendida en menos de 3 minutos en, al menos, 4 de los siguientes 6 días?

6.46 La vida, en años, de cierto interruptor eléctrico tiene una distribución exponencial con una vida promedio de $\beta = 2$. Si 100 de estos interruptores se instalan en diferentes sistemas, ¿cuál es la probabilidad de que a lo más 30 fallen durante el primer año?

6.47 Suponga que la vida de servicio, en años, de la batería de un aparato para reducir la sordera es una variable aleatoria que tiene una distribución de Weibull con $\alpha = 1/2$ y $\beta = 2$.

a) ¿Cuánto tiempo se puede esperar que dure tal batería?

b) ¿Cuál es la probabilidad de que tal batería esté en operación después de 2 años?

6.48 Derive la media y la varianza de la distribución de Weibull.

6.49 Las vidas de ciertas juntas para automóvil tienen la distribución de Weibull con tasa de falla $Z(t) = 1/\sqrt{t}$. Encuentre la probabilidad de que tal junta aún esté en uso después de 4 años.

6.50 La variable aleatoria continua X tiene la distribución beta con parámetros α y β si su función de densidad está dada por

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

donde $\alpha > 0$ y $\beta > 0$. Si la proporción de una marca de televisores que requiere servicio durante el primer año de operación es una variable aleatoria que tiene una distribución beta con $\alpha = 3$ y $\beta = 2$, ¿cuál es la probabilidad de que al menos 80% de los nuevos modelos de esta marca que se vendieron este año requerirán servicio durante su primer año de operación?

6.51 En una actividad de investigación biomédica se determinó que el tiempo de supervivencia, en semanas, de un animal cuando se le somete a cierta exposición de radiación gamma tiene una distribución gamma con $\alpha = 5$ y $\beta = 10$.

a) ¿Cuál es el tiempo medio de supervivencia de un animal seleccionado al azar del tipo que se utilizó en el experimento?

b) ¿Cuál es la desviación estándar del tiempo de supervivencia?

c) ¿Cuál es la probabilidad de que un animal sobreviva más de 30 semanas?

6.52 Se sabe que el tiempo de duración, en semanas, de cierto tipo de transistor sigue una distribución gamma con media de 10 semanas y desviación estándar de $\sqrt{50}$ semanas.

a) ¿Cuál es la probabilidad de que el transistor dure a lo más 50 semanas?

b) ¿Cuál es la probabilidad de que el transistor no sobreviva las primeras 10 semanas?

6.53 El tiempo de respuesta de una computadora es una aplicación importante de las distribuciones gamma y exponencial. Suponga que un estudio de cierto sistema de computadoras revela que el tiempo de respuesta, en segundos, tiene una distribución exponencial con una media de 3 segundos.

a) ¿Cuál es la probabilidad de que el tiempo de respuesta exceda 5 segundos?

b) ¿Cuál es la probabilidad de que el tiempo de respuesta exceda 10 segundos?

6.54 Los datos de porcentaje a menudo siguen una distribución logarítmica normal. Se estudia el uso promedio de potencia (dB por hora) para una compañía específica y se sabe que tiene una distribución logarítmica normal con parámetros $\mu = 4$ y $\sigma = 2$. ¿Cuál es la probabilidad de que la compañía utilice más de 270 dB durante cualquier hora particular?

6.55 Para el ejercicio 6.54, ¿cuál es el uso de potencia media (dB promedio por hora)? ¿Cuál es la varianza?

6.56 El número de automóviles que llegan a cierta intersección por minuto tiene una distribución de Poisson con una media de 5. El interés se centra alrededor del tiempo que transcurre antes de que 10 automóviles aparezcan en la intersección.

a) ¿Cuál es la probabilidad de que más de 10 automóviles aparezcan en la intersección durante cualquier minuto dado?

b) ¿Cuál es la probabilidad de que se requieran más de 2 minutos antes de que lleguen 10 automóviles?

6.57 Considere la información del ejercicio 6.56.

a) ¿Cuál es la probabilidad de que transcurra más de 1 minuto entre llegadas?

b) ¿Cuál es el número medio de minutos que transcurren entre llegadas?

6.58 Muestre que la función de tasa de falla está dada por

$$Z(t) = \alpha\beta t^{\beta-1}, \quad t > 0,$$

y sólo si la distribución del tiempo de operación antes del fallo es la distribución de Weibull

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}, \quad t > 0.$$

Ejercicios de repaso

6.59 De acuerdo con un estudio publicado por un grupo de sociólogos de la Universidad de Massachusetts, aproximadamente 49% de los consumidores de Valium en el estado de Massachusetts son empleados de oficina. ¿Cuál es la probabilidad de que entre 482 y 510, inclusive, de los siguientes 1000 consumidores de Valium seleccionados al azar de dicho estado sean empleados de oficina?

6.60 La distribución exponencial se aplica con frecuencia a los tiempos de espera entre éxitos en un proceso de Poisson. Si el número de llamadas que se reciben por hora en un servicio de contestación telefónica es una variable aleatoria de Poisson con parámetro $\lambda = 6$, sabemos que el tiempo, en horas, entre llamadas sucesivas tiene una distribución exponencial con parámetro $\beta = 1/6$. ¿Cuál es la probabilidad de esperar más de 15 minutos entre cualesquiera dos llamadas sucesivas?

6.61 Cuando α es un entero positivo n, la distribución gamma, también se conoce como **distribución de**

Erlang. Al hacer $\alpha = n$ en la distribución gamma de la página 195, la distribución de Erlang es

$$f(x) = \begin{cases} \frac{x^{n-1}e^{-x/\beta}}{\beta^n(n-1)!}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Se puede mostrar que si los tiempos entre eventos sucesivos son independientes, y cada uno tiene una distribución exponencial con parámetro β, entonces el tiempo de espera total X transcurrido hasta que ocurran n eventos tiene la distribución de Erlang. Con referencia al ejercicio de repaso 6.60, ¿cuál es la probabilidad de que las siguientes 3 llamadas se reciban dentro de los siguientes 30 minutos?

6.62 Un fabricante de cierto tipo de máquina grande desea comprar remaches de uno de dos fabricantes. Es importante que la resistencia a la rotura de cada remache exceda 10,000 psi. Dos fabricantes (A y B) ofrecen este tipo de remache y ambos tienen remaches cuya resistencia a la rotura está distribuida de forma normal. Las resistencias a la rotura medias para los fabricantes A y B son 14,000 y 13,000 psi, respectivamente. Las desviaciones estándar son 2000 y 1000 psi, respectivamente. ¿Cuál fabricante producirá, en promedio, el menor número de remaches defectuosos?

6.63 De acuerdo con un censo reciente, casi 65% de todos los hogares en Estados Unidos se componen de una o dos personas. Suponiendo que este porcentaje aún sea válido en la actualidad, ¿cuál es la probabilidad de que entre 590 y 625 inclusive de los siguientes 1000 hogares seleccionados al azar en Estados Unidos consistan en una o dos personas?

6.64 La vida de cierto tipo de dispositivo tiene una tasa de falla anunciada de 0.01 por hora. La tasa de falla es constante y se aplica la distribución exponencial.

a) ¿Cuál es el tiempo medio de operación antes del fallo?

b) ¿Cuál es la probabilidad de que pasen 200 horas antes de que se observe una falla?

6.65 En una planta de procesamiento químico es importante que el rendimiento de cierto tipo de producto en lote se mantenga por arriba de 80%. Si permanece por debajo de 80% por un tiempo prolongado, la compañía pierde dinero. Los lotes producidos ocasionalmente con defectos son de poco interés. Pero si varios lotes por día resultan defectuosos, la planta se detiene y se llevan a cabo ajustes. Se sabe que el rendimiento se distribuye normalmente con desviación estándar de 4%.

a) ¿Cuál es la probabilidad de una "falsa alarma" (rendimiento por debajo de 80%) cuando el rendimiento medio es de 85%?

b) ¿Cuál es la probabilidad de que un lote producido tenga un rendimiento que exceda 80% cuando de hecho el rendimiento medio es de 79%?

6.66 Considere la tasa de falla de un componente eléctrico de una vez cada 5 horas. Es importante considerar el tiempo que transcurre para que fallen 2 componentes.

a) Suponiendo que se aplica la distribución gamma, ¿cuál es el tiempo medio que transcurre para la falla de 2 componentes?

b) ¿Cuál es la probabilidad de que transcurran 12 horas antes de que fallen 2 componentes?

6.67 Se establece que el alargamiento (elongación) de una barra de acero bajo una carga particular se distribuye normalmente con una media de 0.05 pulgadas y $\sigma = 0.01$ pulgadas. Encuentre la probabilidad de que el alargamiento esté

a) por arriba de 0.1 pulgadas;

b) por abajo de 0.04 pulgadas;

c) entre 0.025 y 0.065 pulgadas.

6.68 Se sabe que un satélite controlado tiene un error (distancia del objetivo) que se distribuye normalmente con media cero y desviación estándar de 4 pies. El fabricante del satélite define un "éxito" como un disparo en el cual el satélite llega a 10 pies del objetivo. Calcule la probabilidad de que el satélite falle.

6.69 Un técnico planea probar cierto tipo de resina desarrollada en el laboratorio para determinar la naturaleza del tiempo que transcurre antes de que se realice la unión. Se sabe que el tiempo medio para la unión es 3 horas y la desviación estándar es 0.5 horas. Un producto se considerará indeseable si el tiempo de unión es menos de 1 hora o más de 4 horas. Comente sobre la utilidad de la resina. ¿Con qué frecuencia su desempeño se considera indeseable? Suponga que el tiempo para la unión se distribuye normalmente.

6.70 Considere la información del ejercicio de repaso 6.64. ¿Cuál es la probabilidad de que transcurran menos de 200 horas antes de que ocurran 2 fallas?

6.71 Para el ejercicio de repaso 6.70, ¿cuál es la media y la varianza del tiempo que transcurre antes de que ocurran 2 fallas?

6.72 Se sabe que la tasa promedio de uso de agua (miles de galones por hora) en cierta comunidad implica la distribución logarítmica normal con parámetros $\mu = 5$ y $\sigma = 2$. Para propósitos de planeación es importante lograr un buen juicio sobre los periodos de alta utilización. ¿Cuál es la probabilidad de que, para cualquier hora dada, se usen 50,000 galones de agua?

6.73 Para el ejercicio de repaso 6.72, ¿cuál es la media del uso de agua por hora promedio en miles de galones?

6.74 En el ejercicio 6.52 de la página 206, se supone que la duración de un transistor tiene una distribución

gamma con media de 10 semanas y desviación estándar de $\sqrt{50}$ semanas. Si la suposición de la distribución gamma es incorrecta y la distribución es normal,

a) ¿cuál es la probabilidad de que el transistor dure a lo más 50 semanas?

b) ¿cuál es la probabilidad de que el transistor no sobreviva las primeras 10 semanas?

c) comente la diferencia entre sus resultados aquí y los que se encontraron en el ejercicio 6.52 de la página 206.

6.75 Considere el ejercicio 6.50 de la página 206. La distribución beta tiene amplia aplicación en problemas de confiabiliad, donde la varaible aleatoria fundamental es una proporción en el contexto práctico que se ilustra en el ejemplo. Al respecto, considere el ejercicio de repaso 3.75 de la página 105. Las impurezas en el lote del producto de un proceso químico reflejan un problema grave. Se sabe que la proporción de impurezas Y en un lote tiene la función de densidad

$$f(y) = \begin{cases} 10(1-y)^9, & 0 \le y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Verifique que la anterior sea una función de densidad válida.

b) ¿Cuál es la probabilidad de que un lote se considere no aceptable (es decir, $Y > 0.6$)?

c) ¿Cuáles son los parámetros α y β de la distribución beta que se ilustra aquí?

d) La media de la distribución beta es $\frac{\alpha}{\alpha+\beta}$. ¿Cuál es la proporción media de impurezas en el lote?

e) La varianza de una variable aleatoria beta distribuida es

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

¿Cuál es la varianza de Y en este problema?

6.76 Considere ahora el ejercicio de repaso 3.76 de la página 105. La función de densidad del tiempo Z en minutos entre las llamadas a un sistema de alimentación eléctrica está dada por

$$f(z) = \begin{cases} \frac{1}{10}e^{-z/10}, & 0 < z < \infty, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) ¿Cuál es el tiempo medio entre llamadas?

b) ¿Cuál es la varianza en el tiempo entre llamadas?

c) ¿Cuál es la probabilidad de que el tiempo entre llamadas supere la media?

6.77 Considere el ejercicio de repaso 6.76. Dada la suposición de la distribución exponencial, ¿cuál es el número medio de llamadas por hora? ¿Cuál es la varianza en el número de llamadas por hora?

6.78 En un proyecto experimental sobre el factor humano, se determinó que el tiempo de reacción de un piloto ante un estímulo visual está distribuido normalmente con una media de 1/2 segundo y una desviación estándar de 2/5 de segundo.

a) ¿Cuál es la probabilidad de que una reacción del piloto tome más de 0.3 segundos?

b) ¿Qué tiempo de reacción se excede 95% de las veces?

6.79 La longitud de tiempo entre fallas de una pieza esencial de equipo es importante en la decisión del uso de equipo auxiliar. Un ingeniero cree que el mejor "modelo" para el tiempo entre fallas de un generador es la distribución exponencial con una media de 15 días.

a) Si el generador acaba de fallar, ¿cuál es la probabilidad de que falle en los siguientes 21 días?

b) ¿Cuál es la probabilidad de que el generador funcionará durante 30 días sin falla?

6.80 El periodo de vida, en horas, de una broca en una operación mecánica tiene una distribución de Weibull con $\alpha = 2$ y $\beta = 50$. Encuentre la probabilidad de que la broca fallará antes de 10 horas de uso.

6.81 Encuentre la fda para la distribución de Weibull. [*Sugerencia:* En la definición de una fda, haga la transformación $z = y^\beta$.]

6.82 En el ejercicio de repaso 6.80, explique porque la naturaleza del escenario probablemente no se preste a la distribución exponencial.

6.83 A partir de la relación entre la variable aleatoria chi cuadrada y la variable aleatoria gamma, demuestre que la media de la variable aleatoria chi cuadrada es v y que la varianza es $2v$.

6.84 La longitud de tiempo, en segundos, que un usuario de computadora lee su correo electrónico se distribuye como una variable aleatoria logarítmica normal con $\mu = 1.8$ y $\sigma^2 = 4.0$.

a) ¿Cuál es la probabilidad de que el usuario lea el correo por más de 20 segundos? ¿Y por más de un minuto?

b) ¿Cuál es la probabilidad de que el usuario lea el correo durante una longitud de tiempo que sea igual a la media de la distribución logarítmica normal subyacente?

6.11 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

Muchos de los riesgos en el uso del material de este capítulo son muy similares a los del capítulo 5. Cuando se realiza un tipo de inferencia estadística, uno de los mayores usos incorrectos de la estadística consiste en suponer una distribución normal subyacente cuando en realidad no es normal. El lector estará expuesto a las prueba de hipótesis en los capítulos 10 a 15, en los cuales se hace la suposición de normalidad. Además, no obstante, se le recordará al lector que hay **pruebas de la bondad del ajuste**, además de las rutinas gráficas que se examinan en los capítulos 8 y 10, que permiten "verificar" los datos para determinar si es razonable la suposición de normalidad.

Capítulo 7

Funciones de variables aleatorias (opcional)

7.1 Introducción

Este capítulo contiene un amplio espectro de material. Los capítulos 5 y 6 tratan con tipos específicos de distribuciones, tanto discretas como continuas. Éstas son distribuciones que encuentran uso en muchas aplicaciones de temas como confiabilidad, control de calidad y muestreo de aceptación. En este capítulo comenzamos con un tema más general: las distribuciones de funciones de variables aleatorias. Se presentan las técnicas generales y se ilustran con ejemplos. Van seguidas por un concepto relacionado, *funciones generadoras de momentos*, que puede ser útil en el estudio de distribuciones de funciones lineales de variables aleatorias.

En los métodos estadísticos estándar, el resultado de la prueba de hipótesis estadísticas, la estimación o incluso las gráficas estadísticas no implica una sola variable aleatoria sino, más bien, *funciones de una o más variables aleatorias*. Como resultado, la inferencia estadística requiere la distribución de tales funciones. Por ejemplo, es común el uso de **promedios de variables aleatorias**. Además, son importantes las sumas y las combinaciones lineales más generales. Con frecuencia nos interesa la distribución de sumas de los cuadrados de variables aleatorias, en particular el uso de las técnicas del análisis de varianza que se estudian en los capítulos 11 a 14.

7.2 Transformaciones de variables

Con frecuencia, en estadística, se encuentra la necesidad de derivar la distribución de probabilidad de una función de una o más variables aleatorias. Por ejemplo, suponga que X es una variable aleatoria discreta con distribución de probabilidad $f(x)$ y suponga, además, que $Y = u(X)$ define una transformación uno a uno entre los valores de X y Y. Queremos encontrar la distribución de probabilidad de Y. Es importante notar que la transformación uno a uno implica que cada valor x está relacionado con un, y sólo un, valor $y = u(x)$, y que cada valor y está relacionado con un, y sólo un, valor $x = w(y)$, donde $w(y)$ se obtiene al resolver $y = u(x)$ para x en términos de y.

De nuestro estudio de las distribuciones de probabilidad discreta en el capítulo 3 resulta claro que la variable aleatoria Y toma el valor y cuando X toma el valor $w(y)$. En consecuencia, la distribución de probabilidad de Y está dada por

$$g(y) = P(Y = y) = P[x = w(y)] = f[w(y)].$$

Teorema 7.1: Suponga que X es una variable aleatoria **discreta** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una transformación uno a uno entre los valores de X y Y, de manera que la ecuación $y = u(x)$ se resuelva unívocamente para x en términos de y, digamos, $x = w(y)$. Entonces, la distribución de probabilidad de Y es

$$g(y) = f[w(y)]$$

Ejemplo 7.1: Sea X una variable aleatoria geométrica con distribución de probabilidad

$$f(x) = \frac{3}{4}\left(\frac{1}{4}\right)^{x-1}, \qquad x = 1, 2, 3, \ldots.$$

Encuentre la distribución de probabilidad de la variable aleatoria $Y = X^2$.

Solución: Como los valores de X son todos positivos, la transformación define una correspondencia uno a uno entre los valores x y y, $y = x^2$ y $x = \sqrt{y}$. De aquí,

$$g(y) = \begin{cases} f(\sqrt{y}) = \frac{3}{4}\left(\frac{1}{4}\right)^{\sqrt{y}-1}, & y = 1, 4, 9, \ldots, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Considere un problema donde X_1 y X_2 son dos variables aleatorias discretas con distribución de probabilidad conjunta $f(x_1, x_2)$ y que deseamos encontrar la distribución de probabilidad conjunta $g(y_1, y_2)$ de las dos variables aleatorias nuevas

$$Y_1 = u_1(X_1, X_2) \quad \text{y} \quad Y_2 = u_2(X_1, X_2).$$

que definen una transformación uno a uno entre el conjunto de puntos (x_1, x_2) y (y_1, y_2). Al resolver las ecuaciones $y_1 = u_1(x_1, x_2)$ y $y_2 = u_2(x_1, x_2)$ de forma simultánea, obtenemos la solución inversa única

$$x_1 = w_1(y_1, y_2) \quad \text{y} \quad x_2 = w_2(y_1, y_2).$$

De aquí las variables aleatorias Y_1 y Y_2 toman los valores y_1 y y_2, respectivamente, cuando X_1 toma el valor $w_1(y_1, y_2)$ y X_2 toma el valor $w_2 = (y_1, y_2)$. La distribución de probabilidad conjunta de Y_1, y Y_2 es, entonces,

$$\begin{aligned} g(y_1, y_2) &= P(Y_1 = y_1, Y_2 = y_2) \\ &= P[X_1 = w_1(y_1, y_2), X_2 = w_2(y_1, y_2)] \\ &= f[w_1(y_1, y_2), w_2(y_1, y_2)]. \end{aligned}$$

Teorema 7.2: Suponga que X_1 y X_2 son variables aleatorias **discretas** con distribución de probabilidad conjunta $f(x_1, x_2)$. Definamos con $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$ una transformación uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), de manera que las ecuaciones

$$y_1 = u_1(x_1, x_2) \quad \text{y} \quad y_2 = u_2(x_1, x_2)$$

se pueden resolver unívocamente para x_1 y x_2 en términos de y_1 y y_2, digamos $x_1 = w_1(y_1, y_2)$ y $x_2 = w_2(y_1, y_2)$. Entonces, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)].$$

El teorema 7.2 es bastante útil para encontrar la distribución de alguna variable aleatoria $Y_1 = u_1(X_1, X_2)$, donde X_1 y X_2 son variables aleatorias discretas con distribución de probabilidad conjunta $f(x_1, x_2)$. Definimos simplemente una segunda función, digamos $Y_2 = u_2(X_1, X_2)$, y mantenemos una correspondencia uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), y obtenemos la distribución de probabilidad conjunta $g(y_1, y_2)$. La distribución de Y_1 es precisamente la distribución marginal de $g(y_1, y_2)$ que se encuentra al sumar los valores y_2. Al denotar la distribución de Y_1 con $h(y_1)$, escribimos

$$h(y_1) = \sum_{y_2} g(y_1, y_2).$$

Ejemplo 7.2: Sean X_1 y X_2 dos variables aleatorias independientes que tienen distribuciones de Poisson con parámetros μ_1 y μ_2, respectivamente. Encuentre la distribución de la variable aleatoria $Y_1 = X_1 + X_2$.

Solución: Como X_1 y X_2 son independientes, podemos escribir

$$f(x_1, x_2) = f(x_1)f(x_2) = \frac{e^{-\mu_1}\mu_1^{x_1}}{x_1!}\frac{e^{-\mu_2}\mu_2^{x_2}}{x_2!} = \frac{e^{-(\mu_1+\mu_2)}\mu_1^{x_1}\mu_2^{x_2}}{x_1!x_2!},$$

donde $x_1 = 0, 1, 2, \ldots$ y $x_2 = 0, 1, 2, \ldots$ Definamos ahora una segunda variable aleatoria, digamos $Y_2 = X_2$. Las funciones inversas están dadas por $x_1 = y_1 - y_2$ y $x_2 = y_2$. Con el teorema 7.2 encontramos que la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = \frac{e^{-(\mu_1+\mu_2)}\mu_1^{y_1-y_2}\mu_2^{y_2}}{(y_1 - y_2)!y_2!},$$

donde $y_1 = 0, 1, 2, \ldots$ Y $y_2 = 0, 1, 2, \ldots,$ y_1. Note que como $x_1 > 0$, la transformación $x_1 = y_1 - x_2$ implica que y_2 y, por lo tanto, x_2 siempre deben ser menores que o iguales a y_1. En consecuencia, la distribución de probabilidad marginal de Y_1 es

$$\begin{aligned} h(y_1) &= \sum_{y_2=0}^{y_1} g(y_1, y_2) = e^{-(\mu_1+\mu_2)} \sum_{y_2=0}^{y_1} \frac{\mu_1^{y_1-y_2}\mu_2^{y_2}}{(y_1 - y_2)!y_2!} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \frac{y_1!}{y_2!(y_1 - y_2)!}\mu_1^{y_1-y_2}\mu_2^{y_2} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \binom{y_1}{y_2}\mu_1^{y_1-y_2}\mu_2^{y_2}. \end{aligned}$$

Al reconocer esta suma como la expansión binomial de $(\mu_1 + \mu_2)^{y_1}$, obtenemos

$$h(y_1) = \frac{e^{-(\mu_1+\mu_2)}(\mu_1 + \mu_2)^{y_1}}{y_1!}, \qquad y_1 = 0, 1, 2, \ldots,$$

de lo cual concluimos que la suma de las dos variables aleatorias independientes que tienen distribuciones de Poisson, con parámetros μ_1 y μ_2, tiene una distribución de Poisson con parámetro $\mu_1 + \mu_2$.

Para encontrar la distribución de probabilidad de la variable aleatoria $Y = u(X)$ cuando X es una variable aleatoria continua y la transformación es uno a uno, necesitaremos el teorema 7.3.

Teorema 7.3: Suponga que X es una variable aleatoria **continua** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una correspondencia uno a uno entre los valores de X y Y, de manera que la ecuación $y = u(x)$ se resuelva unívocamente para x en términos de y, digamos $x = w(y)$. Entonces, la distribución de probabilidad de Y es

$$g(y) = f[w(y)]|J|,$$

donde $J = w'(y)$ y se llama **jacobiano** de la transformación.

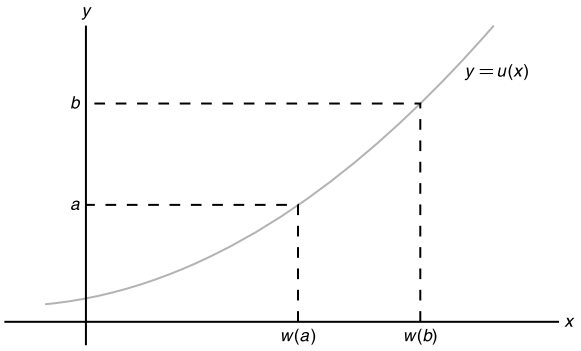


Figura 7.1: Función creciente.

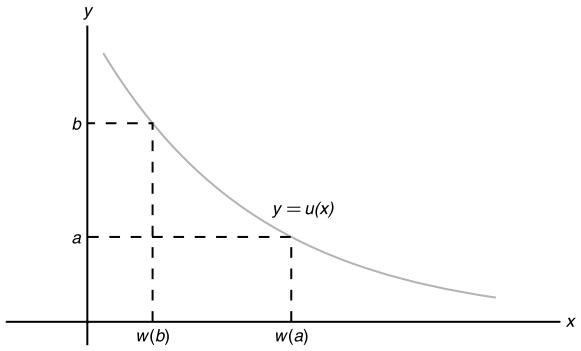


Figura 7.2: Función decreciente.

Prueba: Suponga que $y = u(x)$ es una función creciente como la de la figura 7.1. Entonces, vemos que siempre que Y cae entre a y b, la variable aleatoria X debe caer entre $w(a)$ y $w(b)$. De aquí,

$$P(a < Y < b) = P[w(a) < X < w(b)] = \int_{w(a)}^{w(b)} f(x)\, dx.$$

Al cambiar la variable de integración de x a y mediante la relación $x = w\,(y)$, obtenemos $dx = w'(y)dy$ y, por lo tanto,

$$P(a < Y < b) = \int_a^b f[w(y)]w'(y)\, dy.$$

Como la integral da la probabilidad que se desea para toda $a < b$ dentro del conjunto permisible de valores y, entonces la distribución de probabilidad de Y es

$$g(y) = f[w(y)]w'(y) = f[w(y)]J.$$

Si reconocemos $J = w'(y)$ como el recíproco de la pendiente de la línea tangente a la curva de la función creciente $y = u(x)$, entonces es evidente que $J = |J|$. De aquí,

$$g(y) = f[w(y)]|J|.$$

Suponga que $y = u(x)$ es una función decreciente como la de la figura 7.2. Entonces, escribimos

$$P(a < Y < b) = P[w(b) < X < w(a)] = \int_{w(b)}^{w(a)} f(x)\ dx.$$

De nuevo, al cambiar la variable de integración por y, obtenemos

$$P(a < Y < b) = \int_b^a f[w(y)]w'(y)\ dy = -\int_a^b f[w(y)]w'(y)\ dy,$$

de lo cual concluimos que

$$g(y) = -f[w(y)]w'(y) = -f[w(y)]J.$$

En este caso, la pendiente de la curva es negativa y $J = -|J|$. Entonces,

$$g(y) = f[w(y)]|J|,$$

como antes.

Ejemplo 7.3: Sea X una variable aleatoria continua con distribución de probabilidad

$$f(x) = \begin{cases} \frac{x}{12}, & 1 < x < 5, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la distribución de probabilidad de la variable aleatoria $Y = 2X - 3$.

Solución: La solución inversa de $y = 2x - 3$ da $x = (y + 3)/2$, de la que obtenemos $J = w'(y) = dx/dy = 1/2$. Por lo tanto, usando el teorema 7.3, encontramos que la función de densidad de Y es

$$g(y) = \begin{cases} \frac{(y+3)/2}{12}\left(\frac{1}{2}\right) = \frac{y+3}{48}, & -1 < y < 7, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Para encontrar la distribución de probabilidad conjunta de las variables aleatorias $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$, cuando X_1 y X_2 son continuas, y la transformación es uno a uno, necesitamos un teorema adicional, análogo al teorema 7.2, que establecemos sin demostración.

Teorema 7.4: Suponga que X_1 y X_2 son variables aleatorias **continuas** con distribución de probabilidad conjunta $f(x_1, x_2)$. Definamos con $Y_1 = u_1(X_1, X_2)$ y $Y_2 = u_2(X_1, X_2)$ una transformación uno a uno entre los puntos (x_1, x_2) y (y_1, y_2), de manera que las ecuaciones $y_1 = u_1(x_1, x_2)$ y $y_2 = u_2(x_1, x_2)$ se resuelve unívocamente para x_1 y x_2 en términos de y_1 y y_2, digamos $x_1 = w_1(y_1, y_2)$ y $x_2 = w_2(y_1, y_2)$. Entonces, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]|J|,$$

donde el jacobiano es el determinante 2×2

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

y $\frac{\partial x_1}{\partial y_1}$ es simplemente la derivada de $x_1 = w_1(y_1, y_2)$ con respecto a y_1, y y_2 permanece constante, que en cálculo se denomina derivada parcial de x_1 con respecto a y_1. Las otras derivadas parciales se definen de manera similar.

Ejemplo 7.4: Sean X_1, y X_2 dos variables aleatorias continuas con distribución de probabilidad conjunta

$$f(x_1, x_2) = \begin{cases} 4x_1x_2, & 0 < x_1 < 1,\ 0 < x_2 < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la distribución de probabilidad conjunta de $Y_1 = X_1^2$ y $Y_2 = X_1X_2$.

Solución: Las soluciones inversas de $y_1 = x_1^2$ y $y_2 = x_1x_2$ son $x_1 = \sqrt{y_1}$ y $x_2 = y_2/\sqrt{y_1}$, de las que obtenemos

$$J = \begin{vmatrix} 1/(2\sqrt{y_1}) & 0 \\ -y_2/2y_1^{3/2} & 1/\sqrt{y_1} \end{vmatrix} = \frac{1}{2y_1}.$$

Para determinar el conjunto B de puntos en el plano y_1y_2 en el que se traza (mapea) el conjunto A de puntos en el plano x_1x_2, escribimos

$$x_1 = \sqrt{y_1} \qquad \text{y} \qquad x_2 = y_2/\sqrt{y_1}$$

y después al hacer $x_1 = 0$, $x_2 = 0$, $x_1 = 1$ y $x_2 = 1$, las fronteras del conjunto A se transforman a $y_1 = 0$, $y_2 = 0$, $y_1 = 1$ y $y_2 = \sqrt{y_1}$ o $y_2^2 = y_1$. Las dos regiones se ilustran en la figura 7.3. Claramente, la transformación es uno a uno, al trazar el conjunto $A = \{(x_1, x_2) \mid 0 < x_1 < 1,\ 0 < x_2 < 1\}$ en el conjunto $B = \{(y_1, y_2) \mid y_2^2 < y_1 < 1,\ 0 < y_2 < 1\}$. Del teorema 7.4, la distribución de probabilidad conjunta de Y_1 y Y_2 es

$$g(y_1, y_2) = 4(\sqrt{y_1})\frac{y_2}{\sqrt{y_1}}\frac{1}{2y_1} = \begin{cases} \frac{2y_2}{y_1}, & y_2^2 < y_1 < 1,\ 0 < y_2 < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

A menudo surgen problemas cuando deseamos encontrar la distribución de probabilidad de la variable aleatoria $Y = u(X)$ cuando X es una variable aleatoria con-

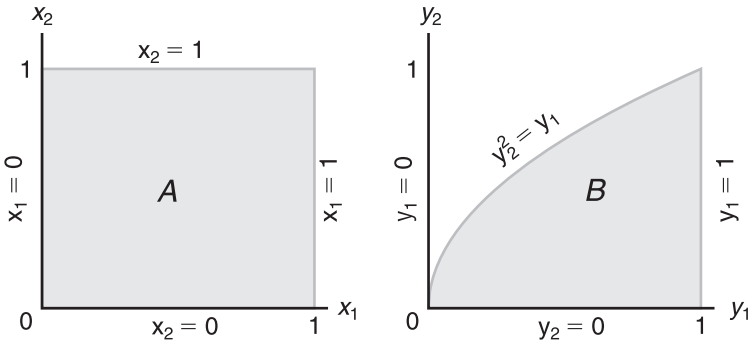


Figura 7.3: Trazo del conjunto A en el conjunto B.

tinua y la transformación no es uno a uno. Es decir, para cada valor x corresponde exactamente un valor y; pero a cada valor y corresponde más de un valor x. Por ejemplo, suponga que $f(x)$ es positiva en el intervalo $-1 < x < 2$ y cero en cualquier otro caso. Considere la transformación $y = x^2$. En este caso $x = \pm\sqrt{y}$ para $0 < y < 1$ y $x = \sqrt{y}$ para $1 < y < 4$. Para el intervalo $1 < y < 4$, la distribución de probabilidad de Y se encuentra como antes, con el teorema 7.3. Es decir,

$$g(y) = f[w(y)]|J| = \frac{f(\sqrt{y})}{2\sqrt{y}}, \qquad 1 < y < 4.$$

Sin embargo, cuando $0 < y < 1$, podemos dividir el intervalo $-1 < x < 1$ para obtener las dos funciones inversas

$$x = -\sqrt{y}, \qquad -1 < x < 0, \qquad \text{y} \qquad x = \sqrt{y}, \qquad 0 < x < 1.$$

Entonces para todo valor y corresponde un solo valor x para cada partición. De la figura 7.4 vemos que

$$\begin{aligned} P(a < Y < b) &= P(-\sqrt{b} < X < -\sqrt{a}) + P(\sqrt{a} < X < \sqrt{b}) \\ &= \int_{-\sqrt{b}}^{-\sqrt{a}} f(x)\, dx + \int_{\sqrt{a}}^{\sqrt{b}} f(x)\, dx. \end{aligned}$$

Al cambiar la variable de integración de x a y, obtenemos

$$\begin{aligned} P(a < Y < b) &= \int_b^a f(-\sqrt{y})J_1\, dy + \int_a^b f(\sqrt{y})J_2\, dy \\ &= -\int_a^b f(-\sqrt{y})J_1\, dy + \int_a^b f(\sqrt{y})J_2\, dy, \end{aligned}$$

donde

$$J_1 = \frac{d(-\sqrt{y})}{dy} = \frac{-1}{2\sqrt{y}} = -|J_1|,$$

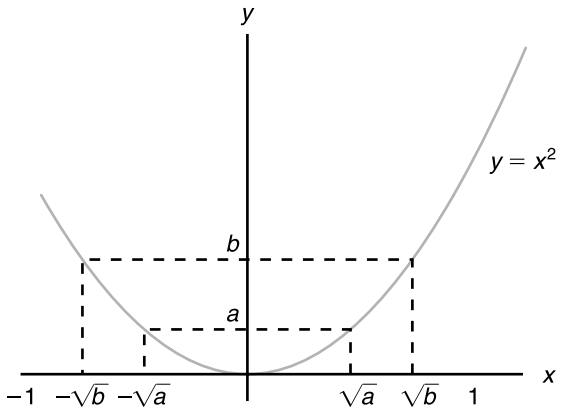


Figura 7.4: Función decreciente y creciente.

y

$$J_2 = \frac{d(\sqrt{y})}{dy} = \frac{1}{2\sqrt{y}} = |J_2|.$$

De aquí, podemos escribir

$$P(a < Y < b) = \int_a^b [f(-\sqrt{y})|J_1| + f(\sqrt{y})|J_2|]\, dy,$$

y entonces

$$g(y) = f(-\sqrt{y})|J_1| + f(\sqrt{y})|J_2| = \frac{f(-\sqrt{y}) + f(\sqrt{y})}{2\sqrt{y}}, \qquad 0 < y < 1.$$

La distribución de probabilidad de Y para $0 < y < 4$ se puede escribir ahora como

$$g(y) = \begin{cases} \frac{f(-\sqrt{y})+f(\sqrt{y})}{2\sqrt{y}}, & 0 < y < 1, \\ \frac{f(\sqrt{y})}{2\sqrt{y}}, & 1 < y < 4, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Este procedimiento para encontrar $g(y)$ cuando $0 < y < 1$ se generaliza en el teorema 7.5 para k funciones inversas. Para transformaciones que no son uno a uno de funciones de varias variables, se recomienda al lector *Introduction to Mathematical Statistics* de Hogg y Craig (véase la bibliografía).

Teorema 7.5: Suponga que X es una variable aleatoria **continua** con distribución de probabilidad $f(x)$. Definamos con $Y = u(X)$ una transformación entre los valores de X y Y que no es uno a uno. Si el intervalo sobre el que se define X se puede dividir en k conjuntos mutuamente disjuntos, de manera que cada una de las funciones inversas

$$x_1 = w_1(y), \quad x_2 = w_2(y), \quad \ldots, \quad x_k = w_k(y)$$

de $y = u(x)$ defina una correspondencia uno a uno, entonces la distribución de probabilidad de Y es

$$g(y) = \sum_{i=1}^{k} f[w_i(y)]|J_i|,$$

donde $J_i = w_i'(y)$, $i = 1, 2, \ldots, k$.

Ejemplo 7.5: Muestre que $Y = (X - \mu)^2/\sigma^2$ tiene una distribución chi cuadrada con 1 grado de libertad cuando X tiene una distribución normal con media μ y varianza σ^2.

Solución: Sea $Z = (X - \mu)/\sigma$, donde la variable aleatoria Z tiene la distribución normal estándar

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad -\infty < z < \infty.$$

Encontraremos ahora la distribución de la variable aleatoria $Y = Z^2$. Las soluciones inversas de $y = z^2$ son $z = \pm\sqrt{y}$. Si designamos $z_1 = -\sqrt{y}$ y $z_2 = \sqrt{y}$, entonces $J_1 = -1/2\sqrt{y}$ y $J_2 = 1/2\sqrt{y}$. De aquí, por el teorema 7.5, tenemos

$$g(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{-1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{\sqrt{2\pi}} y^{1/2-1} e^{-y/2}, \qquad y > 0.$$

Como $g(y)$ es una función de densidad, se sigue que

$$1 = \frac{1}{\sqrt{2\pi}} \int_0^\infty y^{1/2-1} e^{-y/2}\, dy = \frac{\Gamma(1/2)}{\sqrt{\pi}} \int_0^\infty \frac{y^{1/2-1} e^{-y/2}}{\sqrt{2}\Gamma(1/2)}\, dy = \frac{\Gamma(1/2)}{\sqrt{\pi}},$$

la integral es el área bajo una curva de probabilidad gamma con parámetros $\alpha = 1/2$ y $\beta = 2$. Por lo tanto, $\sqrt{\pi} = \Gamma(1/2)$ y la distribución de probabilidad de Y está dada por

$$g(y) = \begin{cases} \frac{1}{\sqrt{2}\Gamma(1/2)} y^{1/2-1} e^{-y/2}, & y > 0, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

que se considera una distribución chi cuadrada con 1 grado de libertad.

7.3 Momentos y funciones generadoras de momentos

En esta sección nos concentramos en aplicaciones de las funciones generadoras de momentos. El propósito evidente de la función generadora de momentos es la deter-

minación de los momentos de variables aleatorias. Sin embargo, la contribución más importante consiste en establecer distribuciones de funciones de variables aleatorias.

Si $g(X) = X^r$ para $r = 0, 1, 2, 3, \ldots$, la definición 7.1 da un valor esperado que se denomina r-ésimo **momento alrededor del origen** de la variable aleatoria X, que denotamos como μ_r'.

Definición 7.1: El r-ésimo **momento alrededor del origen** de la variable aleatoria X está dado por

$$\mu_r' = E(X^r) = \begin{cases} \sum_x x^r f(x), & \text{si } X \text{ es discreta,} \\ \int_{-\infty}^{\infty} x^r f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Como el primer y segundo momentos alrededor del origen están dados por $\mu_1' = E(X)$ y $\mu_2' = E(X^2)$, podemos escribir la media y la varianza de una variable aleatoria como

$$\mu = \mu_1' \qquad \text{y} \qquad \sigma^2 = \mu_2' - \mu^2.$$

Aunque los momentos de una variable aleatoria se pueden determinar directamente de la definición 7.1, existe un procedimiento alternativo, el cual requiere que utilicemos una **función generadora de momentos**.

Definición 7.2: La **función generadora de momentos** de la variable aleatoria X está dada por $E(e^{tX})$ y se denota con $M_X(t)$. De aquí,

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} f(x), & \text{si } X \text{ es discreta,} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Las funciones generadoras de momentos existirán sólo si la suma o integral de la definición 7.2 converge. Si existe una función generadora de momentos de una variable aleatoria X, se puede utilizar para generar todos los momentos de dicha variable. El método se describe en el teorema 7.6.

Teorema 7.6: Sea X una variable aleatoria con función generadora de momentos $M_X(t)$. Entonces,

$$\left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0} = \mu_r'.$$

Prueba: Suponiendo que podemos diferenciar dentro de la sumatoria y los signos de la integral, obtenemos

$$\frac{d^r M_X(t)}{dt^r} = \begin{cases} \sum_x x^r e^{tx} f(x), & \text{si } X \text{ es discreta,} \\ \int_{-\infty}^{\infty} x^r e^{tx} f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Al hacer $t = 0$, vemos que ambos casos se reducen a $E(X^r) = \mu_r'$. ∎

Ejemplo 7.6.: Encuentre la función generadora de momentos de la variable aleatoria binomial X y después utilícela para verificar que $\mu = np$ y $\sigma^2 = npq$.

Solución: De la definición 7.2, tenemos

$$M_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x}.$$

Si reconocemos esta última suma como la expansión binomial de $(pe^t + q)^n$, obtenemos

$$M_X(t) = (pe^t + q)^n.$$

Así,

$$\frac{dM_X(t)}{dt} = n(pe^t + q)^{n-1} pe^t$$

y

$$\frac{d^2 M_X(t)}{dt^2} = np[e^t(n-1)(pe^t + q)^{n-2} pe^t + (pe^t + q)^{n-1} e^t].$$

Al hacer $t = 0$, obtenemos

$$\mu_1^{'} = np$$

y

$$\mu_2^{'} = np[(n-1)p + 1].$$

Por lo tanto,

$$\mu = \mu_1^{'} = np$$

y

$$\sigma^2 = \mu_2^{'} - \mu^2 = np(1-p) = npq,$$

que está de acuerdo con los resultados que se obtuvieron en el capítulo 5.

Ejemplo 7.7: Muestre que la función generadora de momentos de la variable aleatoria X que tiene una distribución de probabilidad normal con media μ y varianza σ^2 está dada por

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

Solución: De la definición 7.2 la función generadora de momentos de la variable aleatoria normal X es

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x^2 - 2(\mu + t\sigma^2)x + \mu^2}{2\sigma^2}\right] dx. \end{aligned}$$

Al completar el cuadrado en el exponente, escribimos

$$x^2 - 2(\mu + t\sigma^2)x + \mu^2 = [x - (\mu + t\sigma^2)]^2 - 2\mu t\sigma^2 - t^2\sigma^4$$

y, entonces,

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[x - (\mu + t\sigma^2)]^2 - 2\mu t\sigma^2 - t^2\sigma^4}{2\sigma^2}\right\} dx \\ &= \exp\left(\frac{2\mu t + \sigma^2 t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[x - (\mu + t\sigma^2)]^2}{2\sigma^2}\right\} dx. \end{aligned}$$

Sea $w = [x - (\mu + t\sigma^2)]/\sigma$; entonces $dx = \sigma\ dw$ y

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\ dw = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right),$$

ya que la última integral representa el área bajo una curva de densidad normal estándar y, por ello, es igual a 1.

Ejemplo 7.8: Muestre que la función generadora de momentos de la variable aleatoria X que tiene una distribución chi cuadrada con v grados de libertad es $M_X(t) = (1 - 2t)^{-v/2}$.

Solución: La distribución chi cuadrada se obtuvo como un caso especial de la distribución gamma al hacer $\alpha = v/2$ y $\beta = 2$. Al sustituir por $f(x)$ en la definición 7.2, obtenemos

$$\begin{aligned} M_X(t) &= \int_0^{\infty} e^{tx} \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}\ dx \\ &= \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^{\infty} x^{v/2-1} e^{-x(1-2t)/2}\ dx. \end{aligned}$$

Al escribir $y = x(1 - 2t)/2$ y $dx = [2/(1 - 2t)]\ dy$, obtenemos para $t < \frac{1}{2}$,

$$\begin{aligned} M_X(t) &= \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^{\infty} \left(\frac{2y}{1-2t}\right)^{v/2-1} e^{-y} \frac{2}{1-2t}\ dy \\ &= \frac{1}{\Gamma(v/2)(1-2t)^{v/2}} \int_0^{\infty} y^{v/2-1} e^{-y}\ dy = (1 - 2t)^{-v/2}, \end{aligned}$$

ya que la última integral es igual a $\Gamma(v/2)$.

Aunque el método de transformación de variables brinda una forma eficaz para encontrar la distribución de una función de diversas variables, hay un procedimiento alternativo y que a menudo se prefiere, cuando la función en cuestión es una combinación lineal de variables aleatorias independientes. Este procedimiento utiliza las propiedades de las funciones generadoras de momentos que se incluyen en los siguientes cuatro teoremas. Para conservar el alcance matemático de este libro, establecemos el teorema 7.7 sin demostración.

Teorema 7.7: (**Teorema de unicidad**) Sean X y Y dos variables aleatorias con funciones generadoras de momentos $M_X(t)$ y $M_Y(t)$, respectivamente. Si $M_X(t) = M_Y(t)$ para todos los valores de t, entonces X y Y tienen la misma distribución de probabilidad.

Teorema 7.8: $M_{X+a}(t) = e^{at}M_X(t)$.

Prueba: $M_{X+a}(t) = E[e^{t(X+a)}] = e^{at}E(e^{tX}) = e^{at}M_X(t)$.

Teorema 7.9: $M_{aX}(t) = M_X(at)$.

Prueba: $M_{aX}(t) = E[e^{t(aX)}] = E[e^{(at)X}] = M_X(at)$.

Teorema 7.10: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes con funciones generadoras de momentos $M_{X_1}(t), M_{X_2}(t), \ldots, M_{X_n}(t)$, respectivamente, y $Y = X_1 + X_2 + \cdots + X_n$, entonces,

$$M_Y(t) = M_{X_1}(t)M_{X_2}(t) \cdots M_{X_n}(t).$$

Prueba: Para el caso continuo

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E[e^{t(X_1+X_2+\cdots+X_n)}] \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t(X_1+X_2+\cdots+X_n)} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n. \end{aligned}$$

Como las variables son independientes, tenemos

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n)$$

y entonces

$$\begin{aligned} M_Y(t) &= \int_{-\infty}^{\infty} e^{tx_1} f_1(x_1)\, dx_1 \int_{-\infty}^{\infty} e^{tx_2} f_2(x_2)\, dx_2 \cdots \int_{-\infty}^{\infty} e^{tx_n} f_n(x_n)\, dx_n \\ &= M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t). \end{aligned}$$

La demostración para el caso discreto se obtiene de manera similar, reemplazando las integrales con sumatorias.

Los teoremas 7.7 a 7.10 son fundamentales para entender las funciones generadoras de momentos. A continuación se presenta un ejemplo como ilustración. Hay muchas situaciones en que necesitamos conocer la distribución de la suma de las variables aleatorias. Podemos utilizar los teoremas 7.7 y 7.10, así como el resultado del ejercicio 7.19 que sigue de esta sección, para encontrar la distribución de una suma de dos variables aleatorias independientes de Poisson, con funciones generadoras de momentos dadas por

$$M_{X_1}(t) = e^{\mu_1(e^t-1)}, \qquad \text{y} \qquad M_{X_2}(t) = e^{\mu_2(e^t-1)},$$

respectivamente. De acuerdo con el teorema 7.10, la función generadora de momentos de la variable aleatoria $Y_1 = X_1 + X_2$ es

$$M_{Y_1}(t) = M_{X_1}(t)M_{X_2}(t) = e^{\mu_1(e^t-1)}e^{\mu_2(e^t-1)} = e^{(\mu_1+\mu_2)(e^t-1)}$$

que de inmediato identificamos como la función generadora de momentos de una variable aleatoria que tiene una distribución de Poisson con el parámetro $\mu_1 + \mu_2$. Por ello, de acuerdo con el teorema 7.7, de nuevo concluimos que la suma de dos variables aleatorias independientes que tienen distribuciones de Poisson, con parámetros μ_1 y μ_2, tiene una distribución de Poisson con parámetro $\mu_1 + \mu_2$.

Combinaciones lineales de variables aleatorias

En estadística aplicada a menudo se necesita conocer la distribución de probabilidad de una combinación lineal de variables aleatorias normales independientes. Obtengamos la distribución de la variable aleatoria $Y = a_1X_1 + a_2\ X_2$ cuando X_1 es una variable normal con media μ_1 y varianza σ_1^2 y X_2 también es una variable normal, pero independiente de X_1, con media μ_2 y varinza σ_2^2. Primero, por el teorema 7.10, encontramos

$$M_Y(t) = M_{a_1X_1}(t)M_{a_2X_2}(t),$$

y después, usando el teorema 7.9,

$$M_Y(t) = M_{X_1}(a_1t)M_{X_2}(a_2t).$$

Al sustituir a_1t por t, y después a_2t por t, en una función generadora de momentos de la distribución normal que se deriva en el ejemplo 7.7, tenemos

$$\begin{aligned} M_Y(t) &= \exp(a_1\mu_1t + a_1^2\sigma_1^2t^2/2 + a_2\mu_2t + a_2^2\sigma_2^2t^2/2) \\ &= \exp[(a_1\mu_1 + a_2\mu_2)t + (a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2/2], \end{aligned}$$

que reconocemos como la función generadora de momentos de una distribución que es normal con media $a_1\mu_1 + a_2\mu_2$ y varianza $a_1^2\sigma_1^2 + a_2^2\sigma_2^2$.

Para generalizar al caso de n variables normales independientes, establecemos el siguiente resultado.

Teorema 7.11: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes que tienen distribuciones normales con medias $\mu_1, \mu_2, \ldots, \mu_n$ y varianzas $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectivamente, entonces la variable aleatoria

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

tiene una distribución normal con media

$$u_Y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

y varianza

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2.$$

Ahora queda claro que la distribución de Poisson y la distribución normal tienen una propiedad reproductiva, en el sentido de que la suma de variables aleatorias independientes que tengan cualquiera de estas distribuciones es una variable alea-

toria que también tiene el mismo tipo de distribución. La distribución chi cuadrada también posee esta propiedad reproductiva.

Teorema 7.12: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias mutuamente independientes que tienen, respectivamente, distribuciones chi cuadrada con $v_1, v_2, \ldots, v_n$ grados de libertad, entonces la variable aleatoria

$$Y = X_1 + X_2 + \cdots + X_n$$

tiene una distribución chi cuadrada con $v = v_1 + v_2 + \cdots v_n$ grados de libertad.

Prueba: Por el teorema 7.10,

$$M_Y(t) = M_{X_1}(t)\ M_{X_2}(t) \cdots M_{X_n}(t).$$

Del ejemplo 7.8,

$$M_{X_i}(t) = (1-2t)^{-v_i/2}, \qquad i = 1, 2, \ldots, n.]$$

Por lo tanto,

$$\begin{aligned} M_Y(t) &= (1-2t)^{-v_1/2}(1-2t)^{-v_2/2} \cdots (1-2t)^{-v_n/2} \\ &= (1-2t)^{-(v_1+v_2+\cdots+v_n)/2}, \end{aligned}$$

que reconocemos como la función generadora de momentos de una distribución chi cuadrada con $v = v_1 + v_2 + \cdots + v_n$ grados de libertad.

Corolario 7.1: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes que tienen distribuciones normales idénticas con media μ y varianza σ^2, entonces la variable aleatoria

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

tiene una distribución chi cuadrada con $v = n$ grados de libertad.

Este corolario es una consecuencia inmediata del ejemplo 7.5, que establece que cada una de las n variables aleatorias independientes $(X_i - \mu)/\sigma]^2$, $i = 1, 2, \ldots, n$, tiene una distribución chi cuadrada con 1 grado de libertad. Este corolario es muy relevante. Establece una relación entre la muy importante distribución chi cuadrada y la distribución normal. También debe proporcionar al lector una idea clara de lo que queremos decir con el parámetro que denominamos grados de libertad. Conforme avancemos en los capítulos siguientes, la noción de grados de libertad jugará un papel de importancia creciente. Del corolario 7.1 vemos que si $Z_1, Z_2, \ldots, Z_n$ son variables aleatorias normales estándar independientes, entonces $\sum_{i=1}^{n} Z_i^2$ tiene una distribución chi cuadrada y el *parámetro único*, v, los grados de libertad, es n, el *número* de variables aleatorias normales estándar. Además, si cada variable aleatoria normal en X_i, en el corolario 7.1, tiene diferentes media y varianza podemos tener el siguiente resultado.

Corolario 7.2: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes y X_i sigue una distribución normal con media μ_i y varianza σ_i^2 para $i = 1, 2, \ldots, n$, entonces, la variable aleatoria

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

tiene una distribución chi cuadrada con $v = n$ grados de libertad.

Ejercicios

7.1 Sea X una variable aleatoria con probabilidad

$$f(x) = \begin{cases} \frac{1}{3}, & x = 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Encuentre la distribución de probabilidad de la variable aleatoria $Y = 2X - 1$.

7.2 Sea X una variable aleatoria binomial con distribución de probabilidad

$$f(x) = \begin{cases} \binom{3}{x} \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{3-x}, & x = 0, 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Encuentre la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.3 Sean X_1 y X_2 variables aleatorias discretas con la distribución multinomial conjunta

$$f(x_1, x_2) = \binom{2}{x_1, x_2, 2 - x_1 - x_2} \left(\frac{1}{4}\right)^{x_1} \left(\frac{1}{3}\right)^{x_2} \left(\frac{5}{12}\right)^{2-x_1-x_2}$$

para $x_1 = 0, 1, 2$; $x_2 = 0, 1, 2$; $x_1 + x_2 \leq 2$; y cero en cualquier otro caso. Encuentre la distribución de probabilidad conjunta de $Y_1 = X_1 + X_2$ y $Y_2 = X_1 - X_2$.

7.4 Sean X_1 y X_2 variables aleatorias discretas con la distribución de probabilidad conjunta

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{18}, & x_1 = 1, 2; \ x_2 = 1, 2, 3, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Encuentre la distribución de probabilidad de la variable aleatoria $Y = X_1 X_2$.

7.5 Si X tiene la distribución de probabilidad

$$f(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

Muestre que la variable aleatoria $Y = -2\ln X$ tiene una distribución chi cuadrada con 2 grados de libertad.

7.6 Dada la variable aleatoria X con distribución de probabilidad

$$f(x) = \begin{cases} 2x, & 0 < x < 1, \\ 0, & \text{en cualquier caso,} \end{cases}$$

encuentre la distribución de probabilidad de $Y = 8X^3$.

7.7 La velocidad de una molécula en un gas uniforme en equilibrio es una variable aleatoria V, cuya distribución de probabilidad está dada por

$$f(v) = \begin{cases} kv^2 e^{-bv^2}, & v > 0, \\ 0, & \text{en cualquier caso,} \end{cases}$$

donde k es una constante adecuada y b depende de la temperatura absoluta y de la masa de la molécula. Encuentre la distribución de probabilidad de la energía cinética de la molécula W, donde $W = mV^2/2$.

7.8 La utilidad de un distribuidor, en unidades de \$5000, sobre un automóvil nuevo está dada por $Y = X^2$, donde X es una variable aleatoria que tiene la función de densidad

$$f(x) = \begin{cases} 2(1 - x), & 0 < x < 1, \\ 0, & \text{en cualquier caso.} \end{cases}$$

a) Encuentre la función de densidad de probabilidad de la variable aleatoria Y.

b) Usando la función de densidad de Y, encuentre la probabilidad de que la utilidad sea menor que \$500 sobre el siguiente automóvil nuevo que venda este distribuidor.

7.9 El periodo hospitalario, en días, para pacientes que siguen un tratamiento para cierto tipo de enfermedad del riñón es una variable aleatoria $Y = X + 4$, donde X tiene la función de densidad

$$f(x) = \begin{cases} \frac{32}{(x+4)^3}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre la función de densidad de probabilidad de la variable aleatoria Y.

b) Con la función de densidad de Y, encuentre la probabilidad de que el periodo hospitalario para un paciente que sigue este tratamiento exceda los ocho días.

7.10 Las variables aleatorias X y Y, que representan los pesos de cremas y chiclosos en cajas de un kilogramo de chocolates, que contienen una combinación de cremas, chiclosos y envinados, tienen la función de densidad conjunta

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1,\ 0 \le y \le 1, \\ & x + y \le 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

a) Encuentre la función de densidad de probabilidad de la variable aleatoria $Z = X + Y$.

b) Usando la función de densidad de Z, encuentre la probabilidad de que en una caja dada la suma de las cremas y los chiclosos sea al menos $1/2$, pero menos que $3/4$ del peso total.

7.11 La cantidad de queroseno, en miles de litros, en un tanque al inicio de cualquier día es una cantidad aleatoria Y, de la cual una cantidad aleatoria X se vende durante ese día. Suponga que la función de densidad conjunta de estas variables está dada por

$$f(x,y) = \begin{cases} 2, & 0 < x < y,\ 0 < y < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la función de densidad de probabilidad para la cantidad de queroseno que queda en el tanque al final del día.

7.12 Sean X_1 y X_2 variables aleatorias independientes que tienen cada una la distribución de probabilidad

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Muestre que las variables aleatorias Y_1 y Y_2 son independientes cuando $Y_1 = X_1 + X_2$ y $Y_2 = X_1/(X_1 + X_2)$.

7.13 Una corriente de I amperes que fluye a través de una resistencia de R ohms varía de acuerdo con la distribución de probabilidad

$$f(i) = \begin{cases} 6i(1-i), & 0 < i < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Si la resistencia varía independientemente de la corriente de acuerdo con la distribución de probabilidad

$$g(r) = \begin{cases} 2r, & 0 < r < 1, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

encuentre la distribución de probabilidad para la potencia $W = I^2R$ watts.

7.14 Sea X una variable aleatoria con distribución de probabilidad

$$f(x) = \begin{cases} \frac{1+x}{2}, & -1 < x < 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.15 Si X tiene la distribución de probabilidad

$$f(x) = \begin{cases} \frac{2(x+1)}{9}, & -1 < x < 2, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Encuentre la distribución de probabilidad de la variable aleatoria $Y = X^2$.

7.16 Muestre que el r-ésimo momento alrededor del origen de la distribución gamma es

$$\mu_r' = \frac{\beta^r \Gamma(\alpha + r)}{\Gamma(\alpha)}.$$

[*Sugerencia:* Sustituya $y = x/\beta$ en la integral que define μ_r' y después utilice la función gamma para evaluar la integral.]

7.17 Una variable aleatoria X tiene la distribución uniforme discreta

$$f(x;k) = \begin{cases} \frac{1}{k}, & x = 1, 2, \ldots, k, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Muestre que la función generadora de momentos de X es

$$M_X(t) = \frac{e^t(1 - e^{kt})}{k(1 - e^t)}.$$

7.18 Una variable aleatoria X tiene la distribución geométrica $g(x;\, p) = pq^{x-1}$ para $x = 1, 2, 3, \ldots$. Muestre que la función generadora de momentos de X es

$$M_X(t) = \frac{pe^t}{1 - qe^t}, \qquad t < \ln q,$$

y después use $M_X(t)$ para encontrar la media y la varianza de la distribución geométrica.

7.19 Una variable aleatoria X tiene la distribución de Poisson $p(x; \mu) = e^{-\mu}\mu^x/x!$ para $x = 0, 1, 2, \ldots$ Muestre que la función generadora de momentos de X es

$$M_X(t) = e^{\mu(e^t-1)}.$$

Usando $M_X(t)$, encuentre la media y la varianza de la distribución de Poisson.

7.20 La función generadora de momentos de cierta variable aleatoria de Poisson X está dada por

$$M_X(t) = e^{\mu(e^t-1)}.$$

Encuentre $P(\mu - 2\sigma < X < \mu + 2\sigma)$.

7.21 Con la función generadora de momentos del ejemplo 7.8, muestre que la media y la varianza de la distribución chi cuadrada con v grados de libertad son, respectivamente, v y $2v$.

7.22 Mediante la expansión de e^{tx} en una serie de Maclaurin y la integración término por término, muestre que

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\ &= 1 + \mu t + \mu_2' \frac{t^2}{2!} + \cdots + \mu_r' \frac{t^r}{r!} + \cdots . \end{aligned}$$

7.23 Si tanto X como Y, distribuidas de manera independiente, siguen distribuciones exponenciales con parámetro medio 1, encuentre las distribuciones de

a) $U = X + Y$, y

b) $V + X/(X + Y)$.

Capítulo 8

Distribuciones de muestreo fundamentales y descripciones de datos

8.1 Muestreo aleatorio

El resultado de un experimento estadístico se puede registrar como un valor numérico o como una representación descriptiva. Cuando se lanza un par de dados y el total es el resultado de interés, registramos un valor numérico. No obstante, si a los estudiantes de cierta escuela se les hacen pruebas de sangre y el tipo sanguíneo es de interés, entonces una representación descriptiva podría ser la más útil. La sangre de un individuo se puede clasificar de 8 maneras. Puede ser AB, A, B u O, con un signo más o uno menos, lo cual depende de la presencia o ausencia del antígeno Rh.

En este capítulo nos enfocamos en el muestreo de distribuciones o poblaciones, y estudiamos cantidades tan importantes como la *media de la muestra* y la *varianza de la muestra*, que son de importancia fundamental para los capítulos siguientes. Además, intentamos dar al lector una introducción al papel que jugarán la media y la varianza de la muestra en los próximos capítulos sobre inferencia estadística. El uso de las computadoras modernas de alta velocidad permite al científico o al ingeniero aumentar enormemente su uso de la inferencia estadística formal con técnicas gráficas. La mayoría de las veces la inferencia formal parece bastante árida, y quizás incluso abstracta para el profesional o el administrador que deseen que el análisis estadístico sea una guía para la toma de decisiones.

Población y muestras

Comenzamos esta sección con la presentación de las nociones de *poblaciones* y *muestras*. Ambas se mencionan de forma extensa en el capítulo 1. Sin embargo, será necesario estudiarlas más ampliamente aquí, en particular en el contexto del concepto de variables aleatorias. La totalidad de observaciones que nos interesan, de número finito o infinito, constituye lo que llamamos **población**. En el pasado el término *población* se refería a observaciones que se obtenían de estudios estadísticos con personas. En la actualidad, el estadístico utiliza la palabra para referirse a observaciones respecto de cualquier cuestión de interés, ya sea de grupos de personas, animales o todos los resultados posibles de algún sistema biológico o de ingeniería complicados.

Definición 8.1: Una **población** consiste en la totalidad de las observaciones en las que estamos interesados.

El número de observaciones en la población se define como el tamaño de la población. Si en la escuela hay 600 estudiantes que clasificamos de acuerdo con su tipo sanguíneo, decimos que tenemos una población de tamaño 600. Los números en las cartas de una baraja, las estaturas de los residentes de cierta ciudad y las longitudes de los peces en un lago específico son ejemplos de poblaciones de tamaño finito. En cada caso, el número total de observaciones es un número finito. Las observaciones que se obtienen al medir diariamente la presión atmosférica desde el pasado hasta el futuro, o todas las mediciones de la profundidad de un lago desde cualquier posición concebible, son ejemplos de poblaciones cuyos tamaños son infinitos. Algunas poblaciones finitas son tan grandes que en teoría las supondríamos infinitas, lo cual es cierto si se considera la población de la duración de cierto tipo de batería de almacenamiento que se fabrica para su distribución masiva en todo el país.

Cada observación en una población es un valor de una variable aleatoria X que tiene alguna distribución de probabilidad $f(x)$. Si se inspeccionan artículos que salen de una línea de ensamble para buscar defectos, entonces cada observación en la población podría ser un valor 0 o 1 de la variable aleatoria de Bernoulli X con distribución de probabilidad

$$b(x;1,\, p) = p^x q^{1-x}, \qquad x = 0,\, 1,$$

donde 0 indica un artículo no defectuoso y 1 indica uno defectuoso. Por supuesto, se supone que p, la probabilidad de que cualquier artículo esté defectuoso, permanece constante de una prueba a otra. En el experimento de tipo sanguíneo la variable aleatoria X representa el tipo de sangre al tomar un valor del 1 al 8. A cada estudiante se le asigna uno de los valores de la variable aleatoria discreta. Las duraciones de las baterías de almacenamiento son valores que toma una variable aleatoria continua que quizá tiene una distribución normal. De ahora en adelante, cuando nos refiramos a una "población binomial", a una "población normal" o, en general, a la "población $f(x)$", aludiremos a una población cuyas observaciones son valores de una variable aleatoria que tiene una distribución binomial, una distribución normal o la distribución de probabilidad $f(x)$. Por ello, a la media y a la varianza de una variable aleatoria o distribución de probabilidad también se les denomina la media y la varianza de la población correspondiente.

En el campo de la inferencia estadística el estadístico se interesa en llegar a conclusiones que tienen que ver con la población, cuando es imposible o poco práctico observar todo el conjunto de observaciones que constituyen la población. Por ejemplo, al intentar determinar la longitud promedio de la vida de cierta marca de bombilla de luz, sería imposible probar todas las bombillas si tenemos que venderlas. Los costos desmesurados también serían un factor prohibitivo para estudiar a toda la población. Por lo tanto, debemos depender de un subconjunto de observaciones de la población para ayudarnos a realizar inferencias con respecto a la misma población. Esto nos lleva a considerar la noción de muestreo.

Definición 8.2: Una **muestra** es un subconjunto de una población.

Si nuestras inferencias a partir de la muestra para la población tienen que ser válidas, debemos obtener muestras que sean representativas de la población. Con mucha

frecuencia nos sentimos tentados a elegir una muestra seleccionando a los miembros más convenientes de la población. Tal procedimiento podría conducir a inferencias erróneas con respecto a la población. Cualquier procedimiento de muestreo que produzca inferencias que sobreestimen, o subestimen, de forma consistente alguna característica de la población se dice que está **sesgado**. Para eliminar cualquier posibilidad de sesgo en el procedimiento de muestreo, es deseable elegir una **muestra aleatoria**, en el sentido de que las observaciones se realicen de forma independiente y al azar.

Para seleccionar una muestra aleatoria de tamaño n de una población $f(x)$, definamos la variable aleatoria X_i, $i = 1, 2, \ldots, n$, que represente la i-ésima medición o valor de la muestra que observemos. Las variables aleatorias $X_1, X_2, \ldots, X_n$ constituirán entonces una muestra aleatoria de la población $f(x)$ con valores numéricos $x_1, x_2, \ldots, x_n$ si las mediciones se obtienen al repetir el experimento n veces independientes bajo esencialmente las mismas condiciones. Debido a las condiciones idénticas bajo las que se seleccionan los elementos de la muestra, es razonable suponer que las n variables aleatorias $X_1, X_2, \ldots, X_n$ son independientes y que cada una tiene la misma distribución de probabilidad $f(x)$. Es decir, las distribuciones de probabilidad de $X_1, X_2, \ldots, X_n$ son, respectivamente, $f(x_1), f(x_2), \ldots, f(x_n)$ y su distribución de probabilidad conjunta es $f(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2) \cdots f(x_n)$. El concepto de muestra aleatoria se describe de manera formal en la siguiente definición.

Definición 8.3: Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes n, cada una con la misma distribución de probabilidad $f(x)$. Definimos $X_1, X_2, \ldots, X_n$ como una **muestra aleatoria** de tamaño n de la población $f(x)$ y escribimos su distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2) \cdots f(x_n).$$

Si se realiza una selección aleatoria de $n = 8$ baterías de almacenamiento de un proceso de fabricación, que mantiene las mismas especificaciones, y registramos la duración de cada batería con la primera medición x_1 como un valor de X_1, la segunda medición x_2 como un valor de X_2, etcétera, entonces, $x_1, x_2, \ldots, x_8$ son los valores de la muestra aleatoria $X_1, X_2, \ldots, X_8$ Si suponemos que la población de duraciones de las baterías es normal, los valores posibles de cualquier X_i, $i = 1, 2, \ldots, 8$, serán precisamente los mismos que los de la población original y, por ello, X_i tiene la misma distribución normal idéntica que X.

8.2 Algunos estadísticos importantes

Nuestro principal propósito al seleccionar muestras aleatorias consiste en obtener información acerca de los parámetros desconocidos de la población. Suponga, por ejemplo, que deseamos llegar a una conclusión con respecto a la proporción de personas bebedoras de café en Estados Unidos que prefieren cierta marca de café. Sería imposible preguntar a cada bebedor de café estadounidense para calcular el valor del parámetro p que representa la proporción de la población. En cambio, se selecciona una muestra aleatoria grande y se calcula la proporción $\hat{p}$ de personas en esta muestra que prefieren la marca de café en cuestión. El valor $\hat{p}$ se utiliza ahora para hacer una inferencia con respecto a la proporción p verdadera.

Ahora, $\hat{p}$ es una función de los valores observados en la muestra aleatoria; como es posible tomar muchas muestras aleatorias a partir de la misma población, esperaríamos

que $\hat{p}$ variara algo de una muestra a otra. Es decir, $\hat{p}$ es un valor de una variable aleatoria que representamos con P. Tal variable aleatoria se llama **estadístico**.

Definición 8.4: Cualquier función de las variables aleatorias que forman una muestra aleatoria se llama **estadístico**.

Tendencia central en la muestra

En el capítulo 4 presentamos los dos parámetros μ y σ^2, que miden el centro de localización y la variabilidad de una distribución de probabilidad. Éstos son parámetros poblacionales constantes y de ninguna manera se ven afectados o influidos por las observaciones de una muestra aleatoria. Definiremos, sin embargo, algunos estadísticos importantes que describen las medidas correspondientes de una muestra aleatoria. Los estadísticos que, por lo general, se utilizan más para medir el centro de un conjunto de datos, acomodados en orden de magnitud, son la **media**, la **mediana** y la **moda**. Los tres estadísticos se expusieron en el capítulo 1; no obstante, la media se define aquí de nuevo.

Definición 8.5: Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de tamaño n, entonces la **media de la muestra** se define mediante el estadístico

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Observe que el estadístico $\bar{X}$ toma el valor $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ cuando X_1 toma el valor x_1, X_2 toma el valor x_2 y así sucesivamente. En la práctica al valor de un estadístico, por lo general, se le da el mismo nombre del estadístico. Por ejemplo, el término *media de la muestra* se aplica tanto al estadístico $\bar{X}$ como a su valor calculado $\bar{x}$.

Hay una referencia previa a la media de la muestra que se hizo en el capítulo 1. Se dieron ejemplos para ilustrar el cálculo de la media de una muestra.

Como se expuso en el capítulo 1, una medida de tendencia central en la muestra no da por sí misma una indicación clara de la naturaleza de la muestra. De manera que también debe considerarse una medición de variabilidad en la muestra.

La varianza de la muestra

La variabilidad en la muestra debería indicar como se dispersan las observaciones a partir del promedio. Se remite al lector al capítulo 1 para un análisis más amplio. Es posible tener dos conjuntos de observaciones con las mismas media o mediana, y que difieran de manera considerable en la variabilidad de sus mediciones alrededor del promedio.

Considere las siguientes mediciones, en litros, para dos muestras de jugo de naranja envasado por las compañías A y B:

Muestra A	0.97	1.00	0.94	1.03	1.06
Muestra B	1.06	1.01	0.88	0.91	1.14

Ambas muestras tienen la misma media, 1.00 litros. Es muy evidente que la compañía A envasa el jugo de naranja con un contenido más uniforme que la B. Decimos

que la variabilidad o la dispersión de las observaciones del promedio es menor para la muestra A que para la muestra B. Por lo tanto, al comprar jugo de naranja, tendríamos más confianza de que el envase seleccionado esté más cerca del promedio anunciado si lo compramos a la compañía A.

En el capítulo 1 presentamos varias mediciones de la variabilidad de una muestra como la **varianza de la muestra** y el **rango de la muestra**. En este capítulo nos enfocaremmos en la varianza de la muestra.

Definición 8.6: Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de tamaño n, entonces la **varianza de la muestra** se define con el estadístico

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

El valor calculado de S^2 para una muestra dada se denota con s^2. Observe que S^2 se define esencialmente como el promedio de los cuadrados de las desviaciones de las observaciones de su media. La razón para utilizar $n-1$ como divisor, en vez de la elección más obvia n, quedará más clara en el capítulo 9.

Ejemplo 8.1: Una comparación de los precios de café en cuatro tiendas de abarrotes, seleccionadas al azar, en San Diego mostró aumentos en comparación con el mes anterior de 12, 15, 17 y 20 centavos para una bolsa de 1 libra. Encuentre la varianza de esta muestra aleatoria de aumentos de precio.

Solución: Al calcular la media de la muestra, obtenemos

$$\bar{x} = \frac{12+15+17+20}{4} = 16 \text{ centavos.}$$

Por lo tanto,

$$\begin{aligned} s^2 &= \frac{1}{3}\sum_{i=1}^{4}(x_i - 16)^2 = \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3} \\ &= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3}. \end{aligned}$$

Mientras que la expresión para la varianza de la muestra de la definición 8.6 ilustra mejor que S^2 es una medida de variabilidad, una expresión alternativa que en verdad tiene algún mérito, de manera que el lector debería estar consciente de ello. El siguiente teorema contiene tal expresión.

Teorema 8.1: Si S^2 es la varianza de una muestra aleatoria de tamaño n, podemos escribir

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}X_i^2 - \left(\sum_{i=1}^{n}X_i\right)^2\right].$$

Prueba: Por definición,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - 2\bar{X}\sum_{i=1}^{n}X_i + n\bar{X}^2\right].$$

Al reemplazar X por $\bar{X}$ por $\sum_{i=1}^{n} X_i/n$ y multiplicar el numerador y el denominador por n, obtenemos la fórmula de cálculo más útil del teorema 8.1.

Definición 8.7: La **desviación estándar de la muestra**, que se denota con S, es la raíz cuadrada positiva de la varianza de la muestra.

Ejemplo 8.2: Encuentre la varianza de los datos 3, 4, 5, 6, 6 y 7, que representan el número de truchas atrapadas por una muestra aleatoria de 6 pescadores el 19 de junio de 1996, en el lago Muskoka.

Solución: Encontramos que $\sum_{i=1}^{6} x_i^2 = 171$, $\sum_{i=1}^{6} x_i = 31$, $n = 6$. De aquí,

$$s^2 = \frac{1}{(6)(5)}[(6)(171) - (31)^2] = \frac{13}{6}.$$

Por lo que la desviación estándar de la muestra $s = \sqrt{13/6} = 1.47$.

Ejercicios

8.1 Defina las poblaciones adecuadas a partir de las cuales se seleccionaron las siguientes muestras:

a) Se llamó por teléfono a personas de 200 casas en la ciudad de Richmond y se les pidió nombrar al candidato por el que votarían en la elección para la mesa directiva de la escuela.

b) Se lanzó 100 veces una moneda y se registraron 34 cruces.

c) Se probaron 200 pares de un nuevo tipo de calzado deportivo en un torneo de tenis profesional y, en promedio, duraron cuatro meses.

d) En cinco ocasiones diferentes tomó a una abogada 21, 26, 24, 22 y 21 minutos manejar desde su casa en los suburbios hasta su oficina en el centro de la ciudad.

8.2 El número de multas emitidas por infracciones de tránsito por 8 oficiales estatales durante el fin de semana del día en Conmemoración de los Caídos es 5, 4, 7, 7, 6, 3, 8 y 6.

a) Si estos valores representan el número de multas levantadas por una muestra aleatoria de 8 oficiales estatales del condado de Montgomery en Virginia, defina una población adecuada.

b) Si los valores representan el número de multas levantadas por una muestra aleatoria de 8 oficiales estatales de Carolina del Sur, defina una población adecuada.

8.3 El número de respuestas incorrectas en un examen de competencia de verdadero-falso para una muestra aleatoria de 15 estudiantes se registraron de la siguiente manera: 2, 1, 3, 0, 1, 3, 6, 0, 3, 3, 5, 2, 1, 4 y 2. Encuentre

a) la media;

b) la mediana;

c) la moda.

8.4 Las longitudes de tiempo, en minutos, que 10 pacientes esperan en un consultorio médico antes de recibir tratamiento se registraron como sigue: 5, 11, 9, 5, 10, 15, 6, 10, 5 y 10. Trate los datos como una muestra aleatoria y encuentre

a) la media;

b) la mediana;

c) la moda.

8.5 Los tiempos de reacción para una muestra aleatoria de 9 individuos ante un estimulante se registraron como 2.5, 3.6, 3.1, 4.3, 2.9, 2.3, 2.6, 4.1 y 3.4 segundos. Calcule

a) la media;

b) la mediana.

8.6 De acuerdo con la escritora ecologista Jacqueline Killeen, los fosfatos que contienen los detergentes de uso casero pasan directamente a través de nuestros sistemas de desagüe, ocasionando que los lagos se conviertan en pantanos, y que a final de cuentas se sequen y se vuelvan desiertos. Los siguientes datos muestran la cantidad de fosfatos por carga de lavado, en gramos, para una muestra aleatoria de diversos tipos de detergentes que se usan de acuerdo con las instrucciones prescritas:

Detergente de lavander?a	**Fosfato por carga (gramos)**
A & P Blue Sail	48
Dash	47
Concentrated All	42
Cold Water All	42
Breeze	41
Oxydol	34
Ajax	31
Sears	30
Fab	29
Cold Power	29
Bold	29
Rinso	26

Para los datos de fosfato dados, encuentre

a) la media;

b) la mediana:

c) la moda.

8.7 Una muestra aleatoria de empleados de una planta de manufactura local prometieron los siguientes donativos, en dólares, al United Fund: 100, 40, 75, 15, 20, 100, 75, 50, 30, 10, 55, 75, 25, 50, 90, 80, 15, 25, 45 y 100. Calcule

a) la media;

b) la moda.

8.8 Encuentre la media, la mediana y la moda para la muestra, cuyas observaciones, 15, 7, 8, 95, 19, 12, 8, 22 y 14 representan el número de días con incapacidad médica reportados en nueve declaraciones federales de impuesto sobre la renta. ¿Qué valor parece ser la mejor medición del centro de nuestros datos? Explique las razones de su preferencia.

8.9 Con referencia a la longitud de los tiempos que esperan 10 pacientes en un consultorio médico antes de recibir tratamiento en el ejercicio 8.4, encuentre

a) el rango;

b) la desviación estándar.

8.10 Con referencia a la muestra de tiempos de reacción para los nueve sujetos que reciben el estimulante en el ejercicio 8.5, calcule

a) el rango;

b) la varianza usando la fórmula de la definición 8.6.

8.11 Con referencia a la muestra aleatoria de respuestas incorrectas en un examen de competencia de verdadero-falso para los 15 estudiantes en el ejercicio 8.3, calcule la varianza usando la fórmula

a) de la definición 8.6;

b) del teorema 8.1.

8.12 El contenido de alquitrán de ocho marcas de cigarrillos que se seleccionan al azar de la lista más reciente publicada por la Comisión de Comercio Federal es como sigue: 7.3, 8.6, 10.4, 16.1, 12.2, 15.1, 14.5 y 9.3 miligramos. Calcule

a) la media;

b) la varianza.

8.13 Los promedios de los puntos por grado de 20 estudiantes universitarios del último año seleccionados al azar de una clase que se va a graduar son los siguientes:

3.2	1.9	2.7	2.4	2.8
2.9	3.8	3.0	2.5	3.3
1.8	2.5	3.7	2.8	2.0
3.2	2.3	2.1	2.5	1.9

Calcule la desviación estándar.

8.14 *a*) Muestre que la varianza de la muestra permanece sin cambio, si se suma o se resta una constante c a cada valor de la muestra.

b) Muestre que la varianza de la muestra se hace c^2 veces su valor original, si cada observación en la muestra se multiplica por c.

8.15 Verifique que la varianza de la muestra 4, 9, 3, 6, 4 y 7 es 5.1, y usando este hecho junto con los resultados del ejercicio 8.14, encuentre

a) la varianza de la muestra 12, 27, 9, 18, 12 y 21;

b) la varianza de la muestra 9, 14, 8, 11, 9 y 12.

8.16 En la temporada 2004-2005 el equipo de fútbol americano de la Universidad del Sur de California tuvo las siguientes diferencias de puntuación para sus 13 juegos disputados.

11 49 32 3 6 38 38 30 8 40 31 5 36

Encuentre

a) la media de las diferencia de puntos;

b) la mediana de las diferencias de puntos.

8.3 Presentación de datos y métodos gráficos

En el capítulo 1 presentamos al lector las distribuciones empíricas. La motivación es el uso de presentaciones creativas para extraer información acerca de las propiedades de un conjunto de datos. Por ejemplo, los diagramas de tallo y hojas brindan al observador una imagen de la simetría y de otras propiedades de los datos. En este capítulo tratamos con muestras que, por supuesto, son agrupaciones de datos experimentales a partir de las cuales obtenemos conclusiones sobre las poblaciones. A menudo la apariencia de la muestra proporciona información acerca de la distribución de la que se toman los datos. Por ejemplo, en el capítulo 1 ilustramos la naturaleza general de pares de muestras con gráficas de puntos que presentan una comparación relativa entre la tendencia central y la variabilidad entre ambas muestras.

En los capítulos siguientes, con frecuencia hacemos la suposición de que la distribución es normal. La información gráfica con respecto a la validez de esta suposición se puede obtener de presentaciones como los diagramas de tallo y hojas y los histogramas de frecuencias. Además, en esta sección presentaremos la noción de *gráficas de probabilidad normal* y *gráficas de cuantiles*. Estas gráficas se utilizan en estudios que tienen grados de complejidad que varían, con el objetivo principal de que las gráficas den una verificación diagnóstica sobre la suposición de que los datos vienen de una distribución normal.

Podemos caracterizar el análisis estadístico como el proceso de extraer conclusiones acerca de los sistemas en presencia de la variabilidad del sistema. El intento de un ingeniero por aprender acerca de un proceso químico a menudo se ve empañado por la *variabilidad del proceso*. Un estudio que implica el número de artículos defectuosos en un proceso de producción con frecuencia se hace más difícil por la variabilidad en el método de fabricación de los artículos. En todo lo anterior, aprendimos acerca de las muestras y los estadísticos que expresan el centro de localización y la variabilidad en la muestra. Tales estadísticos ofrecen medidas simples, en tanto que una presentación gráfica brinda información adicional en términos de una imagen.

Gráfica de caja y extensión o gráfica de caja

Otra presentación que es útil para reflejar propiedades de una muestra es la **gráfica de caja y extensión**, la cual encierra el *rango intercuartil* de los datos en una caja que tiene la mediana representada dentro. El rango intercuartil tiene como extremos el percentil 75 (cuartil superior) y el percentil 25 (cuartil inferior). Además de la caja se prolongan "extensiones", que indican las observaciones alejadas en la muestra. Para muestras razonablemente grandes la presentación indica el centro de localización, la variabilidad y el grado de asimetría.

Además, una variación denominada **gráfica de caja** puede ofrecer al observador información con respecto a cuáles observaciones son **valores extremos**, los cuales son observaciones que se consideran inusualmente alejadas de la masa de datos. Hay muchas pruebas estadísticas diseñadas para detectar valores extremos. Técnicamente, se puede considerar que un valor extremo es una observación que representa un "evento raro" (existe una probabilidad pequeña de obtener un valor tan alejado de la masa de datos). El concepto de valores extremos resurge en el capítulo 12 en el contexto del análisis de regresión.

La información visual en las gráficas de caja y extensión y de caja no intenta ser una prueba formal de valores extremos. Más bien, se ve como una herramienta de diagnóstico. Mientras que la determinación de cuáles observaciones son valores extremos varía con el tipo de software que se emplee, un procedimiento común consiste

en utilizar un **múltiplo del rango intercuartil**. Por ejemplo, si la distancia desde la caja excede 1.5 veces el rango intercuartil (en cualquier dirección) la observación se puede considerar un valor extremo.

Ejemplo 8.3: Considere los datos del ejercicio 1.21 al final del capítulo 1 (página 29). Se midió el contenido de nicotina en una muestra aleatoria de 40 cigarrillos. Los datos se vuelven a presentar en la tabla 8.1.

Tabla 8.1: Valores de nicotina para el ejemplo 8.3

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

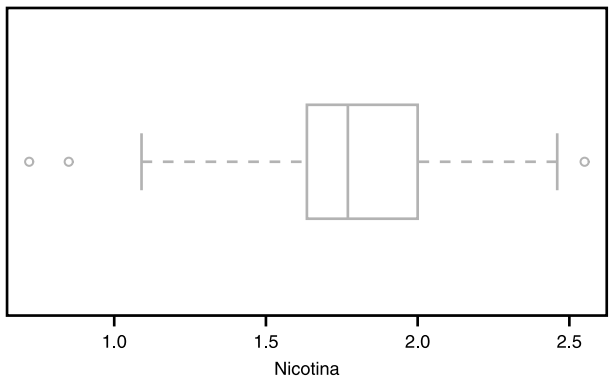


Figura 8.1: Gráfica de caja y extensión para los datos de nicotina del ejercicio 1.21.

La figura 8.1 muestra la gráfica de caja y extensión de los datos que describe las observaciones 0.72 y 0.85 como valores extremos moderados en la cola inferior; en tanto que la observación 2.55 es un valor extremo moderado en la cola superior. En este ejemplo el rango intercuartil es 0.365, y 1.5 veces el rango intercuartil es 0.5475. Por otro lado, la figura 8.2 presenta un diagrama de tallo y hojas.

Ejemplo 8.4: Considere los datos de la tabla 8.2, que consisten en 30 muestras que miden el espesor de las "asas" de latas de pintura (véase el trabajo de Hogg y Ledolter en la bibliografía). La figura 8.3 describe una gráfica de caja y extensión para este conjunto asimétrico de datos. Observe que el bloque izquierdo es considerablemente más grande que el bloque de la derecha. La mediana es 35. El cuartil inferior es 31, mientras que el superior es 36. Note también que la observación alejada de la derecha está más lejos de la caja que la observación alejada de la izquierda. No hay valores extremos en este conjunto de datos.

Existen formas adicionales en las que las gráficas de caja y extensión y otras presentaciones gráficas ayudan al analista. Las muestras múltiples se pueden comparar

```
The decimal point is 1 digit(s) to the left of the |
 7 | 2
 8 | 5
 9 |
10 | 9
11 |
12 | 4
13 | 7
14 | 07
15 | 18
16 | 3447899
17 | 045599
18 | 2568
19 | 0237
20 | 389
21 | 17
22 | 8
23 | 17
24 | 6
25 | 5
```

Figura 8.2: Diagrama de tallo y hojas para los datos de nicotina.

Tabla 8.2: Datos para el ejemplo 8.4

Muestra	Mediciones	Muestra	Mediciones
1	29 36 39 34 34	16	35 30 35 29 37
2	29 29 28 32 31	17	40 31 38 35 31
3	34 34 39 38 37	18	35 36 30 33 32
4	35 37 33 38 41	19	35 34 35 30 36
5	30 29 31 38 29	20	35 35 31 38 36
6	34 31 37 39 36	21	32 36 36 32 36
7	30 35 33 40 36	22	36 37 32 34 34
8	28 28 31 34 30	23	29 34 33 37 35
9	32 36 38 38 35	24	36 36 35 37 37
10	35 30 37 35 31	25	36 30 35 33 31
11	35 30 35 38 35	26	35 30 29 38 35
12	38 34 35 35 31	27	35 36 30 34 36
13	34 35 33 30 34	28	35 30 36 29 35
14	40 35 34 33 35	29	38 36 35 31 31
15	34 35 38 35 30	30	30 34 40 28 30

de forma gráfica. Las gráficas de datos pueden sugerir relaciones entre variables. Las gráficas ayudan en la detección de anomalías o de observaciones de valores extremos en las muestras.

Otro tipo de gráfica que en particular podría ser útil para caracterizar la naturaleza de un conjunto de datos es la *gráfica de cuantiles*. Como en el caso de la gráfica de caja y extensión, se pueden utilizar las ideas básicas de la gráfica de cuantiles para *comparar muestras de datos*, donde el objetivo del analista es encontrar diferencias. Ilustraciones adicionales de este tipo de utilización se darán en los capítulos siguientes, donde se estudia la inferencia estadística formal asociada con la comparación de las muestras. En ese momento se demostrarán estudios de caso, en los cuales

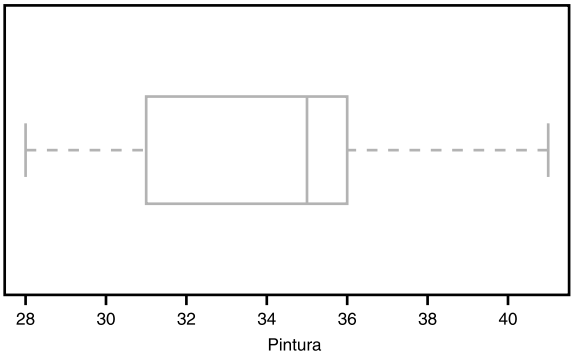


Figura 8.3: Gráfica de caja y extensión para el espesor de las "asas" de latas de pintura.

se expone al lector a la inferencia formal y a las gráficas de diagnóstico para el mismo conjunto de datos.

Gráfica de cuantiles

El propósito de las gráficas de cuantiles consiste en describir, en forma de muestra, la función de distribución acumulada que se presentó en el capítulo 3.

Definición 8.8: Un **cuantil** de una muestra, $q(f)$, es un valor para el que una fracción específica f de los valores de los datos es menor que o igual a $q(f)$.

Evidentemente, un cuantil representa una estimación de una característica de una población o, más bien, la distribución teórica. La mediana de la muestra es $q(0.5)$. El percentil 75 (cuartil superior) es $q(0.75)$ y el cuartil inferior es $q(0.25)$.

Una **gráfica de cuantiles** simplemente *grafica los valores de los datos en el eje vertical contra una evaluación empírica de la fracción de observaciones excedidas por los valores de los datos*. Para propósitos teóricos esta fracción se calcula con

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}},$$

donde i es el orden de las observaciones cuando se clasifican de inferior a superior. En otras palabras, si denotamos las observaciones clasificadas como

$$y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \cdots \leq y_{(n-1)} \leq y_{(n)},$$

entonces, la gráfica de cuantiles describe una gráfica de $y_{(i)}$ contra f_i. En la figura 8.4 se presenta la gráfica de cuantiles para las asas de las latas de pintura analizadas con anterioridad.

A diferencia de la gráfica de caja y extensión, la gráfica de cuantiles realmente muestra todas las observaciones. Todos los cuantiles, incluidos la mediana y los cuantiles superior e inferior, se pueden aproximar de forma visual. Por ejemplo, fácilmente observamos una mediana de 35 y un cuartil superior de alrededor de 36. Las indicaciones de agrupaciones relativamente grandes alrededor de valores específicos se indican por pendientes cercanas a cero; mientras que los datos escasos en ciertas áreas producen pendientes más abruptas. La figura 8.4 muestra la dispersión de datos de los valores 28 a 30, pero una densidad relativamente alta de 36 a 38. En los capítulos 9 y 10 proseguimos con las gráficas de cuantiles mediante la ilustración de formas útiles de comparación de distintas muestras.

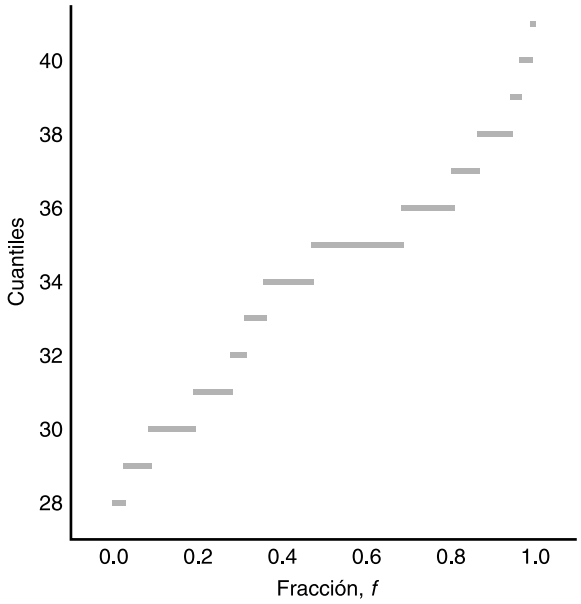


Figura 8.4: Gráfica de cuantiles para los datos de latas de pintura.

Detección de desviaciones de la normalidad

Para el lector debería ser algo evidente que la detección de un conjunto de datos viene o no de una distribución normal que puede ser una herramienta importante para el analista de datos. Como indicamos anteriormente en esta sección, con frecuencia hacemos la suposición de que la totalidad o subconjuntos de las observaciones en un conjunto de datos son realizaciones de variables aleatorias normales independientes e idénticamente distribuidas. Una vez más, la gráfica de diagnóstico a menudo puede agregar (con fines de presentación) una *prueba de la bondad del ajuste* formal de los datos. Las pruebas de bondad del ajuste se estudian en el capítulo 10. Para el lector de un artículo o informe científico, la información de diagnóstico resulta mucho más clara, menos árida y quizá no aburrida. En los capítulos siguientes (9 a 13) nos enfocamos de nuevo en los métodos de detección de desviaciones de la normalidad como un agregado de la inferencia estadística formal. Estos tipos de gráficas son útiles en la detección de los tipos de distribución. En la elaboración de modelos y en el diseño de experimentos también hay situaciones en que las gráficas se utilizan para detectar **términos** o **efectos del modelo** que están activos. En otras situaciones se utilizan para determinar si son razonables o no las suposiciones subyacentes hechas por el científico o por el ingeniero en la construcción del modelo. En los capítulos 11, 12 y 13 se incluyen muchos ejemplos con ilustraciones. La siguiente subsección brinda una presentación e ilustración de una gráfica de diagnóstico que se llama *gráfica de cuantiles-cuantiles normales*.

Gráfica de cuantiles-cuantiles normales

La gráfica de cuantiles-cuantiles normales toma ventaja de lo que se conoce acerca de los cuantiles de la distribución normal. La metodología incluye una gráfica de los cuantiles empíricos recién presentados contra el cuantil correspondiente de la dis-

tribución normal. Entonces, la expresión para un cuantil de una variable aleatoria $N(\mu, \sigma)$ es muy complicada. Sin embargo, una buena aproximación está dada por

$$q_{\mu,\sigma}(f) = \mu + \sigma\{4.91[f^{0.14} - (1-f)^{0.14}]\}.$$

La expresión entre las llaves (el múltiplo de σ) es la aproximación para el cuantil correspondiente para la variable aleatoria $N(0, 1)$, es decir,

$$q_{0,1}(f) = 4.91[f^{0.14} - (1-f)^{0.14}].$$

Definición 8.9: La **gráfica de cuantiles-cuantiles normales** es una gráfica de $y_{(i)}$ (observaciones ordenadas contra $q_{0,1}(f_i)$, donde $f_i = \frac{i-\frac{3}{8}}{n+\frac{1}{4}}$.

Una relación cercana a una línea recta sugiere que los datos provienen de una distribución normal. La intersección en el eje vertical es una estimación de la media de la población μ y la pendiente es una estimación de la desviación estándar σ. La figura 8.5 muestra una gráfica de cuantiles-cuantiles normales para los datos de las latas de pintura.

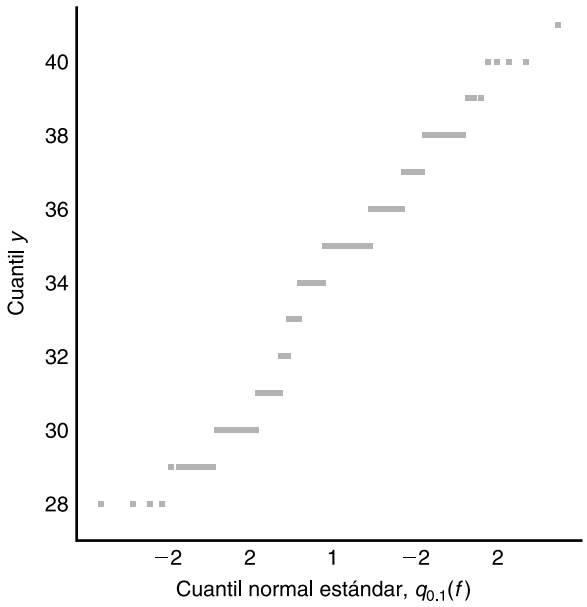


Figura 8.5: Gráfica de cuantiles-cuantiles normales para los datos de la pintura.

Graficación de la probabilidad normal

Observe cómo la desviación de la normalidad se vuelve clara a partir de la apariencia de la gráfica. La asimetría que exhiben los datos tiene como resultado cambios en la pendiente.

Las ideas de graficación de la probabilidad se manifiestan en gráficas diferentes de la gráfica de cuantiles-cuantiles normales que se presentó aquí. Por ejemplo, se pone mucha atención a la llamada **gráfica de probabilidad normal**, donde el eje vertical contiene f graficada en un papel especial y la escala utilizada da como resultado una línea recta, cuando se grafica contra los valores de los datos ordenados. Además, una gráfica alternativa utiliza los valores esperados de las observaciones clasificadas para la distribución normal y grafica las observaciones clasificadas contra su valor esperado, con la suposición de datos de $N(\mu, \sigma)$. Una vez más, la línea recta es el criterio gráfico que se emplea. Continuamos para sugerir que tener como base los métodos analíticos gráficos que se desarrollan en esta sección ayuda en la ilustración de los métodos formales de distinción entre muestras diferentes de datos.

Ejemplo 8.5: Considere los datos del ejercicio 10.41 de la página 359 del capítulo 10. En un estudio, *Retención de nutrientes y respuesta de comunidades de macroinvertebrados ante la presión de aguas residuales en un ecosistema fluvial*, que se llevó a cabo en el departamento de zoología del Instituto Politécnico y la Universidad Estatal de Virginia, se recabaron datos sobre mediciones de densidad (número de organismos por metro cuadrado) en dos diferentes estaciones colectoras. En el capítulo 10 se dan detalles con respecto a los métodos analíticos de comparación de muestras, para determinar si ambas son de la misma distribución $N(\mu, \sigma)$ Los datos se presentan en la tabla 8.3.

Tabla 8.3: Datos para el ejemplo 8.5

N?mero de organismos por metro cuadrado			
Estaci?n 1		**Estaci?n 2**	
5,030	4,980	2,800	2,810
13,700	11,910	4,670	1,330
10,730	8,130	6,890	3,320
11,400	26,850	7,720	1,230
860	17,660	7,030	2,130
2,200	22,800	7,330	2,190
4,250	1,130		
15,040	1,690		

Construya una gráfica de cuantiles-cuantiles normales y obtenga conclusiones con respecto a si es razonable o no suponer que las dos muestras son de la misma distribución $n(x; \mu, \sigma)$.

Solución: La figura 8.5 muestra la gráfica de cuantiles-cuantiles normales para las mediciones de densidad. La gráfica muestra una apariencia que está lejos de una sola línea recta. De hecho, los datos de la estación 1 reflejan pocos valores en la cola inferior de la distribución y varios en la cola superior. El "agrupamiento" de observaciones hace que parezca improbable que las dos muestras vengan de una distribución común $N(\mu, \sigma)$.

Aunque hemos concentrado nuestro desarrollo e ilustración en la graficación de la probabilidad para distribuciones normales, podemos enfocarnos en cualquier distribución. Tan sólo necesitaríamos calcular cantidades de forma analítica para la distribución teórica en cuestión.

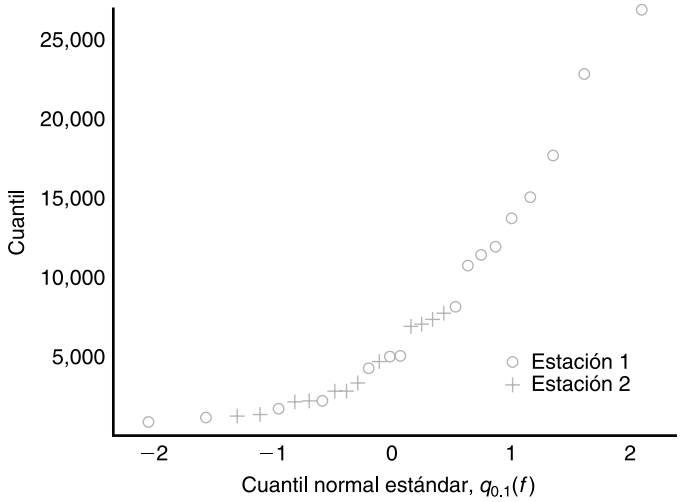


Figura 8.6: Gráfica de cuantiles-cuantiles normales para los datos de densidad del ejemplo 8.5.

8.4 Distribuciones muestrales

El campo de la inferencia estadística trata básicamente con generalizaciones y predicciones. Por ejemplo, podemos afirmar, con base en las opiniones de varias personas entrevistadas en la calle, que en una próxima elección 60% de los votantes en la ciudad de Detroit favorecerían a cierto candidato. En este caso, tratamos con una muestra aleatoria de opiniones de una población finita muy grande. Como segunda ilustración podemos afirmar que el costo promedio de construir una residencia en Charleston, Carolina del Sur, está entre $230,000 y 235,000, con base en las estimaciones de tres contratistas seleccionados al azar de 30 que laboran actualmente en esta ciudad. La población que se va a muestrear aquí nuevamente es finita pero muy pequeña. Finalmente, consideremos una máquina despachadora de bebida gaseosa en la cual la cantidad promedio de bebida servida se mantiene en 240 mililitros. Un funcionario de la compañía calcula la media de 40 bebidas y obtiene $\bar{x} = 236$ mililitros y, con base en este valor, decide que la máquina aún sirve bebidas con un contenido promedio de $\mu = 240$ mililitros. Las 40 bebidas representan una muestra de la población infinita de posibles bebidas que esta máquina servirá.

Inferencias sobre la población a partir de información de la muestra

En cada uno de los ejemplos anteriores calculamos un estadístico a partir de una muestra que se selecciona de la población, y con base en tales estadísticos hacemos varias afirmaciones con respecto a los valores de los parámetros de la población, que pueden ser ciertas o no. El funcionario de la compañía toma la decisión de que la máquina despachadora sirve bebidas con un contenido promedio de 240 mililitros, aun cuando la media de la muestra fue de 236 mililitros, porque sabe de la teoría del muestreo que es probable que ocurra tal valor de la muestra. De hecho, si realiza pruebas similares, digamos cada hora, esperaría que los valores de $\bar{x}$ fluctuaran por arriba y por abajo de $\mu = 240$ mililitros. Sólo cuando el valor de $\bar{x}$ es considerable-

mente diferente de 240 mililitros el funcionario de la compañía iniciará una acción para ajustar la máquina.
Como un estadístico es una variable aleatoria que depende sólo de la muestra observada, debe tener una distribución de probabilidad.

Definición 8.10: La distribución de probabilidad de un estadístico se llama **distribución muestral**.

La distribución de probabilidad de $\bar{X}$ se llama **distribución muestral de la media**.

La distribución muestral de un estadístico depende del tamaño de la población, del tamaño de las muestras y del método de elección de éstas. En el resto de este capítulo estudiaremos varias de las distribuciones muestrales más importantes de los estadísticos que se utilizan con frecuencia. Las aplicaciones de tales distribuciones muestrales a problemas de inferencia estadística se consideran en la mayoría de los capítulos posteriores.

¿Cuál es la distribución muestral de $\bar{X}$?

Se deberían ver las distribuciones muestrales de $\bar{X}$ y S^2 como el mecanismo a partir del cual a final de cuentas realizaremos inferencias de los parámetros μ y σ^2. La distribución muestral de $\bar{X}$ con tamaño muestral n es la distribución que resulta cuando un **experimento se lleva a cabo una y otra vez** (siempre con tamaño de la muestra n) y **resultan los diversos valores de** $\bar{X}$. Esta distribución muestral, entonces, describe la variabilidad de los promedios muestrales alrededor de la media de la población μ. En el caso de la máquina despachadora de bebida gaseosa, el conocimiento de la distribución muestral de $\bar{X}$ ofrece al analista el conocimiento de una discrepancia "típica" entre un valor $\bar{x}$ observado y el verdadero de μ. Se aplica el mismo principio en el caso de la distribución de S^2. La distribución muestral produce información acerca de la variabilidad de los valores de s^2 alrededor de σ^2 en experimentos que se repiten.

8.5 Distribuciones muestrales de medias

La primera distribución muestral importante que se debe considerar es la de la media $\bar{X}$. Suponga que una muestra aleatoria de n observaciones se toma de una población normal con media μ y varianza σ^2. Cada observación X_i, $i = 1, 2, \ldots, n$, de la muestra aleatoria tendrá entonces la misma distribución normal que la población que se muestrea. De aquí, por la propiedad reproductiva de la distribución normal que se establece en el teorema 7.11, concluimos que

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

tiene distribución normal con media

$$\mu_{\bar{X}} = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ términos}}) = \mu,$$

y varianza

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ términos}}) = \frac{\sigma^2}{n}.$$

Si tomamos muestras de una población con distribución desconocida, ya sea finita o infinita, la distribución muestral de $\bar{X}$ aún será aproximadamente normal con media μ y varianza σ^2/n siempre que el tamaño de la muestra sea grande. Este resultado asombroso es una consecuencia inmediata del siguiente teorema, que se conoce como teorema del límite central.

Teorema 8.2:

Teorema del límite central: Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n tomada de una población con media μ y varianza finita σ^2, entonces la forma límite de la distribución de

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

conforme $n \to \infty$, es la distribución normal estándar $n(z; 0, 1)$.

La aproximación normal para $\bar{X}$, por lo general, será buena si $n \geq 30$. Si $n < 30$, la aproximación es buena sólo si la población no es muy diferente de una distribución normal y, como se estableció antes, si se sabe que la población es normal, la distribución muestral de $\bar{X}$ seguirá una distribución normal exacta, no importa qué tan pequeño sea el tamaño de las muestras.

El tamaño de la muestra $n = 30$ es un lineamiento para usar para el teorema del límite central. No obstante, como indica la declaración del teorema, la suposición de normalidad en la distribución de $\bar{X}$ se vuelve más precisa conforme n se hace más grande. De hecho, la figura 8.7 ilustra cómo funciona el teorema. Indica cómo la distribución de $\bar{X}$ se hace más cercana a la normal conforme n aumenta, empezando con la distribución claramente asimétrica de una observación individual ($n = 1$). También ilustra que la media de $\bar{X}$ es μ para cualquier tamaño de la muestra y la varianza de $\bar{X}$ se vuelve más pequeña conforme n aumenta.

Como uno podría esperar, la distribución de $\bar{X}$ estará cercana a la normal para el tamaño de la muestra $n < 30$, si la distribución de una observación individual se acerca a la normal.

Ejemplo 8.6: Una empresa de material eléctrico fabrica bombillas de luz que tienen una duración que se distribuye aproximadamente en forma normal, con media de 800 horas y desviación estándar de 40 horas. Encuentre la probabilidad de que una muestra aleatoria de 16 bombillas tenga una vida promedio de menos de 775 horas.

Solución: La distribución muestral de $\bar{X}$ será aproximadamente normal, con $\mu_{\bar{X}} = 800$ y $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$. La probabilidad que se desea está dada por el área de la región sombreada de la figura 8.8.

En correspondencia con $\bar{x} = 775$, encontramos que

$$z = \frac{775 - 800}{10} = -2.5,$$

y, por lo tanto,

$$P(\bar{X} < 775) = P(Z < -2.5) = 0.0062.$$

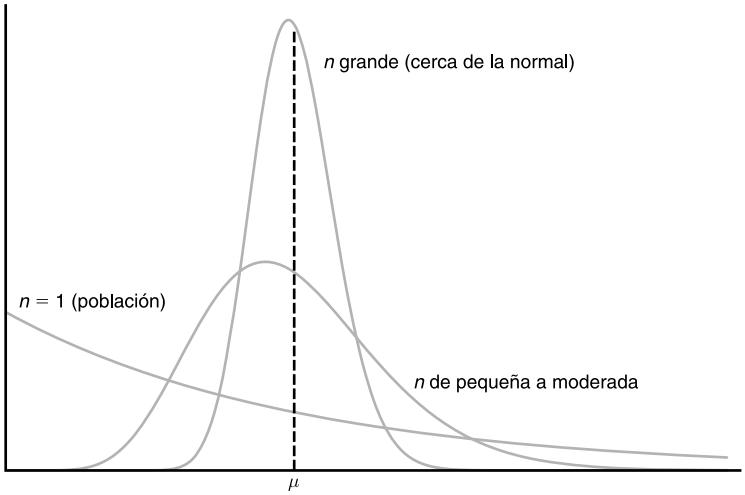


Figura 8.7: Ilustración del teorema del límite central (distribución de $\bar{X}$ para $n = 1$, n moderada y n grande).

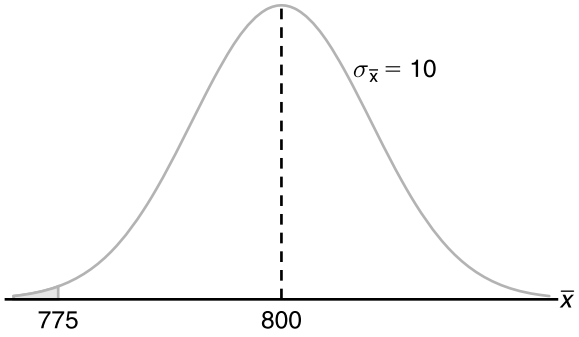


Figura 8.8: Área para el ejemplo 8.6.

Inferencias sobre la media de la población

Una aplicación muy importante del teorema del límite central consiste en determinar valores razonables de la media de la población μ. Temas como prueba de hipótesis, estimación, control de calidad y otros utilizan el teorema del límite central. El siguiente ejemplo ilustra el uso del teorema del límite central a este respecto, aunque la aplicación formal de los temas anteriores se deja para capítulos futuros.

Ejemplo 8.7: Un importante proceso de fabricación produce partes de componentes cilíndricos para la industria automotriz. Es importante que el proceso produzca partes que tengan una media de 5 milímetros. El ingeniero implicado hace la conjetura de que la media de la población es de 5.0 milímetros. Se lleva a cabo un experimento donde se seleccionan al azar 100 partes elaboradas por el proceso y se mide el diámetro de

cada una de ellas. Se sabe que la desviación estándar de la población es $\sigma = 0.1$. El experimento indica un diámetro promedio de la muestra $\bar{x} = 5.027$ milímetros. ¿Esta información de la muestra parece apoyar o refutar la conjetura del ingeniero?

Solución: Este ejemplo refleja la clase de problema que se plantea a menudo y que se resuelve con la maquinaria de la prueba de hipótesis que se presenta en los siguientes capítulos. No utilizaremos aquí el formalismo asociado con la prueba de hipótesis; pero ilustraremos los principios y la lógica que se utilizan.

Si los datos apoyan o refutan la conjetura depende de la probabilidad de que datos similares a los que se obtuvieron en este experimento ($\bar{x} = 5.027$) pueden ocurrir con facilidad cuando de hecho $\mu = 5.0$ (figura 8.9). En otras palabras, ¿qué tan probable es que se pueda obtener $\bar{x} \geq 5.027$ con $n = 100$, si la media de la población $\mu = 5.0$? Si esta probabilidad sugiere que $\bar{x} = 5.027$ no es poco razonable, no se refuta la conjetura. Si la probabilidad es bastante baja, se puede argumentar con certidumbre que los datos no apoyan la conjetura de que $\mu = 5.0$. La probabilidad que elijamos para calcular está dada por $P(|\bar{X} - 5| \geq 0.027)$.

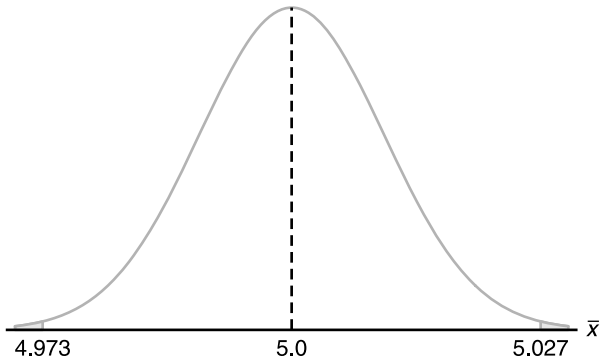


Figura 8.9: Área para el ejemplo 8.7.

En otras palabras, si la media μ es 5, ¿cuál es la probabilidad de que $\bar{X}$ se desvíe a lo más en 0.027 milímetros?

$$\begin{aligned} P(|\bar{X} - 5| \geq 0.027) &= P(\bar{X} - 5 \geq 0.027) + P(\bar{X} - 5 \leq -0.027) \\ &= 2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right). \end{aligned}$$

Aquí simplemente estandarizamos $\bar{X}$ de acuerdo con el teorema del límite central. Si es cierta la conjetura $\mu = 5.0$, $\frac{\bar{X}-5}{0.1/\sqrt{100}}$ debería ser $N(0, 1)$. Así,

$$2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right) = 2P(Z \geq 2.7) = 2(0.0035) = 0.007.$$

De esta manera, se experimentaría por casualidad una $\bar{x}$ que está a 0.027 milímetros de la media en tan sólo 7 de 1000 experimentos. Como resultado, este experimento con $\bar{x} = 5.027$ ciertamente no ofrece evidencia que apoye la conjetura de que $\mu = 5.0$. De hecho, firmemente refuta la conjetura!

Distribución muestral de la diferencia entre dos promedios

La ilustración del ejemplo 8.7 trata con nociones de inferencia estadística sobre una sola media μ. El ingeniero estaba interesado en apoyar una conjetura con respecto a una sola media de la población. Una aplicación mucho más importante incluye dos poblaciones. Un científico o ingeniero se interesa en un experimento comparativo donde se comparan dos métodos de producción: 1 y 2. La base para tal comparación es $\mu_1 - \mu_2$, la diferencia en las medias de las poblaciones.

Suponga que tenemos dos poblaciones, la primera con media μ_1 y varianza σ_1^2, y la segunda con media μ_2 y varianza σ_2^2. Representemos con el estadístico $\bar{X}_1$ la media de una muestra aleatoria de tamaño n_1 seleccionada de la primera población, y con el estadístico $\bar{X}_2$ la media de una muestra aleatoria de tamaño n_2 seleccionada de la segunda población, independiente de la muestra de la primera población. ¿Qué podríamos decir acerca de la distribución de muestreo de la diferencia $\bar{X}_1 - \bar{X}_2$ para muestras repetidas de tamaño n_1 y n_2? De acuerdo con el teorema 8.2, las variables $\bar{X}_1$ y $\bar{X}_2$ están distribuidas aproximadamente de forma normal con medias μ_1 y μ_2 y varianzas σ_1^2/n_1 y σ_2^2/n_2, respectivamente. Esta aproximación mejora conforme n_1 y n_2 aumentan. Al elegir muestras independientes de las dos poblaciones, las variables $\bar{X}_1$ y $\bar{X}_2$ serán independientes y entonces, usando el teorema 7.11, con $a_1 = 1$ y $a_2 = -1$, concluimos que $\bar{X}_1 - \bar{X}_2$ está distribuida aproximadamente de forma normal con media

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

y varianza

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Teorema 8.3: Si se extraen al azar muestras independientes de tamaños n_1 y n_2 de dos poblaciones, discretas o continuas, con medias μ_2 y μ_2 y varianzas σ_1^2 y σ_2^2 respectivamente, entonces la distribución muestral de las diferencias de las medias, $\bar{X}_1 - \bar{X}_2$, está distribuida aproximadamente de forma normal con media y varianza dadas por

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2, \quad \text{y} \qquad \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

De aquí,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

es aproximadamente una variable normal estándar.

Si tanto n_1 como n_2 son mayores que o iguales a 30, la aproximación normal para la distribución de $\bar{X}_1 - \bar{X}_2$ es muy buena cuando las distribuciones subyacentes no están tan alejadas de la normal. Sin embargo, aun cuando n_1 y n_2 sean menores que 30, la aproximación normal es razonablemente buena excepto cuando las poblaciones no son definitivamente normales. Por supuesto, si ambas poblaciones son normales, entonces $\bar{X}_1 - \bar{X}_2$ tiene una distribución normal sin importar cuáles son los tamaños de n_1 y n_2.

Ejemplo 8.8: Se llevan a cabo dos experimentos independientes en los que se comparan dos tipos diferentes de pintura. Se pintan 18 especímenes con el tipo A y en cada uno se registra el tiempo de secado en horas. Lo mismo se hace con el tipo B. Se sabe que las desviaciones estándar de la población son ambas 1.0.

Suponiendo que el tiempo medio de secado es igual para los dos tipos de pintura, encuentre $P(\bar{X}_A - \bar{X}_B > 1.0)$, donde $\bar{X}_A$ y $\bar{X}_B$ son los tiempos promedio de secado para muestras de tamaño $n_A = n_B = 18$.

Solución: De la distribución de muestreo de $\bar{X}_A - \bar{X}_B$, sabemos que la distribución es aproximadamente normal con media

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B = 0,$$

y varianza

$$\sigma^2_{\bar{X}_A - \bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

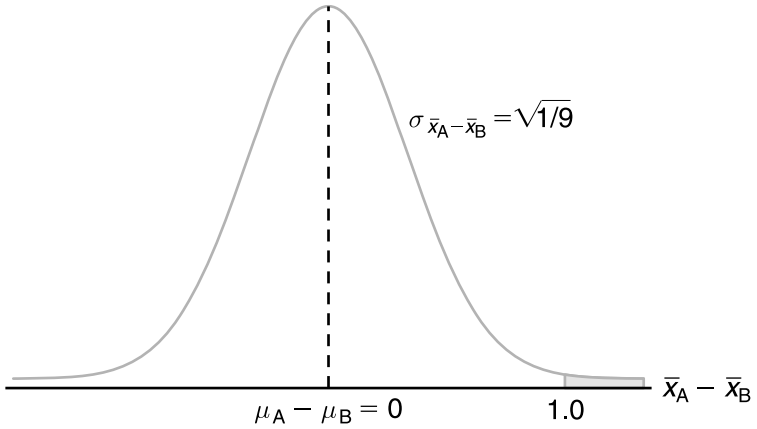


Figura 8.10: Área para el ejemplo 8.8.

La probabilidad que se desea está dada por la región sombreada en la figura 8.10. En correspondencia con el valor $\bar{X}_A - \bar{X}_B = 1.0$, tenemos

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3.0;$$

por lo que

$$P(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013.$$

¿Qué aprendemos con este ejemplo?

La maquinaria de cálculo se basa en la suposición de que $\mu_A = \mu_B$. Suponga, sin embargo, que el experimento realmente se lleva a cabo con la finalidad de obtener

una inferencia con respecto a la igualdad de μ_A y μ_B, los tiempos medios de secado de las dos poblaciones. Si los dos promedios difieren por una hora (o más), esto claramente es una evidencia que nos llevaría a concluir que el tiempo medio de secado de la población no es igual para los dos tipos de pintura. Por otro lado, suponga que la diferencia en los dos promedios muestrales es tan pequeña como, digamos, 15 minutos. Si $\mu_A = \mu_B$,

$$P[(\bar{X}_A - \bar{X}_B) > 0.25 \text{ horas}] = P\left(\frac{\bar{X}_A - \bar{X}_B - 0}{\sqrt{1/9}} > \frac{3}{4}\right)$$
$$= P\left(Z > \frac{3}{4}\right) = 1 - P(Z < 0.75) = 1 - 0.7734 = 0.2266.$$

Como esta probabilidad no es baja, se concluiría que una diferencia de 15 minutos en las medias de las muestras puede ocurrir por casualidad (es decir, sucede con frecuencia aun cuando $\mu_A = \mu_B$). Como resultado, este tipo de diferencia en el tiempo promedio de secado ciertamente *no es una señal clara* de que $\mu_A \neq \mu_B$.

Como indicamos al principio, en los capítulos siguientes se proporcionará más formalismo con respecto a éste y a otros tipos de inferencia estadística (por ejemplo, la prueba de hipótesis). El teorema del límite central y las distribuciones de muestreo que se presentan en las siguientes tres secciones también jugarán un papel fundamental.

Ejemplo 8.9: Los cinescopios para televisión del fabricante A tienen una duración media de 6.5 años y una desviación estándar de 0.9 años; mientras que los del fabricante B tienen una duración media de 6.0 años y una desviación estándar de 0.8 años. ¿Cuál es la probabilidad de que una muestra aleatoria de 36 cinescopios del fabricante A tengan una duración media que sea al menos de 1 año más que la duración media de una muestra de 49 cinescopios del fabricante B?

Solución: Se nos da la siguiente información:

Población 1	Población 2
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

Si utilizamos el teorema 8.3, la distribución muestral de $\bar{X}_1 - \bar{X}_2$ será aproximadamente normal y tendrá una media y una desviación estándar de

$$\mu_{\bar{X}_1 - \bar{X}_2} = 6.5 - 6.0 = 0.5 \quad \text{y} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189.$$

La probabilidad de que la media de 36 cinescopios del fabricante A sea al menos 1 año mayor que la media de 49 cinescopios del fabricante B está dada por el área de la región sombreada de la figura 8.11. Con respecto al valor $\bar{x}_1 - \bar{x}_2 = 1.0$, encontramos que

$$z = \frac{1.0 - 0.5}{0.189} = 2.65,$$

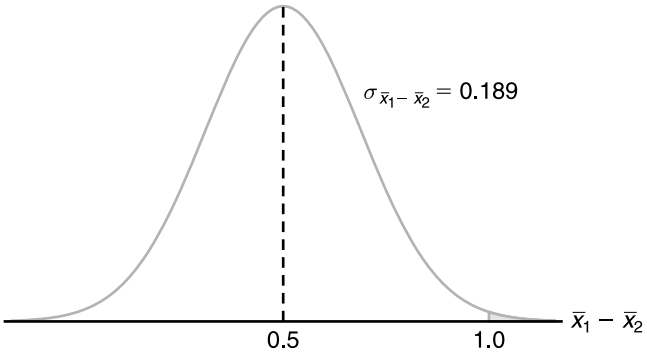


Figura 8.11: Área para el ejemplo 8.9.

y de aquí

$$P(\bar{X}_1 - \bar{X}_2 \geq 1.0) = P(Z > 2.65) = 1 - P(Z < 2.65)$$
$$= 1 - 0.9960 = 0.0040.$$

Más sobre la distribución muestral de medias: Aproximación normal a la binomial

En la sección 6.5, se discutió mucho acerca de la aproximación normal a la distribución binomial. Estaban dadas las condiciones sobre los parámetros n y p, para los cuales la distribución de una variable aleatoria binomial puede aproximarse mediante la distribución normal. Los ejemplos y los ejercicios reflejaron la importancia de la herramienta a la que nos referimos como la "aproximación normal". Resulta que el teorema del límite central arroja aún más luz sobre cómo y por qué funciona esta aproximación. Sabemos con certeza que una variable aleatoria binomial es el número X de éxitos en n pruebas independientes, donde el resultado de cada prueba es binario. En el capítulo 1 también vimos que la proporción calculada en un experimento así es un promedio de un conjunto de ceros y unos. De hecho, mientras que la proporción X/n es un promedio, X es la suma de este conjunto de ceros y unos, y tanto X como X/n son aproximadamente normales si n es suficientemente grande. Por supuesto, a partir de lo que aprendimos en el capítulo 6, hay condiciones de n y p que afectan la calidad de la aproximación; a saber, $np \geq 5$ y $nq \geq 5$.

Ejercicios

8.17 Si se extraen todas las muestras posibles de tamaño 16 de una población normal con media igual a 50 y desviación estándar igual a 5, ¿cuál es la probabilidad de que una media muestral $\bar{X}$ caiga en el intervalo que va de $\mu_{\bar{X}} - 1.9\sigma_{\bar{X}}$ a $\mu_{\bar{X}} - 0.4\sigma_{\bar{X}}$? Suponga que las medias muestrales se pueden medir con cualquier grado de precisión.

8.18 Dada la población uniforme discreta

$$f(x) = \begin{cases} \frac{1}{3}, & x = 2, 4, 6, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

encuentre la probabilidad de que una muestra aleatoria de tamaño 54, seleccionada con reemplazo, dé una media muestral mayor que 4.1 pero menor que 4.4. Suponga que las medias se miden al décimo más cercano.

8.19 Se fabrica cierto tipo de hilo con una resistencia a la tensión media de 78.3 kilogramos y una desviación estándar de 5.6 kilogramos. ¿Cómo cambia la varianza de la media muestral cuando el tamaño de la muestra

a) aumenta de 64 a 196?

b) disminuye de 784 a 49?

8.20 Si la desviación estándar de la media para la distribución muestral de muestras aleatorias de tamaño 36 de una población grande o infinita es 2, ¿qué tan grande debe ser el tamaño de la muestra si la desviación estándar se reduce a 1.2?

8.21 Una máquina de bebida gaseosa se ajusta de manera que la cantidad de bebida que sirve promedie 240 mililitros con una desviación estándar de 15 mililitros. La máquina se verifica periódicamente tomando una muestra de 40 bebidas y se calcula el contenido promedio. Si la media de las 40 bebidas es un valor dentro del intervalo $\mu_{\bar{X}} \pm 2\sigma_{\bar{X}}$, se piensa que la máquina opera satisfactoriamente; de otra forma, se ajusta. En la sección 8.4, el funcionario de la compañía encuentra que la media de 40 bebidas es $\bar{x} = 236$ mililitros y concluye que la máquina no necesita un ajuste. ¿Fue ésta una decisión razonable?

8.22 Las estaturas de 1000 estudiantes están distribuidas aproximadamente de forma normal con una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros. Si se extraen 200 muestras aleatorias de tamaño 25 de esta población y las medias se registran al décimo más cercano de centímetro, determine

a) la media y la desviación estándar de la distribución muestral de $\bar{X}$;

b) el número de las medias muestrales que caen entre 172.5 y 175.8 centímetros inclusive;

c) el número de medias muestrales que caen por debajo de 172.0 centímetros.

8.23 La variable aleatoria X, que representa el número de cerezas en una tarta, tiene la siguiente distribución de probabilidad:

x	4	5	6	7
$P(X = x)$	0.2	0.4	0.3	0.1

a) Encuentre la media μ y la varianza σ^2 de X.

b) Encuentre la media $\mu_{\bar{X}}$ y la varianza $\sigma^2_{\bar{X}}$ de la media $\bar{X}$ para muestras aleatorias de 36 tartas de cereza.

c) Encuentre la probabilidad de que el número promedio de cerezas en 36 tartas sea menor que 5.5.

8.24 Si cierta máquina fabrica resistencias eléctricas que tienen una resistencia media de 40 ohms y una desviación estándar de 2 ohms, ¿cuál es la probabilidad de que una muestra aleatoria de 36 de estas resistencias tenga una resistencia combinada de más de 1458 ohms?

8.25 La vida media de una máquina para elaborar pasta es de 7 años, con una desviación estándar de 1 año. Suponiendo que las vidas de estas máquinas siguen aproximadamente una distribución normal, encuentre

a) la probabilidad de que la vida media de una muestra aleatoria de 9 de estas máquinas caiga entre 6.4 y 7.2 años;

b) el valor de x a la derecha del cual caería el 15% de las medias calculadas de muestras aleatorias de tamaño 9.

8.26 El tiempo en que el cajero de un banco con servicio en el automóvil atiende a un cliente es una variable aleatoria con una media $\mu = 3.2$ minutos y una desviación estándar $\sigma = 1.6$ minutos. Si se observa una muestra aleatoria de 64 clientes, encuentre la probabilidad de que su tiempo medio con el cajero sea

a) a lo más 2.7 minutos;

b) más de 3.5 minutos;

c) al menos 3.2 minutos pero menos de 3.4 minutos.

8.27 En un proceso químico, la cantidad de cierto tipo de impurezas en el producto es difícil de controlar y por ello es una variable aleatoria. Se especula que la cantidad media de la población de impurezas es 0.20 gramos por gramo del producto. Se sabe que la desviación estándar es 0.1 gramos por gramo. Se realiza un experimento para aprender más con respecto a la especulación de que $\mu = 0.2$. El proceso se lleva a cabo 50 veces en un laboratorio y el promedio de la muestra $\bar{x}$ resulta ser 0.23 gramos por gramo. Comente sobre la especulación de que la cantidad media de impurezas es 0.20 gramos por gramo. Utilice el teorema del límite central en su respuesta.

8.28 Se toma una muestra aleatoria de tamaño 25 de una población normal que tiene una media de 80 y una desviación estándar de 5. Una segunda muestra aleatoria de tamaño 36 se toma de una población normal diferente que tiene una media de 75 y una desviación estándar de 3. Encuentre la probabilidad de que la media muestral calculada de las 25 mediciones exceda la media muestral calculada de las 36 mediciones por al menos 3.4 pero menos de 5.9. Suponga que las diferencias de las medias se miden al décimo más cercano.

8.29 La distribución de alturas de cierta raza de perros terrier tiene una altura media de 72 centímetros y una desviación estándar de 10 centímetros; en tanto que la distribución de alturas de cierta raza de poodles tiene una altura media de 28 centímetros con una desviación estándar de 5 centímetros. Suponiendo que las medias muestrales se pueden medir con cualquier grado de precisión, encuentre la probabilidad de que la media muestral para una muestra aleatoria de alturas de 64 terriers exceda la media muestral para una muestra aleatoria de alturas de 100 poodles a lo más en 44.2 centímetros.

8.30 La calificación media de estudiantes de primer año en un examen de aptitudes en cierta universidad es 540, con una desviación estándar de 50. ¿Cuál es la probabilidad de que dos grupos de estudiantes seleccionados al azar, que consisten en 32 y 50 estudiantes, respectivamente, difieran en sus calificaciones medias por

a) más de 20 puntos?

b) una cantidad entre 5 y 10 puntos?

Suponga que las medias se miden con cualquier grado de precisión.

8.31 Construya una gráfica de cuantiles para estos datos. Las duraciones, en horas, de cincuenta lámparas incandescentes de 40 watts a 110 volts enfriadas internamente tomadas de pruebas en condiciones adversas son:

919	1196	785	1126	936	918
1156	920	948	1067	1092	1162
1170	929	950	905	972	1035
1045	855	1195	1195	1340	1122
938	970	1237	956	1102	1157
978	832	1009	1157	1151	1009
765	958	902	1022	1333	811
1217	1085	896	958	1311	1037
702	923				

8.32 Considere el ejemplo 8.8 de la página 249. Suponga que se utilizan 18 especímenes para cada tipo de pintura en un experimento y que $\bar{x}_A - \bar{x}_B$, la diferencia real en el tiempo medio de secado resulta ser 1.0.

a) ¿Éste parece un resultado razonable si los tiempos medios de secado de las dos poblaciones en verdad son iguales? Utilice el resultado en la solución del ejemplo 8.8.

b) Si alguien hizo el experimento 10,000 veces bajo la condición de que $\mu_A = \mu_B$, ¿en cuántos de estos 10,000 experimentos habría una diferencia $\bar{x}_A - \bar{x}_B$ que fuera tan grande como (o más grande que) 1.0?

8.33 Dos máquinas diferentes de llenado de cajas se utilizan para llenar cajas de cereal en la línea de ensamble. La medición fundamental que está influida por tales máquinas es el peso del producto en las máquinas. Los ingenieros están bastante seguros de que la varianza en el peso del producto es $\sigma^2 = 1$ onza. Se realizan experimentos usando ambas máquinas con tamaños muestrales de 36 cada una. Los promedios muestrales para las máquinas A y B son $\bar{x}_A = 4.5$ onzas y $\bar{x}_B = 4.7$ onzas. Los ingenieros parecen sorprendidos de que los dos promedios maestrales para las máquinas de llenado sean tan diferentes.

a) Utilice el teorema del límite central para determinar

$$P(\bar{X}_B - \bar{X}_A \geq 0.2)$$

con la condición de que $\mu_A = \mu_B$.

b) ¿Parece como si los experimentos mencionados, de cualquier forma, apoyaran consistentemente una conjetura de que las dos medias de las poblaciones para las dos máquinas son diferentes? Explique utilizando su respuesta en el inciso *a*).

8.34 Construya una gráfica de cuantiles-cuantiles normales de estos datos. Los diámetros de 36 cabezas de remaches en 1/100 de pulgada son:

6.72	6.77	6.82	6.70	6.78	6.70	6.62
6.75	6.66	6.66	6.64	6.76	6.73	6.80
6.72	6.76	6.76	6.68	6.66	6.62	6.72
6.76	6.70	6.78	6.76	6.67	6.70	6.72
6.74	6.81	6.79	6.78	6.66	6.76	6.76
6.72						

8.35 El benceno es una sustancia química altamente tóxica para los seres humanos. Sin embargo, se le utiliza en la fabricación de medicamentos, tintes, en la industria del cuero y en la fabricación de recubrimientos. En cualquier proceso de producción en que participe el benceno, el agua en el resultado del proceso no debe exceder 7950 partes por millón (ppm) de benceno, de acuerdo con la regulación gubernamental. Para un proceso particular de interés, un fabricante recolectó la muestra de agua 25 veces de manera aleatoria y el promedio muestral $\bar{x}$ fue de 7960 ppm. A partir de los datos históricos, se sabe que la desviación estándar σ es 100 ppm.

a) ¿Cuál es la probabilidad de que el promedio muestral en este experimento exceda el límite gubernamental, si la media poblacional es igual al límite? Utilice el teorema del límite central.

b) La cifra $\bar{x} = 7960$ observada en este experimento ¿es firme evidencia de que la media poblacional para el proceso excede el límite gubernamental? Responda calculando

$$P(\bar{X} \geq 7960 \mid \mu = 7950).$$

Suponga que la distribución de la concentración de benceno es normal.

8.36 Dos aleaciones, A y B, se utilizan en la fabricación de cierto producto de acero. Se necesita diseñar un experimento para comparar las dos aleaciones en términos de la capacidad de carga máxima en toneladas, es decir, el máximo que pueden soportar sin romperse. Se sabe que las dos desviaciones estándar de la capacidad de carga son iguales a 5 toneladas cada una. Se realiza un experimento en el que se prueban 30 muestras de cada aleación (A y B), y los resultados son

$$\bar{x}_A = 49.5, \quad \bar{x}_B = 45.5; \qquad \bar{x}_A - \bar{x}_B = 4.$$

Los fabricantes de la aleación A están convencidos de que esta evidencia demuestra de forma concluyente que $\mu_A > \mu_B$ y que apoya sólidamente su aleación. Los fabricantes de la aleación B afirman que el experimento fácilmente podría haber dado $\bar{x}_A - \bar{x}_B = 4$ *incluso si* las dos medias poblacionales son iguales. En otras palabras, "¡los resultados no son concluyentes!"

a) Argumente que los fabricantes de la aleación B están equivocados. Para ello, calcule

$$P(\bar{X}_A - \bar{X}_B > 4 \mid \mu_A = \mu_B).$$

b) ¿Considera que estos datos apoyan fuertemente la aleación A?

8.37 Considere la situación del ejemplo 8.6 de la página 245. ¿Estos resultados lo hacen cuestionar la premisa de que $\mu = 800$ horas? Dé un resultado probabilístico que indique qué tan *raro* es un evento de que $\bar{X} \leq 775$ cuando $\mu = 800$. Por otro lado, ¿qué tan raro sería si μ fuera verdaderamente, digamos, 760 horas?

8.38 Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria a partir de una distribución que sólo pueda adoptar valores positivos. Utilice el teorema del límite central para argumentar que si n es suficientemente grande, entonces $Y = X_1 X_2 \cdots X_n$ tiene aproximadamente una distribución logarítmica normal.

8.6 Distribución muestral de S^2

En la sección anterior aprendimos acerca de la distribución de muestreo de $\bar{X}$. El teorema del límite central nos permitió utilizar el hecho de que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tiende a $N(0, 1)$ conforme crece el tamaño de la muestra. Los ejemplos 8.6 a 8.9 ilustran las aplicaciones del teorema del límite central. *Las distribuciones muestrales de estadísticos importantes* nos permiten conocer información sobre los parámetros. Por lo general, los parámetros son la contraparte del estadístico en cuestión. Si un ingeniero se interesa en la resistencia media de la población de cierto tipo de resistencia, la distribución muestral de $\bar{X}$ se explotará una vez que se reúna la información de la muestra. Por otro lado, si se estudia la variabilidad en la resistencia, claramente la distribución muestral de S^2 se utilizará para conocer la contraparte paramétrica, la varianza de la población σ^2.

Si se extrae una muestra aleatoria de tamaño n de una población normal con media μ y varianza σ^2, y se calcula la varianza muestral, obtenemos un valor del estadístico S^2. Procederemos a considerar la distribución del estadístico $(n-1)S^2/\sigma^2$.

Mediante la suma y la resta de la media muestral $\bar{X}$, es fácil ver que

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}[(X_i - \bar{X}) + (\bar{X} - \mu)]^2 \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X}) \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.\end{aligned}$$

Al dividir cada término de la igualdad entre σ^2 y sustituir $(n-1)S^2$ por $\sum_{i=1}^{n}(X_i - \bar{X})^2$, obtenemos

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2/n}.$$

Ahora, de acuerdo con el corolario del teorema 7.12 sabemos que

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

es una variable aleatoria chi cuadrada con n grados de libertad. Tenemos una variable aleatoria chi cuadrada con n grados de libertad dividida en dos componentes. El segundo término del lado derecho es Z^2, que es una variable aleatoria chi cuadrada con 1 grado de libertad y resulta que $(n-1)S^2/\sigma^2$ es una variable aleatoria chi cuadrada con $n-1$ grados de libertad. Formalizamos esto en el siguiente teorema.

Teorema 8.4: Si S^2 es la varianza de una muestra aleatoria de tamaño n que se toma de una población normal que tiene la varianza σ^2, entonces el estadístico

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v = n - 1$ grados de libertad.

Los valores de la variable aleatoria X^2 se calculan de cada muestra mediante la fórmula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}.$$

La probabilidad de que una muestra aleatoria produzca un valor χ^2 mayor que algún valor específico es igual al área bajo la curva a la derecha de este valor. Se acostumbra representar con χ^2_α el valor χ^2 por arriba del cual encontramos un área de α. Esto se ilustra mediante la región sombreada de la figura 8.12.

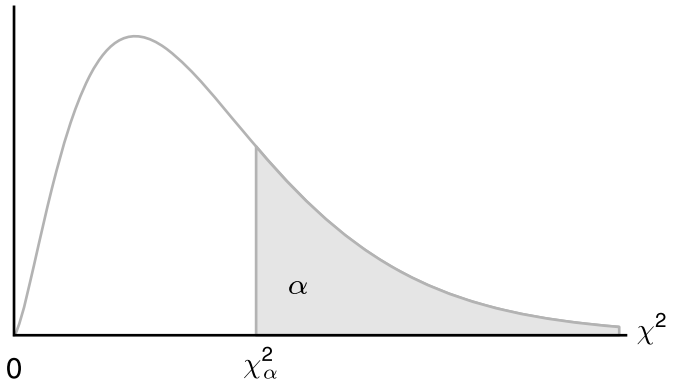


Figura 8.12: La distribución chi cuadrada.

La tabla A.5 da los valores de χ^2_α para diversos valores de α y v. Las áreas, α, son los encabezados de las columnas; los grados de libertad, v, se dan en la columna

izquierda, y las entradas de la tabla son los valores χ^2. De aquí, el valor χ^2 con 7 grados de libertad, que deja un área de 0.05 a la derecha, es $\chi^2_{0.05} = 14.067$. Debido a la falta de simetría, debemos usar también las tablas para encontrar $\chi^2_{0.95} = 2.167$ para $v = 7$.

Exactamente 95% de una distribución chi cuadrada yace entre $\chi^2_{0.975}$ y $\chi^2_{0.025}$. Un valor χ^2 que cae a la derecha de $\chi^2_{0.025}$ no es probable que ocurra, a menos que nuestro valor supuesto de σ^2 sea demasiado pequeño. Asimismo, un valor χ^2 que cae a la izquierda de $\chi^2_{0.975}$ es improbable, a menos que nuestro valor supuesto de σ^2 sea demasiado grande. En otras palabras, es posible tener un valor χ^2 a la izquierda de $\chi^2_{0.975}$ o a la derecha de $\chi^2_{0.025}$ cuando σ^2 es correcta; pero si esto debería ocurrir, es más probable que el valor supuesto de σ^2 esté equivocado.

Ejemplo 8.10: Un fabricante de baterías para automóvil garantiza que sus baterías durarán, en promedio, 3 años con una desviación estándar de 1 año. Si cinco de estas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años, ¿el fabricante aún está convencido de que sus baterías tienen una desviación estándar de 1 año? Suponga que la duración de la batería sigue una distribución normal.

Solución: Usando el teorema 8.1 encontramos primero la varianza de la muestra,

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815.$$

Entonces,

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

es un valor de una distribución chi cuadrada con 4 grados de libertad. Como 95% de los valores χ^2 con 4 grados de libertad caen entre 0.484 y 11.143, el valor calculado con $\sigma^2 = 1$ es razonable y, por lo tanto, el fabricante no tiene razón para sospechar que la desviación estándar sea diferente de 1 año.

Grados de libertad como medición de la información muestral

El lector puede obtener algunos conocimientos al considerar el teorema 8.4 y el corolario 7.1 en la sección 7.3. Sabemos que con las condiciones del teorema 7.12, es decir, una muestra aleatoria que se toma de una distribución normal, que la variable aleatoria

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

tiene una distribución χ^2 con n *grados de libertad*. Observe ahora que el teorema 8.4 indica que con las mismas condiciones del teorema 7.12, la variable aleatoria

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

tiene una distribución χ^2 con $n - 1$ *grados de libertad*. El lector debe recordar que el término *grados de libertad*, que se utiliza en este contexto idéntico, se estudió en el capítulo 1.

Como indicamos anteriormente, no se dará la demostración del teorema 8.4. Sin embargo, el lector puede ver el teorema 8.4 como una indicación de que cuando no se conoce μ y se considera la distribución de

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2},$$

hay **1 grado de libertad menos**, o se pierde un grado de libertad en la estimación de μ (es decir, cuando μ se reemplaza por $\bar{x}$). En otras palabras, hay n grados de libertad o *piezas de información* independientes en la muestra aleatoria de la distribución normal. Cuando los datos (los valores en la muestra) se utilizan para calcular la media, hay 1 grado de libertad menos en la información que se utiliza para estimar σ^2.

8.7 Distribución *t*

En la sección 8.5 se presentó la utilidad del teorema del límite central. Sus aplicaciones giran alrededor de las inferencias sobre una media de la población o la diferencia entre dos medias de población. El uso del teorema del límite central y la distribución normal es evidentemente útil en este contexto. Sin embargo, se supuso que se conoce la desviación estándar de la población. Esta suposición quizá sea razonable en situaciones donde el ingeniero esté bastante familiarizado con el sistema o proceso. No obstante, en muchos escenarios experimentales el conocimiento de σ ciertamente no es más razonable que el conocimiento de la media de la población μ. A menudo, de hecho, una estimación de σ la debe proporcionar la misma información muestral que produce el promedio muestral $\bar{x}$. Como resultado, un estadístico natural a considerar para tratar con las inferencias sobre μ es

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

puesto que S es el análogo de la muestra para σ. Si el tamaño de la muestra es pequeño, los valores de S^2 fluctúan de forma considerable de una muestra a otra (véase el ejercicio 8.45 de la página 265) y la distribución de T se desvía de forma apreciable de la de una distribución normal estándar.

Si el tamaño de la muestra es suficientemente grande, digamos $n \geq 30$, la distribución de T no difiere mucho de la normal estándar. Sin embargo, para $n < 30$, es útil tratar con la distribución exacta de T. Para desarrollar la distribución muestral de T supondremos que nuestra muestra aleatoria se seleccionó de una población normal. Podemos escribir, entonces,

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}},$$

donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tiene la distribución normal estándar y

$$V = \frac{(n-1)S^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v = n - 1$ grados de libertad. Al muestrear a partir de poblaciones normales, se puede mostrar que $\bar{X}$ y S^2 son independientes y, en consecuencia, también lo son Z y V. El siguiente teorema da la definición de una variable aleatoria T como una función de Z (normal estándar) y una χ^2. Para completar se da la función de densidad de la distribución t.

Teorema 8.5: Sea Z una variable aleatoria normal estándar y V una variable aleatoria chi cuadrada con v grados de libertad. Si Z y V son independientes, entonces, la distribución de la variable aleatoria T, donde

$$T = \frac{Z}{\sqrt{V/v}}$$

está dada por la función de densidad

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

Ésta se conoce como la **distribución *t*** con v grados de libertad.

De lo anterior y del teorema 8.5 tenemos el siguiente corolario:

Corolario 8.1: Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes que son todas normales con media μ y desviación estándar σ. Sea

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{y} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Entonces, la variable aleatoria $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ tiene una distribución t con $v = n - 1$ grados de libertad.

La distribución de probabilidad de T se publicó por primera vez en 1908 en un artículo de W. S. Gosset. En esa época, Gosset era empleado de una cervecería irlandesa que no autorizaba la publicación de investigaciones de sus empleados. Para evadir tal prohibición, publicó su trabajo en secreto bajo el nombre "Student". En consecuencia, la distribución de T normalmente se llama distribución t de Student, o simplemente distribución t. Para derivar la ecuación de esta distribución, Gosset supone que las muestras se seleccionan de una población normal. Aunque esto parecería una suposición muy restrictiva, se puede mostrar que las poblaciones no normales que poseen distribuciones en forma casi de campana aún proporcionan valores de T que se aproximan muy de cerca a la distribución t.

¿A qué se parece la distribución *t*?

La distribución de T es similar a la distribución de Z en que ambas son simétricas alrededor de una media de cero. Ambas distribuciones tienen forma de campana; pero la distribución t es más variable, debido al hecho de que los valores T dependen de

las fluctuaciones de dos cantidades, $\bar{X}$ y S^2; mientras que los valores Z dependen sólo de los cambios de $\bar{X}$ de una muestra a otra. La distribución de T difiere de la de Z en que la varianza de T depende del tamaño de la muestra n y siempre es mayor que 1. Únicamente cuando el tamaño de la muestran $n \to \infty$ las dos distribuciones serán las mismas. En la figura 8.13 mostramos la relación entre una distribución normal estándar ($v = \infty$) y las distribuciones t con 2 y 5 grados de libertad. Los puntos porcentuales de la distribución t se dan en la tabla A.4.

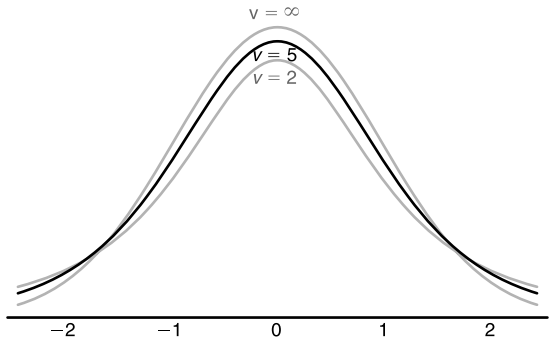


Figura 8.13: Curvas de la distribución t para $v = 2$, 5 y ∞.

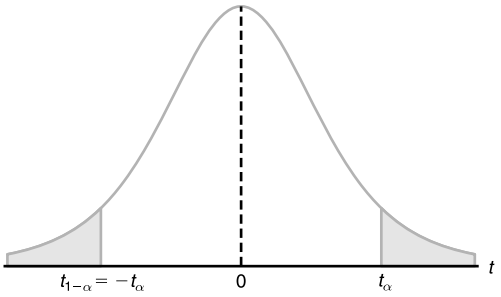


Figura 8.14: Propiedad de simetría de la distribución t.

Se acostumbra representar con t_α el valor t por arriba del cual encontramos un área igual a α. De aquí, el valor t con 10 grados de libertad que deja un área de 0.025 a la derecha es $t = 2.228$. Como la distribución t es simétrica alrededor de una media de cero, tenemos $t_{1-\alpha} = -t_\alpha$; es decir, el valor t que deja un área de $1 - \alpha$ a la derecha y, por lo tanto, un área de α a la izquierda es igual al valor t negativo que deja un área de α en la cola derecha de la distribución (véase la figura 8.14). Esto es, $t_{0.95} = -t_{0.05}$, $t_{0.99} = -t_{0.01}$, etcétera.

Ejemplo 8.11: El valor t con $v = 14$ grados de libertad que deja un área de 0.025 a la izquierda y, por lo tanto, un área de 0.975 a la derecha, es

$$t_{0.975} = -t_{0.025} = -2.145$$

Ejemplo 8.12: Encuentre $P(-t_{0.025} < T < t_{0.05})$.

Solución: Como $t_{0.05}$ deja un área de 0.05 a la derecha, y $-t_{0.025}$ deja un área de 0.025 a la izquierda, encontramos un área total de

$$1 - 0.05 - 0.025 = 0.925$$

entre $-t_{0.025}$ y $t_{0.05}$. De aquí,

$$P(-t_{0.025} < T < t_{0.05}) = 0.925.$$

Ejemplo 8.13: Encuentre k tal que $P(k < T < -1.761) = 0.045$, para una muestra aleatoria de tamaño 15 que se selecciona de una distribución normal y $\frac{\overline{X}-\mu}{s/\sqrt{n}}$.

Solución: De la tabla A.4 notamos que 1.761 corresponde a $t_{0.05}$ cuando $v = 14$. Por lo tanto, $-t_{0.05} = -1.761$. Como k en el enunciado de probabilidad original está a la izquierda de $-t_{0.05} = -1.761$, sea $k = -t_\alpha$. Entonces, de la figura 8.15, tenemos

$$0.045 = 0.05 - \alpha \quad \text{o} \quad \alpha = 0.005.$$

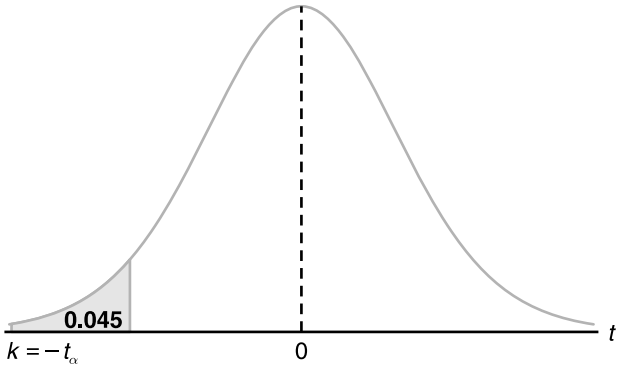


Figura 8.15: Valores t para el ejemplo 8.13.

Por ello, de la tabla A.4 con $v = 14$,

$$k = -t_{0.005} = -2.977 \text{ y } P(-2.977 < T < -1.761) = 0.045.$$

Exactamente 95% de los valores de una distribución t con $v = n - 1$ grados de libertad caen entre $-t_{0.025}$ y $t_{0.025}$. Por supuesto, hay otros valores t que contienen 95% de la distribución, como $-t_{0.02}$ y $t_{0.03}$, pero estos valores no aparecen en la tabla A.4 y, además, el intervalo más corto posible se obtiene al elegir valores t que dejen exactamente la misma área en las dos colas de nuestra distribución. Un valor t que caiga por debajo de $-t_{0.025}$ o por arriba de $t_{0.025}$ tendería a hacernos creer que ha ocurrido un evento muy raro, o que quizá nuestra suposición acerca de μ está equivocada. Si esto ocurre, tomaremos la última decisión y afirmaremos que nuestro valor supuesto de μ es erróneo. De hecho, un valor t que cae por debajo de $-t_{0.01}$ o por arriba de $t_{0.01}$ proporcionaría incluso evidencia más sólida de que nuestro valor supuesto de μ es bastante improbable. En el capítulo 10 se tratarán procedimientos generales para probar afirmaciones con respecto al valor del parámetro μ. El siguiente ejemplo ilustra un aspecto preliminar del fundamento de tales procedimientos.

Ejemplo 8.14: Un ingeniero químico afirma que el rendimiento medio de la población de cierto proceso en lotes es 500 gramos por milímetro de materia prima. Para verificar dicha afirmación muestrea 25 lotes cada mes. Si el valor t calculado cae entre $-t_{0.05}$ y $t_{0.05}$, queda satisfecho con su afirmación. ¿Qué conclusión debería obtener de una muestra que tiene una media $\bar{x} = 518$ gramos por milímetro y una desviación estándar muestral $s = 40$ gramos? Suponga que la distribución de rendimientos es aproximadamente normal.

Solución: De la tabla A.4 encontramos que $t_{0.05} = 1.711$ para 24 grados de libertad. Por lo tanto, el fabricante queda satisfecho con esta afirmación si una muestra de 25 lotes

produce un valor t entre -1.711 y 1.711. Si $\mu = 500$, entonces,

$$t = \frac{518 - 500}{40/\sqrt{25}} = 2.25,$$

un valor muy por arriba de 1.711. La probabilidad de obtener un valor t, con $v = 24$, igual o mayor que 2.25 es aproximadamente 0.02. Si μ ¿ 500, el valor de t calculado de la muestra sería más razonable. De aquí que sea probable que el fabricante concluya que el proceso produce un mejor producto del que pensaba.

¿Para qué se utiliza la distribución *t*?

La distribución t se usa de manera extensa en problemas que tienen que ver con inferencia acerca de la media de la población (como se ilustra en el ejemplo 8.14) o en problemas que implican muestras comparativas (es decir, en casos donde se trata de determinar si las medias de dos muestras son significativamente diferentes). El uso de la distribución se ampliará en los capítulos 9 a 12. El lector debería notar que el uso de la distribución t para el estadístico

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

requiere que $X_1, X_2, \ldots, X_n$ sea normal. El uso de la distribución t y la consideración del tamaño de la muestra no se relacionan con el teorema del límite central. El uso de la distribución normal estándar en vez de T para $n \geq 30$ solamente implica, en este caso, que S es un estimador suficientemente bueno de σ. En los siguientes capítulos la distribución t encuentra un uso extenso.

8.8 Distribución *F*

Motivamos la distribución t en parte sobre la base de la aplicación a problemas en los que hay muestreo comparativo (es decir, comparación entre dos medias muestrales). Algunos de nuestros ejemplos en los siguientes capítulos brindarán el formalismo. Un ingeniero químico reúne datos de dos catalizadores. Un biólogo colecta datos sobre dos medias de crecimiento. Un químico reúne datos sobre dos métodos de recubrimiento de material para prevenir la corrosión. Aunque es de interés que la información muestral arroje luz sobre dos medias de poblaciones, es frecuente el caso en que una comparación, en algún sentido, de la variabilidad sea igualmente importante, si no es que más. La distribución F encuentra enorme aplicación en la comparación de varianzas muestrales. Las aplicaciones de la distribución F se encuentran en problemas que implican dos o más muestras.

El estadístico F se define como la razón de dos variables aleatorias chi cuadradas independientes, dividida cada una entre su número de grados de libertad. De aquí, podemos escribir

$$F = \frac{U/v_1}{V/v_2},$$

donde U y V son variables aleatorias independientes que tienen distribuciones chi cuadradas con v_1 y v_2 grados de libertad, respectivamente. Estableceremos ahora la distribución muestral de F.

Teorema 8.6:

Sean U y V dos variables aleatorias independientes que tienen distribuciones chi cuadradas con v_1 y v_2 grados de libertad, respectivamente. Entonces, la distribución de la variable aleatoria $F = \frac{U/v_1}{V/v_2}$ está dada por la densidad

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0, \\ 0, & f \leq 0. \end{cases}$$

Ésta se conoce como la **distribución *F*** con v_1 y v_2 grados de libertad (g.l).

De nuevo haremos un uso considerable de la variable aleatoria F en capítulos posteriores. Sin embargo, no se utilizará la función de densidad y se dará sólo como complemento. La curva de la distribución F depende no sólo de los dos parámetros v_1 y v_2 sino también del orden en el que se establecen. Una vez que se dan estos dos valores, podemos identificar la curva. En la figura 8.16 se presentan distribuciones F típicas.

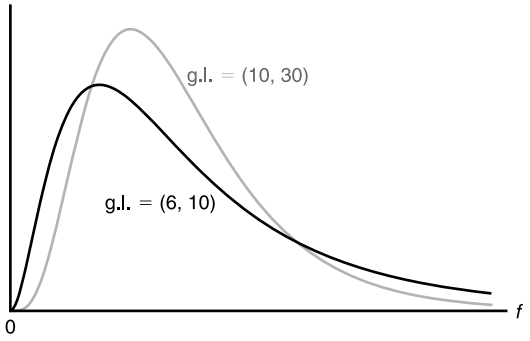


Figura 8.16: Distribuciones F típicas.

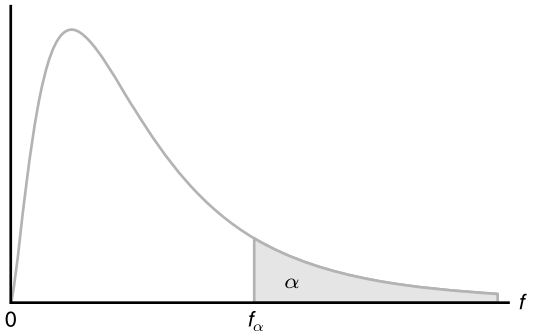


Figura 8.17: Ilustración de la f_α para la distribución F.

Sea f_α el valor f por arriba del cual encontramos un área igual a α. Esto se ilustra mediante la región sombreada de la figura 8.17. La tabla A.6 da valores de f_α sólo para $\alpha = 0.05$ y $\alpha = 0.01$ para varias combinaciones de los grados de libertad v_1 y v_2. De aquí, el valor f con 6 y 10 grados de libertad, que deja un área de 0.05 a la derecha, es $f_{0.05} = 3.22$. Por medio del siguiente teorema, la tabla A.6 también se puede utilizar para encontrar valores de $f_{0.95}$ y $f_{0.99}$. La demostración se deja al lector.

Teorema 8.7:

Al escribir $f_\alpha\,(v_1, v_2)$ para f_α con v_1 y v_2 grados de libertad, obtenemos

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$

Así, el valor f con 6 y 10 grados de libertad, que deja un área de 0.95 a la derecha, es

$$f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10, 6)} = \frac{1}{4.06} = 0.246.$$

La distribución *F* con dos varianzas muestrales

Suponga que las muestras aleatorias de tamaño n_1 y n_2 se seleccionan de dos poblaciones normales con varianzas σ_1^2 y σ_2^2, respectivamente. Del teorema 8.4, sabemos que

$$X_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \quad \text{y} \quad X_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

son variables aleatorias que tienen distribuciones chi cuadradas con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Además, como las muestras se seleccionan al azar, tratamos con variables aleatorias independientes y, entonces, usando el teorema 8.6 con $X_1^2 = U$ y $X_2^2 = V$, obtenemos el siguiente resultado.

Teorema 8.8: Si S_1^2 y S_2^2 son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones normales con varianzas σ_1^2 y σ_2^2, respectivamente, entonces,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

¿Para qué se utiliza la distribución *F*?

Contestamos esta pregunta, parcialmente, al inicio de esta sección. La distribución F se usa en situaciones de dos muestras para realizar inferencias acerca de las varianzas de población, lo cual implica la aplicación del resultado del teorema 8.8. Sin embargo, la distribución F se aplica a muchos otros tipos de problemas en los cuales están relacionadas las varianzas muestrales. De hecho, la distribución F se llama *distribución de razón de varianzas*. Como ilustración, considere el ejemplo 8.8. Se compararon dos pinturas, A y B, con respecto a su tiempo medio de secado. La distribución normal se aplica bien (suponiendo que se conocen σ_A y σ_B). Sin embargo, considere que hay tres tipos de pinturas para comparar, digamos A, B y C. Queremos determinar si las medias de las poblaciones son equivalentes. Suponga, de hecho, que importante información resumida del experimento es la siguiente:

Pintura	Media muestral	Varianza muestral	Tamaño muestral
A	$\bar{X}_A = 4.5$	$s_A^2 = 0.20$	10
B	$\bar{X}_B = 5.5$	$s_B^2 = 0.14$	10
C	$\bar{X}_C = 6.5$	$s_C^2 = 0.11$	10

El problema se centra alrededor de si los promedios muestrales ($\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$) están suficientemente alejados o no. La implicación de "suficientemente alejados" resulta muy importante. Parecería razonable que si la variabilidad entre los promedios muestrales es mayor que lo que se esperaría por casualidad, los datos no apoyan la conclusión de que $\mu_A = \mu_B = \mu_C$. Que estos promedios muestrales pudieran ocurrir por casualidad depende de la *variabilidad dentro de las muestras*, como cuantifican s_A^2, s_B^2 y s_C^2. La noción de los componentes importantes de la variabilidad se ve

mejor utilizando algunas gráficas sencillas. Considere la gráfica de los datos originales de las muestras A, B y C, que se presenta en la figura 8.18. Estos datos podrían generar con facilidad la información de resumen anterior.

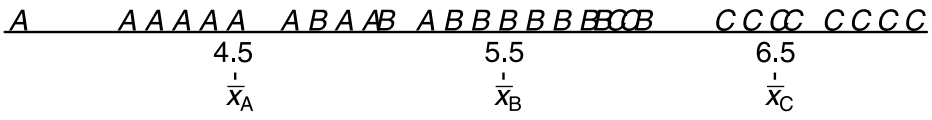


Figura 8.18: Datos de tres muestras diferentes.

Parece evidente que los datos vienen de distribuciones con diferentes medias de población, aunque hay alguna superposición entre las muestras. Un análisis que incluya todos los datos intentaría determinar si la variabilidad entre los promedios muestrales y la variabilidad dentro de las muestras podría haber ocurrido conjuntamente *si, de hecho, las poblaciones tienen una media común.* Observe que la clave para este análisis se centra alrededor de las dos siguientes fuentes de variabilidad.

1. Variabilidad dentro de las muestras (entre observaciones en muestras distintas).

2. Variabilidad entre muestras (entre promedios muestrales).

Claramente, si la variabilidad en **1.** es considerablemente mayor que la de **2.**, habrá una superposición considerable en los datos de la muestra y una señal de que los datos podrían provenir de una distribución común. Se encuentra un ejemplo en el conjunto de datos que contienen tres muestras, y que se presenta en la figura 8.19. Por otro lado, es muy improbable que los datos de una distribución con una media común puedan tener variabilidad entre promedios muestrales que sea considerablemente mayor que la variabilidad dentro de las muestras.

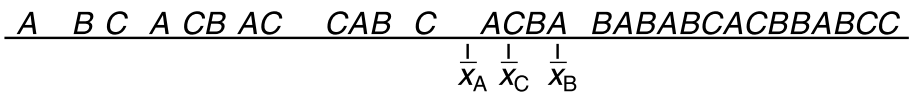


Figura 8.19: Datos que fácilmente provendrían de la misma población.

Las fuentes de variabilidad en **1.** y **2.** anteriores generan importantes razones de *varianzas muestrales* y las razones se utilizan junto con la distribución F. El procedimiento general implicado se llama **análisis de varianza**. Es interesante que en el ejemplo de la pintura que se describe aquí tratamos con inferencias acerca de tres medias de población; aunque se utilizan dos fuentes de variabilidad. No proporcionaremos detalles aquí; pero en los capítulos 13, 14 y 15 utilizaremos de manera extensa el análisis de varianza y, desde luego, la distribución F juega un papel importante.

Ejercicios

8.39 Para una distribución chi cuadrada encuentre
a) $\chi^2_{0.025}$ cuando $v = 15$;
b) $\chi^2_{0.01}$ cuando $v = 7$;
c) $\chi^2_{0.05}$ cuando $v = 24$.

8.40 Para una distribución chi cuadrada encuentre lo siguiente:
a) $\chi^2_{0.005}$ cuando $v = 5$;
b) $\chi^2_{0.05}$ cuando $v = 19$;
c) $\chi^2_{0.01}$ cuando $v = 12$.

8.41 Para una distribución chi cuadrada encuentre χ^2_α tal que
a) $P(X^2 > \chi^2_\alpha) = 0.99$ cuando $v = 4$;
b) $P(X^2 > \chi^2_\alpha) = 0.025$ cuando $v = 19$;
c) $P(37.652 < X^2 < \chi^2_\alpha) = 0.045$ cuando $v = 25$.

8.42 Para una distribución chi cuadrada encuentre χ^2_α tal que
a) $P(X^2 > \chi^2_\alpha) = 0.01$ cuando $v = 21$;
b) $P(X^2 < \chi^2_\alpha) = 0.95$ cuando $v = 6$;
c) $P(\chi^2_\alpha < X^2 < 23.209) = 0.015$ cuando $v = 10$.

8.43 Encuentre la probabilidad de que una muestra aleatoria de 25 observaciones, de una población normal con varianza $\sigma^2 = 6$, tenga una varianza s^2
a) mayor que 9.1;
b) entre 3.462 y 10.745.
Suponga que las varianzas muestrales son mediciones continuas.

8.44 Las calificaciones de un examen de colocación que se aplicó a estudiantes de primer año de una universidad durante los últimos cinco años están distribuidas aproximadamente de forma normal con una media $\mu = 74$ y una varianza $\sigma^2 = 8$. ¿Consideraría aún que $\sigma^2 = 8$ es un valor válido de la varianza si una muestra aleatoria de 20 estudiantes, quienes realizan tal examen de colocación este año, obtienen un valor de $s^2 = 20$?

8.45 Muestre que la varianza de S^2 para muestras aleatorias de tamaño n de una población normal disminuye conforme n se hace grande. [*Sugerencia:* primero encuentre la varianza de $(n-1)S^2/\sigma^2$.]

8.46 *a*) Encuentre $t_{0.025}$ cuando $v = 14$.
b) Encuentre $-t_{0.10}$ cuando $v = 10$.
c) Encuentre $t_{0.995}$ cuando $v = 7$.

8.47 *a*) Encuentre $P(T < 2.365)$ cuando $v = 7$.
b) Encuentre $P(T > 1.318)$ cuando $v = 24$.
c) Encuentre $P(-1.356 < T < 2.179)$ cuando $v = 12$.
d) Encuentre $P(T > -2.567)$ cuando $v = 17$.

8.48 *a*) Encuentre $P(-t_{0.005} < T < t_{0.01})$ para $v = 20$.
b) Encuentre $P(T > -t_{0.025})$.

8.49 Dada una muestra aleatoria de tamaño 24 de una distribución normal, encuentre k tal que
a) $P(-2.069 < T < k) = 0.965$;
b) $P(k < T < 2.807) = 0.095$;
c) $P(-k < T < k) = 0.90$.

8.50 Una empresa manufacturera afirma que las baterías que utiliza en sus juegos electrónicos duran un promedio de 30 horas. Para mantener este promedio, se prueban 16 baterías cada mes. Si el valor t que se calcula cae entre $-t_{0.025}$ y $t_{0.025}$, la empresa queda satisfecha con su afirmación. ¿Qué conclusiones debería obtener la empresa de una muestra que tiene una media $\bar{x} = 27.5$ horas y una desviación estándar $s = 5$ horas? Suponga que la distribución de las duraciones de las baterías es aproximadamente normal.

8.51 Una población normal con varianza desconocida tiene una media de 20. ¿Se tiene posibilidad de obtener una muestra aleatoria de tamaño 9 de esta población con una media de 24 y una desviación estándar de 4.1? Si no, ¿qué conclusión obtendría?

8.52 Un fabricante de cierta marca de barras de cereal bajo en grasa afirma que su contenido promedio de grasa saturada es 0.5 gramos. En una muestra aleatoria de 8 barras de cereal de esta marca, el contenido de grasa saturada fue 0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4 y 0.2. ¿Estaría de acuerdo con la afirmación? Suponga una distribución normal.

8.53 Para una distribución F encuentre:
a) $f_{0.05}$ con $v_1 = 7$ y $v_2 = 15$;
b) $f_{0.05}$ con $v_1 = 15$ y $v_2 = 7$;
c) $f_{0.01}$ con $v_1 = 24$ y $v_2 = 19$;
d) $f_{0.95}$ con $v_1 = 19$ y $v_2 = 24$;
e) $f_{0.99}$ con $v_1 = 28$ y $v_2 = 12$.

8.54 Pruebas de resistencia a la tracción sobre 10 cables conductores soldados para un dispositivo semiconductor dan los siguientes resultados en libras fuerza requeridas para romper la unión:

19.8	12.7	13.2	16.9	10.6
18.8	11.1	14.3	17.0	12.5

Otro conjunto de ocho cables conductores se probó después del encapsulado para determinar si la resistencia a la tracción había aumentado debido al encapsulado del dispositivo, con los siguientes resultados:

24.9 22.8 23.6 22.1 20.4 21.6 21.8 22.5

Haga comentarios sobre la evidencia disponible con respecto a la igualdad de las dos varianzas de la población.

8.55 Considere las siguientes mediciones de la capacidad de producción de calor del carbón producido por dos minas (en millones de calorías por tonelada):

Mina 1: 8260 8130 8350 8070 8340

Mina 2: 7950 7890 7900 8140 7920 7840

¿Se puede concluir que son iguales las dos varianzas de población?

Ejercicios de repaso

8.56 Considere los datos que se muestran en el ejercicio 1.20 de la página 29. Construya una gráfica de caja y extensión, y comente la naturaleza de la muestra. Calcule la media muestral y la desviación estándar de la muestra.

8.57 Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes que tienen distribuciones exponenciales idénticas con parámetro θ, muestre que la función de densidad de la variable aleatoria $Y = X_1 + X_2 \cdots + X_n$ es la de una distribución gamma con parámetros $\alpha = n$ y $\beta = \theta$.

8.58 Al probar el monóxido de carbono en cierta marca de cigarrillos, los datos, en miligramos por cigarrillo, se codificaron al restar 12 de cada observación. Utilice los resultados del ejercicio 8.14 de la página 235 para encontrar la desviación estándar del contenido de monóxido de carbono de una muestra aleatoria de 15 cigarrillos de esta marca, si las mediciones codificadas son 3.8, -0.9, 5.4, 4.5, 5.2, 5.6, 2.7, -0.1, -0.3, -1.7, 5.7, 3.3, 4.4, -0.5, y 1.9.

8.59 Si S_1^2 y S_2^2 representan las varianzas de muestras aleatorias independientes de tamaño $n_1 = 8$ y $n_2 = 12$, tomadas de poblaciones normales con varianzas iguales, encuentre $P(S_1^2/S_2^2 < 4.89)$.

8.60 Una muestra aleatoria de 5 presidentes de bancos indicó sueldos anuales de \$395,000, \$521,000, \$483.000, \$479,000 y \$510,000. Encuentre la varianza de este conjunto.

8.61 Si el número de huracanes que azotan cierta área del este de Estados Unidos por año es una variable aleatoria que tiene una distribución de Poisson con $\mu = 6$, encuentre la probabilidad de que esta área sea azotada por

a) exactamente 15 huracanes en 2 años;

b) a lo más 9 huracanes en 2 años.

8.62 Una compañía de taxis prueba una muestra aleatoria de 10 neumáticos radiales con bandas tensoras de acero de cierta marca y registra los siguientes desgastes de la banda: 48,000, 53,000, 45,000, 61,000, 59,000, 56,000, 63,000, 49,000, 53,000, y 54,000 kilómetros. Utilice los resultados del ejercicio 8.14 en la página 235 para encontrar la desviación estándar de este conjunto de datos al dividir primero cada observación entre 1000 y después restar 55.

8.63 Considere los datos del ejercicio 1.19 de la página 28. Construya una gráfica de caja y extensión. Haga comentarios. Calcule la media muestral y la desviación estándar muestral.

8.64 Si S_1^2 y S_2^2 representan las varianzas de muestras aleatorias independientes de tamaño $n_1 = 25$ y $n_2 = 31$, tomadas de poblaciones normales con varianzas $\sigma_1^2 = 10$ y $\sigma_2^2 = 15$, respectivamente, encuentre

$$P(S_1^2/S_2^2 > 1.26).$$

8.65 Considere el ejercicio 1.21 de la página 29. Comente cualquier valor extremo.

8.66 Considere el ejercicio de repaso 8.56. Comente cualquier valor extremo en los datos.

8.67 La resistencia a la rotura X de cierto remache utilizado en el motor de una máquina tiene una media de 5000 psi y una desviación estándar de 400 psi. Se toma una muestra aleatoria de 36 remaches. Considere la distribución de $\bar{X}$, la resistencia a la rotura de la media muestral.

a) ¿cuál es la probabilidad de que la media de la muestra caiga entre 4800 psi y 5200 psi?

b) ¿Qué muestra n sería necesaria para tener

$$P(4900 < \bar{X} < 5100) = 0.99?$$

8.68 Considere la situación del ejercicio de repaso 8.62. Si la población de la cual se tomó la muestra tiene una media $\mu = 53{,}000$ kilómetros, ¿aquí la información de la muestra parece apoyar esa afirmación? En su respuesta calcule

$$t = \frac{\bar{x} - 53{,}000}{s/\sqrt{10}}$$

y determine, consultado la tabla A.4 (con 9 g.l.), si el valor t calculado es razonable ¿o parece ser un evento raro?

8.69 Dos propulsores de combustible sólido distintos, tipo A y tipo B, se consideran en una actividad del programa espacial. Las velocidades de combustión en el propulsor son fundamentales. Se toman muestras aleatorias de 20 especímenes de los dos propulsores con medias de muestra dadas por 20.5 cm/s para el propulsor A y 24.50 cm/s para el propulsor B. Por lo general, se supone que la variabilidad en la velocidad de combustión es aproximadamente la misma para los dos propulsores y está dada por una desviación estándar de 5 cm/s. Suponga que la velocidad de combustión para cada propulsor es aproximadamente normal y, por lo tanto, utilice el teorema del límite central. Nada se sabe acerca de las dos velocidades de combustión medias de la poblacion y se espera que este experimento podría arrojar alguna luz.

a) Si, de hecho, $\mu_A = \mu_B$, ¿cuál será $P(\bar{X}_B - \bar{X}_A \geq 4.0)$?

b) Utilice su respuesta en el inciso *a*) para arrojar alguna luz sobre la proposición de que $\mu_A = \mu_B$.

8.70 La concentración de un ingrediente activo en el producto de una reacción química está influido fuertemente por el catalizador que se usa en la reacción. Se considera que cuando se utiliza el catalizador A, la concentración media de la población excede 65%. Se sabe que la desviación estándar es $\sigma = 5\%$. Una muestra de productos tomada de 30 experimentos independientes da la concentración promedio de $\bar{x}_A = 64.5\%$.

a) ¿Esta información muestral con una concentración promedio de $\bar{x}_A = 64.5\%$ ofrece información inquietante de que quizá μ_A no sea 65%, sino menos de 65%? Apoye su respuesta con una declaración de probabilidad.

b) Suponga que se realiza un experimento similar utilizando otro catalizador, uno B. Aún se supone que la desviación estándar σ es 5% y $\bar{x}_B$ resulta ser de 70%. Comente sobre si la información del catalizador B parece dar o no evidencia sólida que sugiera que μ_B es en realidad mayor que μ_A. Apoye su respuesta calculando

$$P(\bar{X}_B - \bar{X}_A \geq 5.5 \mid \mu_B = \mu_A).$$

c) Con la condición de que $\mu_A = \mu_B = 65\%$, determine la distribución aproximada de los siguientes cuantiles (con la media y la varianza de cada uno). Utilice el teorema del límite central.

i) $\bar{X}_B$;
ii) $\bar{X}_A - \bar{X}_B$;
iii) $\frac{\bar{X}_A - \bar{X}_B}{\sigma\sqrt{2/30}}$.

8.71 De la información del ejercicio de repaso 8.70, calcule (suponiendo $\mu_B = 65\%$)

$$P(\bar{X}_B \geq 70).$$

8.72 Dada una variable aleatoria normal X con media 20 y varianza 9, y una muestra aleatoria de tamaño n tomada de la distribución, ¿qué tamaño de la muestra n se necesita para que

$$P(19.9 \leq \bar{X} \leq 20.1) = 0.95?$$

8.73 En el capítulo 9 se estudiará con detenimiento el concepto de **estimación de parámetros**. Suponga que X es una variable aleatoria con media μ y varianza $\sigma^2 = 1.0$. Además, suponga que se toma una muestra aleatoria de tamaño n y que $\bar{x}$ se utiliza como un *estimado* de μ. Cuando se toman los datos y se mide la media de la muestra, deseamos que ésta esté dentro de 0.05 unidades de media real con probabilidad de 0.99. Es decir, aquí queremos que haya una buena probabilidad de que la $\bar{x}$ calculada de la muestra esté "muy cercana" a la media de la población (¡dondequiera que se encuentre!), de manera que deseamos

$$P(|\bar{X} - \mu| > 0.05) = 0.99.$$

¿Qué tamaño de muestra se requiere?

8.74 Suponga que una máquina de llenado se utiliza para llenar envases de cartón con un producto líquido. La especificación que es estrictamente indispensable cumplir para el llenado de la máquina es 9 ± 1.5 onzas. Si cualquier envase cartón se produce fuera de tales límites de peso, el proveedor lo considera como defectuoso. Se espera que al menos 99% de los envases de cartón cumplirá tales especificaciones. Con las condiciones $\mu = 9$ y $\sigma = 1$, ¿qué proporción de envases de cartón del proceso están defectuosos? Si se hacen cambios para reducir la variabilidad, ¿cuánto debe reducirse σ para cumplir con las especificaciones con una probabilidad de 0.99? Suponga una distribución normal para el peso.

8.75 Considere la situación del ejercicio de repaso 8.74. Suponga que se realiza un esfuerzo de calidad considerable para "apretar" la variabilidad del sistema. Siguiendo el esfuerzo, se toma una muestra aleatoria de tamaño 40 de la nueva línea de ensamble y la varianza de la muestra $s^2 = 0.188$ onzas^2. ¿Tenemos evidencia numérica sólida de que σ^2 se redujo por debajo de 1.0? Considere la probabilidad

$$P(S^2 \leq 0.188 \mid \sigma^2 = 1.0),$$

y dé una conclusión.

8.9 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

El teorema del límite central es una de las herramientas más poderosa de la estadística, y aún cuando este capítulo es relativamente corto, contiene una riqueza de información fundamental que se relaciona con las herramientas que se utilizan para lograr el equilibrio de todo el texto.

La noción de una distribución muestral es uno de los conceptos fundamentales más importantes de la estadística y, en este punto, el estudiante debería obtener por su estudio una comprensión clara antes de continuar más allá de este capítulo. Todos los capítulos que siguen continuarán utilizando ampliamente las distribuciones muestrales. Suponga que se quiere utilizar el estadístico $\bar{X}$ para obtener inferencias acerca de la media de la población μ, lo cual se hace utilizando el valor observado $\bar{x}$ de una sola muestra de tamaño n. Luego cualquier inferencia que se realice debe lograrse tomando en cuenta no sólo el valor único, sino más bien la estructura teórica o la **distribución de todos los valores $\bar{x}$ que podrían observarse de las muestras de tamaño n**. Así se presentó el término *distribución muestrales*, que es la base del teorema del límite central. Las distribuciones t, χ^2 y F también se utilizan en el contexto de las distribuciones muestrales. Por ejemplo, la distribución t, que se ilustra en la figura 8.13, representa la estructura que ocurre si se forman todos los valores de $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, donde $\bar{x}$ y s se toman de las muestras de tamaño n de una distribución $n(x; \mu, \sigma)$. Se pueden hacer comentarios similares con respecto de χ^2 y F, y el lector no debería olvidar que la información muestral que forma el estadístico para todas estas distribuciones es la normal. De manera que **se puede afirmar que donde haya una t, F, χ^2 la fuente era una muestra de una distribución normal**.

Puede parecer que las tres distribuciones antes descritas se presentaron de una forma bastante autosuficiente, sin una indicación de a qué se refieren. No obstante, aparecerán en la resolución de problemas prácticos a lo largo del texto.

Entonces, hay cuestiones que se deben tener presentes para evitar que haya confusión respecto de tales distribuciones muestrales fundamentales:

i. No se puede usar el teorema del límite central a menos que se conozca σ. Cuando no se conoce σ, se debería remplazar con s, la desviación estándar de la muestra, para usar el teorema del límite central.

ii. El estadístico T **no** es un resultado del teorema del límite central y $x_1, x_2, \ldots, x_n$ deben provenir de una distribución $n(x; \mu, \sigma)$ para que $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ sea una distribución t, y s, desde luego, es tan sólo una estimación de σ.

iii. En tanto que la noción de **grados de libertad** es nueva en este punto, el concepto debería ser muy intuitivo, ya que es razonable que la naturaleza de la distribución de S y también t debería depender de la cantidad de información en la muestra $x_1, x_2, \ldots, x_n$.

Capítulo 9

Problemas de estimación de una y dos muestras

9.1 Introducción

En los capítulos anteriores resaltamos las propiedades del muestreo de la media y de la varianza muestrales. También destacamos las representaciones de datos en varias formas. El propósito de tales presentaciones es establecer las bases que permitan a los estadísticos extraer conclusiones acerca de los parámetros de la población a partir de datos experimentales. Por ejemplo, el teorema del límite central brinda información sobre la distribución de la media muestral $\bar{X}$. La distribución incluye la media de la población μ. Así, cualesquiera conclusiones que se obtengan con respecto a μ, a partir de un promedio muestral observado, deben depender del conocimiento de su distribución muestral. Comentarios similares se podrían aplicar a S^2 y σ^2. Resulta claro que cualesquiera conclusiones que extraigamos acerca de la varianza de una distribución normal probablemente implicarían la distribución muestral de S^2.

En este capítulo comenzaremos por esbozar de manera formal el propósito de la inferencia estadística. Seguimos con la presentación del problema de la **estimación de los parámetros de la población**. Restringiremos nuestros desarrollos formales de los procedimientos de estimación específicos a problemas que tengan una y dos muestras.

9.2 Inferencia estadística

En el capítulo 1 presentamos la filosofía general de la inferencia estadística formal. La teoría de la **inferencia estadística** consiste en aquellos métodos por los que se realizan inferencias o generalizaciones acerca de una población. La tendencia actual es la distinción entre el **método clásico** de estimación de un parámetro de la población, mediante el cual las inferencias se basan estrictamente en información obtenida de una muestra aleatoria seleccionada de la población, y el **método bayesiano**, que utiliza el conocimiento subjetivo previo sobre la distribución de probabilidad de los parámetros desconocidos junto con la información que proporcionan los datos de la muestra. A lo largo de la mayoría de este capítulo utilizaremos los métodos clásicos para estimar los parámetros de la población desconocidos, como la media,

proporción y la varianza, calculando estadísticos de muestras aleatorias y aplicando la teoría de las distribuciones muestrales, mucho de lo cual se estudió en el capítulo 8. La estimación bayesiana se examina en el capítulo 18.

La inferencia estadística se puede dividir en dos áreas principales: **estimación** y **pruebas de hipótesis**. Trataremos estas dos áreas por separado: en este capítulo veremos la teoría y las aplicaciones de la estimación; y la prueba de hipótesis, en el capítulo 10. Para distinguir claramente entre ambas áreas, considere los siguientes ejemplos. Un candidato a un cargo público puede desear estimar la verdadera proporción de votantes que lo favorecerán mediante la obtención de las opiniones de una muestra aleatoria de 100 de ellos. La fracción de votantes en la muestra que favorecerán al candidato se podría utilizar como una estimación de la verdadera proporción en la población de votantes. El conocimiento de la distribución muestral de una proporción nos permite establecer el grado de precisión de nuestra estimación. Este problema cae en el área de la estimación.

Considere ahora el caso en que alguien está interesado en averiguar si la marca A de cera para piso es más resistente al desgaste que la marca B. Se puede establecer la hipótesis de que la marca A es mejor que la marca B y, después de la prueba adecuada, aceptar o rechazar dicha hipótesis. En este ejemplo no intentamos estimar un parámetro, sino que en realidad tratamos de llegar a una decisión correcta acerca de una hipótesis preestablecida. Una vez más, dependemos de la teoría del muestreo y del uso de datos que nos proporcionen alguna medición de la precisión de nuestra decisión.

9.3 Métodos clásicos de estimación

Una **estimación puntual** de algún parámetro de la población θ es un solo valor $\hat{\theta}$ de un estadístico $\hat{\Theta}$. Por ejemplo, el valor $\bar{x}$ del estadístico $\bar{X}$, que se calcula a partir de una muestra de tamaño n, es una estimación puntual del parámetro poblacional μ. De manera similar, $\hat{p} = x/n$ es una estimación puntual de la verdadera proporción p para un experimento binomial.

No se espera que un estimador realice la estimación del parámetro poblacional sin error. No esperamos que $\bar{X}$ estime μ exactamente, sino que en realidad esperamos que no esté muy alejado. Para una muestra específica es posible obtener un estimado más cercano de μ utilizando la mediana de la muestra $\tilde{X}$ como un estimador. Considere, por ejemplo, una muestra que consista en los valores 2, 5 y 11 de una población cuya media es 4, pero que supuestamente se desconoce. Estimaríamos μ como $\bar{x} = 6$, con la media muestral como nuestra estimación, o $\tilde{x} = 5$, con la mediana muestral como nuestra estimación. En este caso, el estimador $\tilde{X}$ produce una estimación más cercana al verdadero parámetro que la del estimador $\bar{X}$. Por otro lado, si nuestra muestra aleatoria contiene los valores 2, 6 y 7, entonces $\bar{x} = 5$ y $\tilde{x} = 6$, por lo que $\bar{X}$ ahora es el mejor estimador. Al no conocer el valor real de μ, debemos decidir de antemano si se utiliza $\bar{X}$ o $\tilde{X}$ como nuestro estimador.

Estimador insesgado

¿Cuáles son las propiedades deseables de una "buena" función de decisión que influirían sobre nosotros para elegir un estimador en vez de otro? Sea $\hat{\Theta}$ un estimador cuyo valor $\hat{\theta}$ es una estimación puntual de algún parámetro poblacional desconocido B. Ciertamente, desearíamos que la distribución muestral de $\hat{\Theta}$ tuviera una media igual al parámetro estimado. Se dice que un estimador que posee esta propiedad es **insesgado**.

Definición 9.1: Se dice que un estadístico $\hat{\Theta}$ es un estimador **insesgado** del parámetro θ si

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Ejemplo 9.1: Muestre que S^2 es un estimador insesgado del parámetro σ^2.
Solución: Escribamos

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.\end{aligned}$$

Entonces,

$$\begin{aligned}E(S^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right] \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma_{X_i}^2 - n\sigma_{\bar{X}}^2\right).\end{aligned}$$

Sin embargo,

$$\sigma_{X_i}^2 = \sigma^2 \text{ para } i = 1, 2, \ldots, n, \quad \text{y} \quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Por lo tanto,

$$E(S^2) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2.$$

Aunque S^2 es un estimador insesgado de σ^2, S, es, por otro lado, un estimador sesgado de σ y el sesgo se vuelve insignificante en muestras grandes. Este ejemplo ilustra **por qué dividimos entre $n - 1$** en vez de n cuando se estima la varianza.

Varianza de un estimador puntual

Si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores insesgados del mismo parámetro poblacional θ, elegiríamos el estimador cuya distribución muestral tuviera la menor varianza. De aquí, si $\sigma_{\hat{\theta}_1}^2 < \sigma_{\hat{\theta}_2}^2$, decimos que $\hat{\Theta}_1$ es un **estimador más eficaz** de θ que $\hat{\Theta}_2$.

Definición 9.2: Si consideramos todos los posibles estimadores insesgados de algún parámetro θ, el de menor varianza se llama **estimador más eficaz** de θ.

En la figura 9.1 ilustramos las distribuciones muestrales de 3 estimadores diferentes $\hat{\Theta}_1$, $\hat{\Theta}_2$ y $\hat{\Theta}_3$, todos para θ. Resulta claro que sólo $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son insesgados, pues

sus distribuciones están centradas en θ. El estimador $\hat{\Theta}_1$ tiene una varianza menor que $\hat{\Theta}_2$ y, por lo tanto, es más eficaz. De aquí que nuestra elección de un estimador de θ, entre los tres que se consideran, Θería $\hat{\Theta}_1$.

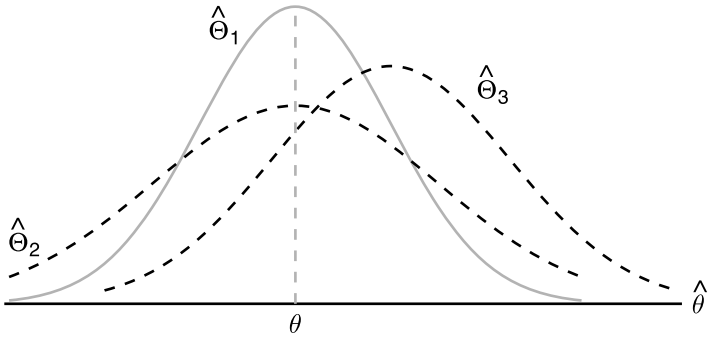


Figura 9.1: Distribuciones muestrales de estimadores diferentes de θ.

Para poblaciones normales se puede mostrar que $\bar{X}$ y $\tilde{X}$ son estimadores insesgados de la media poblacional μ, pero la varianza de $\bar{X}$ es más pequeña que la varianza de $\tilde{X}$. De esta manera las estimaciones $\bar{x}$ y $\tilde{x}$ serán, en promedio, iguales a la media poblacional μ, aunque es probable que $\bar{x}$ esté más cerca de μ para una muestra dada y, por ello, $\bar{X}$ es más eficaz que $\tilde{X}$.

La noción de una estimación por intervalo

Es improbable que incluso el estimador insesgado más eficaz estime con exactitud el parámetro poblacional. Es cierto que nuestra precisión aumenta con muestras grandes; pero no hay razón por la cual deberíamos esperar que una estimación **puntual** de una muestra dada sea exactamente igual al parámetro poblacional que se supone estima. Hay muchas situaciones en que es preferible determinar un intervalo dentro del cual esperaríamos encontrar el valor del parámetro. Tal intervalo se llama **estimación por intervalo**.

Estimación por intervalo

La estimación por intervalo de un parámetro poblacional θ es un intervalo de la forma $\hat{\theta}_L < \theta < \hat{\theta}_U$, donde $\hat{\theta}_L$ y $\hat{\theta}_U$ dependen del valor del estadístico $\hat{\Theta}$ para una muestra específica, y también de la distribución de muestreo de $\hat{\Theta}$. Así, una muestra aleatoria de calificaciones verbales SAT para estudiantes universitarios de una clase de primer año produciría un intervalo de 530 a 550, dentro del cual esperamos encontrar el promedio real de todas las calificaciones verbales del SAT para tal clase. Los valores de los puntos extremos, 530 y 550, dependerán de la media muestral calculada $\bar{x}$ y de la distribución de muestreo de $\bar{X}$. A medida de que se incrementa el tamaño de la muestra, sabemos que $\sigma^2_{\bar{X}} = \sigma^2/n$ disminuye y, en consecuencia, es probable que nuestra estimación esté más cercana al parámetro μ, lo cual tiene como resultado un intervalo más pequeño. De esta manera, el intervalo estimado indica,

por su longitud, la precisión de la estimación puntual. Un ingeniero obtendrá una idea de la proporción de la población de artículos defectuosos al tomar una muestra y calcular la *proporción de defectuosos de la muestra*. No obstante, una estimación por intervalo podría resultar más informativa.

Interpretación de la estimación por intervalo

Como muestras distintas, por lo general, darán valores diferentes de $\hat{\Theta}$ y, por lo tanto, valores diferentes de $\hat{\theta}_L$ y $\hat{\theta}_U$, estos puntos extremos del intervalo son valores de las variables aleatorias correspondientes $\hat{\Theta}_L$ y $\hat{\Theta}_U$. De la distribución muestral de $\hat{\Theta}$ seremos capaces de determinar $\hat{\Theta}_L$ y $\hat{\Theta}_U$, de manera que $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U)$ sea igual a algún valor fraccional positivo que queramos especificar. Si, por ejemplo, encontramos $\hat{\Theta}_L$ y $\hat{\Theta}_U$ tales que

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha$$

para $0 < \alpha < 1$, tenemos entonces una probabilidad de $1 - \alpha$ de seleccionar una variable aleatoria que produzca un intervalo que contenga θ. El intervalo $\hat{\theta}_L < \theta < \hat{\theta}_U$, que se calcula a partir de la muestra seleccionada, se llama entonces **intervalo de confianza** de $(1 - \alpha)100\%$, la fracción $1 - \alpha$ se llama **coeficiente de confianza** o **grado de confianza**, y los extremos, $\hat{\theta}_L$ y $\hat{\theta}_U$, se denominan **límites de confianza** inferior y superior. Así, cuando $\alpha = 0.05$, tenemos un intervalo de confianza de 95%, y cuando $\alpha = 0.01$ obtenemos un intervalo de confianza más amplio de 99%. Cuanto más amplio sea el intervalo de confianza, tendremos mayor confianza de que el intervalo dado contenga el parámetro desconocido. Desde luego, es mejor tener una confianza de 95% de que la vida promedio de cierto transistor de televisor está entre 6 y 7 años, que tener una confianza de 99% de que esté entre 3 y 10 años. Idealmente, preferimos un intervalo corto con un grado de confianza alto. Algunas veces las restricciones en el tamaño de nuestra muestra nos impiden tener intervalos cortos sin sacrificar algo de nuestro grado de confianza.

En las secciones que siguen estudiaremos las nociones de estimación puntual y por intervalo, donde cada sección representa un caso especial diferente. El lector debería notar que mientras la estimación puntual y por intervalo representan diferentes aproximaciones para obtener información con respecto a un parámetro, también se relacionan en el sentido de que los estimadores del intervalo de confianza se basan en estimadores puntuales. En la siguiente sección, por ejemplo, veremos que $\bar{X}$ es un estimador puntual de μ muy razonable. Como resultado, el importante estimador del intervalo de confianza de μ depende del conocimiento de la distribución muestral de $\bar{X}$.

En la siguiente sección empezamos con el caso más sencillo de un intervalo de confianza, donde el escenario es simple e incluso irreal. Nos interesa estimar una media de la población μ y se desconoce σ. Evidentemente, si se desconoce μ es bastante improbable que se conozca σ. Cualquier información histórica que produzca información suficiente para permitir la suposición de que se conoce σ probablemente habría ofrecido información similar acerca de μ. A pesar de este argumento, iniciamos con este caso porque los conceptos y, de hecho, los mecanismos resultantes asociados con la estimación del intervalo de confianza permanecen constantes cuando se presenten situaciones más realistas en la sección 9.4 y en las siguientes.

9.4 Una sola muestra: Estimación de la media

La distribución muestral de $\bar{X}$ está centrada en μ y en la mayoría de las aplicaciones la varianza es más pequeña que la de cualesquiera otros estimadores de μ. Así, la media muestral $\bar{x}$ se utilizará como una estimación puntual para la media de la población μ. Recuerde que $\sigma^2_{\bar{X}} = \sigma^2/n$, por lo que una muestra grande dará un valor de $\bar{X}$ que provenga de una distribución muestral con varianza pequeña. De aquí que $\bar{x}$ es probablemente una estimación muy precisa de μ cuando n es grande.

Consideremos ahora la estimación por intervalo de μ. Si nuestra muestra se selecciona a partir de una población normal o, a falta de ésta, si n es suficientemente grande, podemos establecer un intervalo de confianza para μ al considerar la distribución muestral de $\bar{X}$.

De acuerdo con el teorema del límite central, podemos esperar que la distribución muestral de $\bar{X}$ esté distribuida de forma aproximadamente normal con media $\mu_{\bar{X}} = \mu$ y desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Al escribir $z_{\alpha/2}$ para el valor z por arriba del cual encontramos un área de $\alpha/2$, de la figura 9.2 podemos ver que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Por ello,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

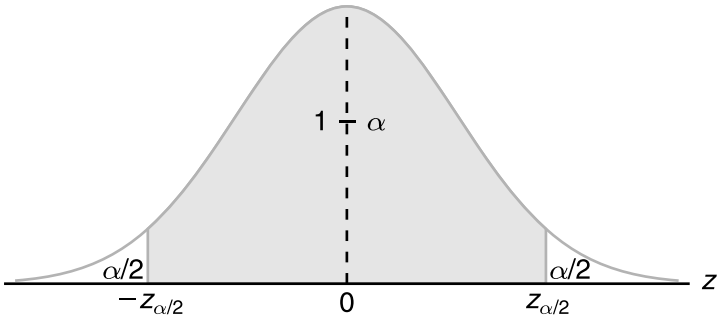


Figura 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Al multiplicar cada término en la desigualdad por $\sigma/\sqrt{n}$, y después restar $\bar{X}$ de cada término y multiplicar por -1 (para invertir el sentido de las desigualdades), obtenemos

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Se selecciona una muestra aleatoria de tamaño n de una población cuya varianza σ^2 se conoce y se calcula la media $\bar{x}$ para obtener el siguiente intervalo de confianza de $(1 - \alpha)100\%$. Es importante enfatizar que recurrimos al teorema del límite central.

Como resultado es importante destacar las condiciones para las aplicaciones que siguen.

Intervalo de confianza de μ; con σ conocida

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una población con varianza σ^2 conocida, un intervalo de confianza de $(1-\alpha)100\%$ para μ está dado por

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

donde $z_{\alpha/2}$ es el valor z que deja un área de $\alpha/2$ a la derecha.

Para muestras pequeñas que se seleccionan de poblaciones no normales, no podemos esperar que nuestro grado de confianza sea preciso. Sin embargo, para muestras de tamaño $n \geq 30$, donde la forma de las distribuciones no esté muy sesgada, la teoría de muestreo garantiza buenos resultados.

Claramente, los valores de las variables aleatorias $\hat{\Theta}_L$ y $\hat{\Theta}_U$, que se definen en la sección 9.3, son los límites de confianza

$$\hat{\theta}_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{y} \quad \hat{\theta}_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Muestras diferentes darán valores diferentes de $\bar{x}$ y, por lo tanto, producirán diferentes estimaciones por intervalo del parámetro μ como se observa en la figura 9.3. Los puntos circulares al centro de cada intervalo indican la posición de la estimación puntual $\bar{x}$ para cada muestra aleatoria. Se ve que la mayoría de los intervalos contienen μ, pero no en todos los casos. Note que todos los intervalos son del mismo ancho, pues esto depende sólo de la elección de $z_{\alpha/2}$ una vez que se determina $\bar{x}$. Cuanto más grande sea el valor de $z_{\alpha/2}$ que elijamos, más anchos haremos todos los intervalos, y podremos tener más confianza en que la muestra particular que se seleccione producirá un intervalo que contenga el parámetro desconocido μ.

Ejemplo 9.2: Se encuentra que la concentración promedio de zinc que se obtiene a partir de una muestra de mediciones de zinc en 36 sitios diferentes es 2.6 gramos por mililitro. Encuentre los intervalos de confianza de 95 y 99% para la concentración media de zinc en el río. Suponga que la desviación estándar de la población es 0.3.

Solución: La estimación puntual de μ es $\bar{x} = 2.6$. El valor z, que deja un área de 0.025 a la derecha y, por lo tanto, un área de 0.975 a la izquierda, es $z_{0.025} = 1.96$ (tabla A.3). De aquí que el intervalo de confianza de 95% sea

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right),$$

que se reduce a $2.50 < \mu < 2.70$. Para encontrar un intervalo de confianza de 99%, encontramos el valor z que deja un área de 0.005 a la derecha y de 0.995 a la izquierda. Por lo tanto, usando la tabla A.3 de nuevo, $z_{0.005} = 2.575$ y el intervalo de confianza de 99% es

$$2.6 - (2.575)\frac{0.3}{\sqrt{36}} < \mu < 2.6 + (2.575)\frac{0.3}{\sqrt{36}},$$

o simplemente

$$2.47 < \mu < 2.73.$$

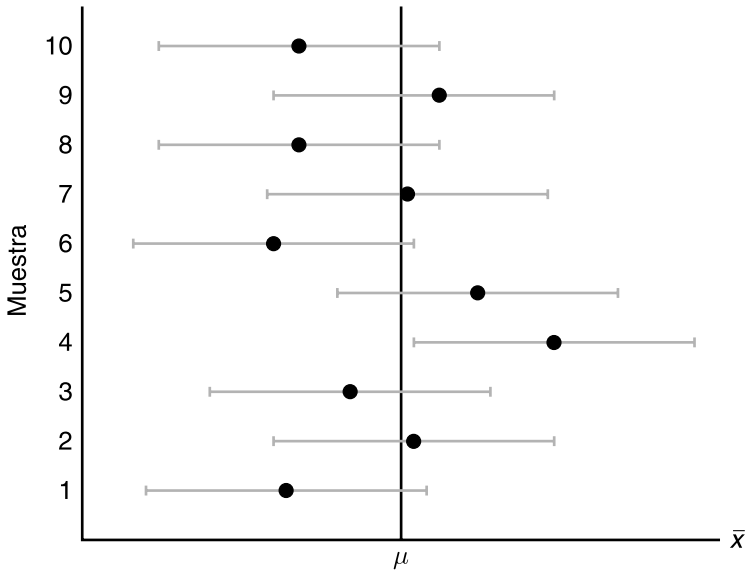


Figura 9.3: Estimaciones por intervalo de μ para muestras diferentes.

Vemos ahora que se requiere un intervalo más grande para estimar μ con un mayor grado de confianza.

El intervalo de confianza de $(1-\alpha)100\%$ ofrece una estimación de la precisión de nuestra estimación puntual. Si μ es realmente el valor central del intervalo, entonces $\bar{x}$ estima μ sin error. La mayoría de las veces, sin embargo, $\bar{x}$ no será exactamente igual a μ y la estimación puntual será errónea. La magnitud de este error será el valor absoluto de la diferencia entre μ y $\bar{x}$, de manera que podemos tener $(1-\alpha)100\%$ de confianza de que esta diferencia no excederá $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Esto se puede ver con facilidad si elaboramos un diagrama de un intervalo de confianza hipotético, como el de la figura 9.4.

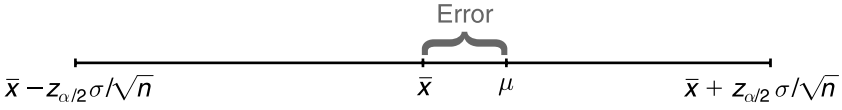


Figura 9.4: Error en la estimación de μ mediante $\bar{x}$.

Teorema 9.1: Si se utiliza $\bar{x}$ como una estimación de μ, podemos tener una confianza de $(1-\alpha)100\%$ de que el error no excederá $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

En el ejemplo 9.2 tenemos el 95% de confianza de que la media muestral $\bar{x} = 2.6$ difiere de la media real μ en una cantidad menor que 0.1, y 99% de confianza de que la diferencia es menor que 0.13.

Con frecuencia, queremos saber qué tan grande necesita ser una muestra para asegurarnos de que el error al estimar μ será menor que una cantidad específica e. Por el teorema 9.1, esto significa que debemos elegir n de manera que $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = e$. Al resolver esta ecuación se obtiene la siguiente fórmula para n.

Teorema 9.2: Si $\bar{x}$ se usa como estimación de μ, podemos tener $(1 - \alpha)100\%$ de confianza de que el error no excederá una cantidad específica e cuando el tamaño de la muestra sea

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Cuando se resuelve para el tamaño de la muestra n, todos los valores fraccionales se redondean al siguiente número entero. Si se sigue este principio, podemos estar seguros de que nuestro grado de confianza nunca caerá por debajo de $(1 - \alpha)100\%$.

Estrictamente hablando, la fórmula del teorema 9.2 se aplica sólo si conocemos la varianza de la población de la cual seleccionamos nuestra muestra. A falta de tal información, podríamos tomar una muestra preliminar de tamaño $n \geq 30$ que proporcione una estimación de σ. Después, usando s como aproximación para σ en el teorema 9.2 podemos determinar aproximadamente cuántas observaciones se necesitan para brindar el grado de precisión que se desea.

Ejemplo 9.3: ¿Qué tan grande se requiere una muestra en el ejemplo 9.2 si queremos tener 95% de confianza de que nuestra estimación de μ difiera por menos de 0.05?

Solución: La desviación estándar de la población es $\sigma = 0.3$. Entonces, por el teorema 9.2,

$$n = \left[\frac{(1.96)(0.3)}{0.05}\right]^2 = 138.3.$$

Por lo tanto, podemos tener una confianza de 95% de que una muestra aleatoria de tamaño 139 proporcionará una estimación $\bar{x}$ que difiera de μ por una cantidad menor que 0.05.

Límites de confianza unilaterales

Los intervalos de confianza y los límites de confianza resultantes analizados hasta ahora son en realidad de bilaterales (esto es, se dan tanto el límite superior como el inferior). Sin embargo, hay muchas aplicaciones en que sólo se requiere un límite. Por ejemplo, si la medida de interés es la resistencia a la tensión, el ingeniero recibe más información del límite inferior solamente. Este límite comunica el escenario del "peor caso". Por otro lado, si para la medida un valor relativamente grande de μ no es provechoso o deseable, entonces resultará de interés el límite de confianza superior. Un ejemplo sería el caso en el que se necesita hacer inferencias acerca de la composición media de mercurio en un río. Un límite superior sería muy informativo en este caso.

Los límites de confianza unilaterales se desarrollan de la misma forma que los intervalos bilaterales. Sin embargo, la fuente es un enunciado de probabilidad unilateral que utiliza el teorema del límite central

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha.$$

Entonces, es posible manipular el enunciado de probabilidad de forma muy similar a como se hizo anteriormente, para obtener

$$P(\mu > \bar{X} - z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Una manipulación similar de $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > -z_\alpha\right) = 1 - \alpha$ da

$$P(\mu < \bar{X} + z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Como resultado, se obtienen los siguientes límites unilaterales superior e inferior.

Límites de confianza unilaterales en μ; σ desconocida

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n a partir de una población con varianza σ^2, los límites de confianza unilaterales de $(1 - \alpha)100\%$ para μ están dados por

límite unilateral superior: $\bar{x} + z_\alpha \sigma/\sqrt{n}$;

límite unilateral inferior: $\bar{x} - z_\alpha \sigma/\sqrt{n}$.

Ejemplo 9.4: En un estudio de pruebas psicológicas, se seleccionan al azar 25 sujetos y se mide su tiempo de reacción, en segundos, ante un experimento particular. La experiencia pasada sugiere que la varianza en el tiempo de reacción a estos tipos de estímulos es de 4 s^2 y que el tiempo de reacción es aproximadamente normal. El tiempo promedio para los sujetos fue de 6.2 segundos. Dé un límite superior de 95% para el tiempo medio de reacción.

Solución: El límite superior de 95% está dado por

$$\bar{x} + z_\alpha \sigma/\sqrt{n} = 6.2 + (1.645)\sqrt{4/25} = 6.2 + 0.658$$
$$= 6.858 \text{ segundos.}$$

De esta forma, tenemos un nivel de confianza de 95% de que la reacción media sea menor que 6.858 segundos.

El caso de σ desconocida

Con frecuencia intentamos estimar la media de una población cuando se desconoce la varianza. El lector debería recordar que, en el capítulo 8, aprendimos que si tenemos una muestra aleatoria a partir de una *distribución normal*, entonces la variable aleatoria

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tiene una distribución t de Student con $n - 1$ grados de libertad. Aquí S es la desviación estándar de la muestra. En esta situación en que se desconoce σ se puede utilizar T para construir un intervalo de confianza de μ. El procedimiento es el mismo que cuando se conoce σ excepto en que σ se reemplaza con S y la distribución normal estándar se reemplaza con la distribución t. Con referencia a la figura 9.5, podemos asegurar que

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

donde $t_{\alpha/2}$ es el valor t con $n - 1$ grados de libertad, arriba del cual encontramos un área de $\alpha/2$. Debido a la simetría, un área igual de $\alpha/2$ caerá a la izquierda de $-t_{\alpha/2}$. Y sustituyendo por T, escribimos

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Al multiplicar cada término en la desigualdad por $S/\sqrt{n}$, y después restar $\bar{X}$ de cada término y multiplicar por −l, obtenemos

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Para nuestra muestra aleatoria particular de tamaño n, se calculan la media $\bar{x}$ y la desviación estándar s y se obtiene el siguiente intervalo de confianza de $(1 - \alpha)100\%$ para μ.

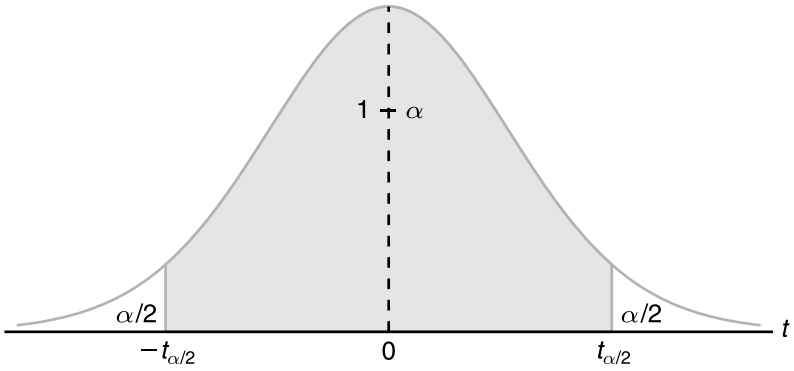


Figura 9.5: $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$.

Intervalo de confianza de μ; con σ desconocida

Si $\bar{x}$ y s son la media y la desviación estándar de una muestra aleatoria de una población con varianza σ^2 desconocida, un intervalo de confianza de $(1 - \alpha)100\%$ para μ es

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad que deja un área de $\alpha/2$ a la derecha.

Hacemos una distinción entre los casos de σ conocida y σ desconocida al calcular las estimaciones del intervalo de confianza. Deberíamos resaltar que para el caso de σ conocida se utiliza el teorema del límite central; mientras que para σ desconocida usamos la distribución muestral de la variable aleatoria T. Sin embargo, el uso de la distribución t se basa en la premisa de que el muestreo se realiza de una distribución normal. En tanto que la distribución tenga aproximadamente forma de campana, los intervalos de confianza se pueden calcular cuando σ^2 se desconoce utilizando la distribución t y se esperarían muy buenos resultados.

Los límites de confianza unilaterales calculados para μ con σ desconocida son como el lector esperaría; a saber:

$$\bar{x} + t_{\alpha}\frac{s}{\sqrt{n}} \qquad \text{y} \qquad \bar{x} - t_{\alpha}\frac{s}{\sqrt{n}}.$$

Son, respectivamente los límites superior e inferior de $(1 - \alpha)100\%$. Aquí t_α es el valor t que tiene un área α a la derecha.

Concepto de intervalo de confianza de una muestra grande

Con mucha frecuencia los estadísticos recomiendan que aun cuando no se pueda suponer la normalidad, con σ desconocida y $n \geq 30$, s puede reemplazar a σ y utilizar el intervalo de confianza

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Por lo general, éste se denomina como un *intervalo de confianza de muestra grande*. La justificación yace sólo en la presunción de que con una muestra tan grande como 30 y la distribución de la población no sesgada, s estará muy cerca de la σ real y, de esta manera, el teorema del límite central continúa siendo válido. Se debería destacar que esto es sólo una aproximación y que la calidad de este enfoque mejora conforme el tamaño de la muestra crece más.

Ejemplo 9.5: El contenido de 7 contenedores similares de ácido sulfúrico es de 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, y 9.6 litros. Encuentre un intervalo de confianza de 95% para la media de todos los contenedores, si se supone una distribución aproximadamente normal.

Solución: La media muestral y la desviación estándar para los datos dados son

$$\bar{x} = 10.0 \qquad \text{y} \qquad s = 0.283.$$

Con la tabla A.4, encontramos $t_{0.025} = 2.447$ para $v = 6$ grados de libertad. De aquí, el intervalo de confianza de 95% para μ es

$$10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right),$$

que se reduce a $9.74 < \mu < 10.26$.

9.5 Error estándar de una estimación puntual

Hacemos una distinción bastante clara entre los objetivos de las estimaciones puntuales y las estimaciones del intervalo de confianza. Las primeras proporcionan un solo número que se extrae de un conjunto de datos experimentales, y las últimas brindan un intervalo, *dados los datos experimentales*, que sea razonable para el parámetro; es decir, $(1 - \alpha)100\%$ de tales intervalos que se calculan "cubren" el parámetro.

Estas dos aproximaciones a la estimación se relacionan entre sí. El "hilo común" es la distribución muestral del estimador puntual. Considere, por ejemplo, el estimador $\bar{X}$ de μ con σ conocida. Indicamos antes que una medida de la calidad de un estimador insesgado es su varianza. La varianza de $\bar{X}$ es

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.$$

De esta forma la desviación estándar de $\bar{X}$ *o error estándar* de $\bar{X}$ es $\sigma/\sqrt{n}$. De manera simple, el error estándar de un estimador es su desviación estándar. Para el caso de $\bar{X}$, el límite de confianza que se calcula

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \text{se escribe como} \quad \bar{x} \pm z_{\alpha/2}\ \text{e.e.}(\bar{x}),$$

donde "e.e." es el error estándar. El punto importante a considerar es que el ancho del intervalo de confianza de μ depende de la calidad del estimador puntual a través de su error estándar. En el caso donde σ se desconoce y el muestreo es sobre una distribución normal, s reemplaza a σ y se incluye el *error estándar estimado* $s/\sqrt{n}$. De esta forma, los límites de confianza de μ son

Límites de confianza de μ para σ desconocida

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm t_{\alpha/2}\text{e.e.}(\bar{x})$$

De nuevo, el intervalo de confianza *no es mejor* (en términos de anchura) *que la calidad de la estimación puntual*, en este caso a través de su error estándar estimado. A menudo el software computacional se refiere a los errores estándar estimados simplemente como "errores estándar".

Conforme nos movemos hacia intervalos de confianza más complejos, prevalece el concepto de que el ancho de los intervalos de confianza se vuelve más corto conforme mejora la calidad de la estimación puntual correspondiente; aunque ello no siempre sea tan sencillo como aquí se ilustra. Se puede argumentar que un intervalo de confianza es tan sólo una ampliación de la estimación puntual para tomar en cuenta la precisión de ésta.

9.6 Intervalos de predicción

Las estimaciones puntual y por intervalos de la media en las secciones 9.4 y 9.5 ofrecen excelente información sobre el parámetro desconocido μ de una distribución normal, o de una distribución no normal a partir de la cual se toma una muestra grande. Algunas veces, aparte de la media de la población, quizás el experimentador esté interesado en predecir los posibles **valores de una observación futura**. Por ejemplo, en un caso de control de calidad, el experimentador necesitaría utilizar los datos observados para predecir una nueva observación. Un proceso que produce una pieza de metal podría evaluarse con base en si la pieza cumple con las especificaciones del proceso en cuanto a resistencia a la tensión. En ciertas ocasiones tal vez un cliente se interese en comprar una **sola pieza**. En este caso, un intervalo de confianza de la resistencia media a la tensión no cubre el requerimiento. El cliente requiere un enunciado con respecto a la incertidumbre de una **sola observación**. El tipo de requerimiento se satisface muy bien mediante la construcción de un **intervalo de predicción**.

Es bastante sencillo obtener un intervalo de predicción para las situaciones que hemos considerado hasta el momento. Suponga que la muestra aleatoria se tomó de una población normal con media desconocida μ y varianza conocida σ^2. Un estimador puntual natural de una nueva observación es $\bar{X}$. De la sección 8.5 se sabe que la varianza de $\bar{X}$ es σ^2/n. Sin embargo, para predecir una nueva observación, no únicamente necesitamos dar cuenta de la variación debida a la estimación de la media, sino también deberíamos dar cuenta de la **variación de una observación futura**. Por la suposición sabemos que la varianza del error aleatorio en una nueva observación es σ^2. El desarrollo de un intervalo de predicción se representa mejor empezando con una variable aleatoria normal $x_0 - \bar{x}$, donde x_0 es la nueva observación y $\bar{x}$ se toma de la muestra. Como x_0 y $\bar{x}$ son independientes, sabemos que

$$z = \frac{x_0 - \bar{x}}{\sqrt{\sigma^2 + \sigma^2/n}} = \frac{x_0 - \bar{x}}{\sigma\sqrt{1 + 1/n}}$$

es $n(z; 0, 1)$. Como resultado, si utilizamos el enunciado de probabilidad

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

con el estadístico z anterior, colocamos x_0 en el centro del enunciado de probabilidad, tenemos que el siguiente evento ocurre con probabilidad $1 - \alpha$:

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n}.$$

Como resultado, el intervalo de predicción calculado se formaliza como sigue:

Intervalo de predicción para una observación futura: σ conocida

Para una distribución normal de mediciones con media desconocida μ y varianza conocida σ^2, un **intervalo de predicción** de $(1 - \alpha)100\%$ de una observación futura x_0 es

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n},$$

donde $z_{\alpha/2}$ es el valor z que deja un área de $\alpha/2$ a la derecha.

Ejemplo 9.6: A causa de la disminución en las tasas de interés, el First Citizens Bank recibió muchas solicitudes para hipoteca. Una muestra reciente de 50 créditos hipotecarios resultó en un promedio de \$257,300. Suponga una desviación estándar de la población de \$25,000. Si el siguiente cliente llamó para una solicitud de crédito hipotecario, encuentre un intervalo de predicción de 95% para la cantidad del crédito de este cliente.

Solución: La predicción puntual de la cantidad del crédito del siguiente cliente es $\bar{x} = \$257{,}300$. El valor z aquí es $z_{0.025} = 1.96$. Por lo tanto, un intervalo de predicción de 95% para un crédito futuro es

$$257300 - (1.96)(25000)\sqrt{1 + 1/50} < x_0 < 257300 + (1.96)(25000)\sqrt{1 + 1/50},$$

que da el intervalo (\$207,812.43, \$306,787.57).

El intervalo de predicción brinda una buena estimación de la ubicación de una observación futura. Lo cual es bastante diferente de la estimación del valor medio de la muestra. Debería notarse que la variación de esta predicción es la suma de la variación debida a una estimación de la media y la variación de una sola observación. No obstante, como antes, consideramos primero el caso de la varianza conocida. De manera que resulta importante tratar con el intervalo de predicción de una observación futura en la situación en que se desconoce la varianza. De hecho, en este caso podría utilizarse una distribución t de Student como en el siguiente resultado. Aquí simplemente se reemplaza la distribución normal con la distribución t.

Intervalo de predicción de una observación futura: σ desconocida

Para una distribución normal de mediciones con media desconocida μ y varianza desconocida σ^2, un **intervalo de predicción** de $(1 - \alpha)100\%$ de una observación futura x_0 es

$$\bar{x} - t_{\alpha/2}s\sqrt{1 + 1/n} < x_0 < \bar{x} + t_{\alpha/2}s\sqrt{1 + 1/n},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad, que deja un área de $\alpha/2$ a la derecha.

Pueden utilizarse los intervalos de predicción unilaterales, pues ciertamente se aplican en casos donde, digamos, es necesario enfocarse en observaciones futuras

grandes. Aquí se aplican los límites de predicción superiores. Concéntrese en las observaciones pequeñas futuras que sugieren el uso de límites de predicción más bajos. El límite superior está dado por

$$\bar{x} + t_\alpha s\sqrt{1 + 1/n}$$

y el límite inferior por

$$\bar{x} - t_\alpha s\sqrt{1 + 1/n}.$$

Ejemplo 9.7: Un inspector de alimentos midió aleatoriamente 30 paquetes de carne de res 95% sin grasa. La muestra resultó en una media de 96.2% con la desviación estándar muestral de 0.08%. Encuentre un intervalo de predicción de 99% para un paquete nuevo. Suponga normalidad.

Solución: Para $v = 29$ grados de libertad, $t_{0.005} = 2.756$. Por lo tanto, un intervalo de predicción de 99% para una observación nueva x_0 es

$$96.2 - (2.756)(0.8)\sqrt{1 + \frac{1}{30}} < x_0 < 96.2 + (2.756)(0.8)\sqrt{1 + \frac{1}{30}},$$

que se reduce a (93.96, 98.44).

Uso de límites de predicción para detectar valores extremos

Hasta el momento hemos dado poca atención al concepto de **valores extremos** u observaciones aberrantes. La mayoría de los investigadores son bastante sensibles ante la existencia de observaciones de valores extremos o también llamados datos defectuosos o "malos". Estudiaremos con detalle el concepto en el capítulo 12, donde se ilustra la detección de valores extremos en el análisis de regresión. No obstante, en efecto, resulta de interés considerarlos aquí, pues existen relaciones importantes entre la detección de los valores extremos y los intervalos de predicción.

Para nuestros propósitos es conveniente observar un valor extremo como aquel en que la observación proviene de una población con una media que es diferente de la que determina el resto de la muestra de tamaño n que se estudia. El intervalo de predicción produce un límite que "cubre" una sola observación futura con probabilidad $1 - \alpha$, si viene de la población a partir de la cual se tomó la muestra. Entonces, una metodología para la detección de valores extremos implica la regla de que una **observación es un valor extremo si cae fuera del intervalo de predicción calculado sin incluir la observación cuestionable en la muestra**. Como resultado. Para el intervalo de predicción del ejemplo 9.7, si se observa un nuevo paquete y tiene un contenido porcentual de grasa fuera del intervalo (93.96, 98.44), como se muestra en esta página, podría considerarse un valor extremo.

9.7 Límites de tolerancia

Al estudiar la sección 9.6 aprendimos que el científico o el ingeniero pueden interesarse menos en la estimación de parámetros y más en obtener una noción sobre dónde caerían *observaciones* o mediciones individuales. Entonces, el interés está en los intervalos de predicción. Sin embargo, aun hay un tercer tipo de intervalo que es

útil en muchas aplicaciones. Una vez más, suponga que el interés se centra en torno de la fabricación de la pieza de un componente y que existen especificaciones sobre una dimensión de esa parte. Interesa poco la media de tal dimensión. No obstante, a diferencia del escenario de la sección 9.6, se podría estar menos interesado en una sola observación y más en dónde cae la mayoría de la población. Si las especificaciones del proceso son importantes, entonces el administrador del proceso se interesará en el desempeño a largo plazo, **no en la siguiente observación**. Se debe intentar determinar los límites que en sentido probabilístico "cubren" los valores en la población (es decir, los valores que se miden de la dimensión).

Un método para establecer el límite deseado consiste en determinar un intervalo de confianza sobre una *proporción fija* de las mediciones. Esto se motiva mejor al visualizar una situación en la que hacemos un muestreo aleatorio de una distribución normal con media conocida μ y varianza σ^2. Evidentemente, un límite que cubre el 95% central de la población de observaciones es

$$\mu \pm 1.96\,\sigma.$$

Esto se llama **intervalo de tolerancia** y, en realidad, es exacta la cobertura de 95% de las observaciones medidas. Sin embargo, en la práctica μ y σ rara vez se conocen; por ello, el usuario debe aplicar

$$\bar{x} \pm ks,$$

y ahora, por supuesto, el intervalo es una variable aleatoria y, por consiguiente, la *cobertura* de una proporción de la población disfrutada por el intervalo no es exacta. Como resultado se aplica un intervalo de confianza de $(1-\gamma)100\%$ al planteamiento, ya que no se puede esperar que todo el tiempo $\bar{x} \pm ks$ cubra cualquier proporción específica. Como resultado tenemos la siguiente definición.

Límites de tolerancia

Para una distribución normal de mediciones con media μ y desviación estándar σ, ambas desconocidas, los **límites de tolerancia** están dados por $\bar{x} \pm ks$, donde k se determina de manera que se pueda asegurar con una confianza de $(1-\gamma)100\%$ que los límites dados contienen al menos la proporción $1-\alpha$ de las mediciones.

La tabla A.7 da valores de k para $1-\alpha = 0.90, 0.95, 0.99$; $\gamma = 0.05, 0.01$; y para valores seleccionados de n de 2 a 1000.

Ejemplo 9.8: Una máquina produce piezas de metal que tienen forma cilíndrica. Se toma una muestra de tales piezas y se encuentra que los diámetros son 1.01 , 0.97. 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 y 1.03 centímetros. Encuentre los límites de tolerancia de 99% que contendrán 95% de las piezas de metal que produce esta máquina. Suponga una distribución aproximadamente normal.

Solución: La media muestral y la desviación estándar para los datos dados son

$$\bar{x} = 1.0056 \quad \text{y} \quad s = 0.0246.$$

De la tabla A.7 para $n = 9$, $1-\gamma = 0.99$, y $1-\alpha = 0.95$, encontramos $k = 4.550$ para los límites de los dos lados. De aquí que los límites de tolerancia de 99% sean

$$1.0056 \pm (4.550)(0.0246).$$

Es decir, tenemos 99% de confianza de que el intervalo de tolerancia de 0.894 a 1.117 contendrá 95% de las piezas de metal que produce esta máquina. Es interesante notar que el correspondiente intervalo de confianza de 99% para μ (véase el ejercicio 9.13 de la página 286) tiene un límite inferior de 0.978 y un límite superior de 1.033, lo cual verifica nuestro planteamiento anterior de que un intervalo de tolerancia debe necesariamente ser mayor que un intervalo de confianza con el mismo grado de confianza.

Distinción entre intervalos de confianza, intervalos de predicción e intervalos de tolerancia

Es importante resaltar la diferencia entre los tres tipos de intervalos que estudiamos e ilustramos en las secciones anteriores. Los cálculos son sencillos, aunque la interpretación podría resultar confusa. En aplicaciones de la vida real, tales intervalos no son intercambiables, ya que sus interpretaciones son bastante distintas.

En el caso de los intervalos de confianza, sólo se pone interés en la **media de la población**. Por ejemplo, en el ejercicio 9.15 de la página 286 hay un proceso de ingeniería que produce los alfileres para costura. Se establece una especificación sobre la dureza de Rockwell por debajo de la cual el cliente no aceptará ningún alfiler. Aquí, un parámetro poblacional debe tener un respaldo. Es importante que el ingeniero sepa dónde *van a estar la mayoría de los valores de la dureza de Rockwell*. De manera que deberían utilizarse los límites de tolerancia. Seguramente cuando los límites de tolerancia en cualquier producto del proceso son más rigurosos que las especificaciones del proceso, entonces las noticias son buenas para el administrador del proceso.

Es verdad que la interpretación del límite de tolerancia se relaciona un poco con el intervalo de confianza. El intervalo de tolerancia de $(1 - \alpha)100\%$ sobre, digamos, la proporción 0.95 se puede ver como un intervalo de confianza **sobre el 95 % central** de la distribución normal correspondiente. Los límites de tolerancia unilaterales también son relevantes. En el caso del problema de dureza de Rockwell se desearía tener un límite inferior de la forma $\bar{x} = ks$, tal que tengamos el "99% de confianza de que al menos 99% de los valores de la dureza de Rockwell excederá el valor calculado".

Los límites de predicción se aplican cuando es importante determinar un límite para un **solo valor**. Ni la media ni la ubicación de la mayoría de la población son la cuestión clave. Más bien se requiere la ubicación de una sola nueva observación.

Ejercicios

9.1 Definamos $S'^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$. Muestre que

$$E(S'^2) = [(n-1)/n]\sigma^2,$$

y de aquí que S'^2 es un estimador sesgado para σ^2.

9.2 Si X es una variable aleatoria binomial, demuestre que

a) $P = X/n$ es un estimador insesgado de p;

b) $P' = \frac{X+\sqrt{n}/2}{n+\sqrt{n}}$ es un estimador sesgado de p.

9.3 Muestre que el estimador P' del ejercicio 9.2*b*) se vuelve insesgado conforme $n \to \infty$.

9.4 Una empresa de material eléctrico fabrica bombillas de luz que tienen una duración aproximadamente distribuida de forma normal, con una desviación estándar de 40 horas. Si una muestra de 30 bombillas tiene una duración promedio de 780 horas, encuentre un intervalo de confianza de 96% para la media de la población de todas las bombillas que produce esta empresa.

9.5 A muchos pacientes con problemas cardiacos se les implantó un marcapasos para controlar su ritmo cardiaco.

Se monta un módulo conector de plástico sobre la parte superior del marcapasos. Suponiendo una desviación estándar de 0.0015 y una distribución aproximadamente normal, encuentre un intervalo de confianza de 95% para la media de todos los módulos conectores que fabrica cierta compañía de manufactura. Una muestra aleatoria de 75 módulos tiene un promedio de 0.310 pulgadas.

9.6 Las estaturas de una muestra aleatoria de 50 estudiantes de una universidad muestra una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros.

a) Construya un intervalo de confianza de 98% para la estatura media de todos los estudiantes de la universidad.

b) ¿Qué podemos afirmar con 98% de confianza sobre el tamaño posible de nuestro error, si estimamos que la estatura media de todos los estudiante de la universidad es 174.5 centímetros?

9.7 Una muestra aleatoria de l00 propietarios de automóviles muestra que, en el estado de Virginia, un automóvil se maneja, en promedio, 23,500 kilómetros por año con una desviación estándar de 3900 kilómetros. Suponga que la distribución de las mediciones es aproximadamente normal.

a) Construya un intervalo de confianza de 99% para el número promedio de kilómetros que se maneja un automóvil anualmente en Virginia.

b) ¿Qué puede afirmar con 99% de confianza acerca del tamaño posible de nuestro error, si estimamos que el número promedio de kilómetros manejados por lo propietarios de automóviles en Virginia es 23,500 kilómetros por año?

9.8 ¿De qué tamaño se necesita una muestra en el ejercicio 9.4 si deseamos tener 96% de confianza de que nuestra media muestral esté dentro de 10 horas de la media real?

9.9 ¿De qué tamaño se necesita una muestra en el ejercicio 9.5 si deseamos tener 95% de confianza de que nuestra media muestral esté dentro de 0.0005 pulgada de la media real?

9.10 Un experto en eficiencia desea determinar el tiempo promedio que toma perforar tres hoyos en cierta placa metálica. ¿De qué tamaño se necesita una muestra para tener 95% de confianza de que esta media muestral esté dentro de 15 segundos de la media real? Suponga que por estudios previo se sabe que $\sigma = 40$ segundos.

9.11 Un investigador de la UCLA afirma que la vida de los ratones se puede extender hasta en 25% cuando se reducen las calorías en su alimento en aproximadamente 40%, desde el momento en que se les desteta. Las dietas restringidas se enriquecen a niveles normales con vitaminas y proteínas. Suponiendo que por estudios previos se sabe que $\sigma = 5.8$ meses, ¿cuántos ratones se deberían incluir en nuestra muestra, si deseamos tener 99% de confianza de que la vida media de la muestra esté dentro de 2 meses de la media de la población para todos los ratones sujetos a la dieta reducida?

9.12 El consumo regular de cereales preendulzados contribuye a la caída de los dientes, a las enfermedades cardiacas y a otras enfermedades degenerativas, según estudios realizados por el doctor W. H. Bowen del Instituto Nacional de Salud y el doctor J. Yudben, profesor de nutrición y dietética de la Universidad de Londres. En una muestra aleatoria de 20 porciones sencillas similares del cereal Alpha-Bits, el contenido promedio de azúcar fue de 11.3 gramos con una desviación estándar de 2.45 gramos. Suponiendo que el contenido de azúcar está distribuido normalmente, construya un intervalo de confianza de 95% para el contenido medio de azúcar para porciones sencillas de Alpha-Bits.

9.13 Una máquina produce piezas metálicas de forma cilíndrica. Se toma una muestra de las piezas y los diámetros son 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 y 1.03 centímetros. Encuentre un intervalo de confianza de 99% para el diámetro medio de las piezas de esta máquina. Suponga una distribución aproximadamente normal.

9.14 Una muestra aleatoria de 10 barras de chocolate energético de cierta marca tiene, en promedio, 230 calorías con una desviación estándar de 15 calorías. Construya un intervalo de confianza de 99% para el contenido medio de calorías real de esta marca de barras de chocolate energético. Suponga que la distribución de las calorías es aproximadamente normal.

9.15 Se toma una muestra aleatoria de 12 alfileres para costura en un estudio de dureza de Rockwell en la cabeza de los alfileres. Se realizaron mediciones de la dureza de Rockwell para cada una de las 12, lo cual dio un valor promedio de 48.50 con una desviación estándar muestral de 1.5. Suponiendo que las mediciones se distribuyen de forma normal, construya un intervalo de confianza de 90% para la dureza de Rockwell media.

9.16 Una muestra aleatoria de 12 graduadas de cierta escuela secretarial teclearon un promedio de 79.3 palabras por minuto, con una desviación estándar de 7.8 palabras por minuto. Suponiendo una distribución normal para el número de palabras que se teclea por minuto, encuentre un intervalo de confianza de 95% para el número promedio de palabras tecleadas por todas las graduadas de esta escuela.

9.17 Una muestra aleatoria de 25 botellas de aspirinas contiene, en promedio, 325.05 mg de aspirina con una desviación estándar de 0.5. Encuentre los límites de tolerancia de 95% que contendrán 90% del contenido de aspirina para esta marca. Suponga que el contenido de aspirina se distribuye normalmente.

9.18 Las siguientes mediciones se registraron para el tiempo de secado, en horas, de cierta marca de pintura látex:

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponiendo que las mediciones representan una muestra aleatoria de una población normal, encuentre los

límites de tolerancia de 99% que contendrán 95% de los tiempos de secado.

9.19 Refiérase al ejercicio 9.7 y construya un intervalo de tolerancia de 99% que contenga 99% de las millas que recorren los automóviles anualmente en Virginia.

9.20 Refiérase al ejercicio 9.15, construya un intervalo de tolerancia de 95% que contenga 90% de las mediciones.

9.21 En la sección 9.3 destacamos la noción del "estimador más eficaz" comparando la varianza de dos estimadores insesgados $\hat{\Theta}_1$ y $\hat{\Theta}_2$. Sin embargo, ello no toma en cuenta el sesgo en el caso de que uno o ambos estimadores no sean insesgados. Considere la cantidad

$$\text{ECM} = E(\hat{\Theta} - \theta),$$

donde ECM denota el **error cuadrado medio**. El error cuadrado medio a menudo se utiliza para comparar dos estimadores $\hat{\Theta}_1$ y $\hat{\Theta}_2$ de θ, cuando uno o ambos son insesgados porque **i.** ello es intuitivamente razonable y **ii.** se toma en cuenta para el sesgo. Demuestre que el ECM se puede escribir como

$$\begin{aligned}\text{ECM} &= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [E(\hat{\Theta} - \theta)]^2 \\ &= Var(\hat{\Theta}) + [sesgo(\hat{\Theta})]^2.\end{aligned}$$

9.22 Considere el ejercicio 9.1 y S'^2, el estimador de σ^2. El analista a menudo utiliza S'^2 en vez de dividir $\sum_{i=1}^{n}(X_i - \bar{X})^2$ entre $n - 1$, los *grados de libertad* en la muestra.

a) ¿Cuál es el sesgo de S'^2?
b) Demuestre que el sesgo de S'^2 se aproxima a cero conforme $n \to \infty$.

9.23 Compare S^2 y S'^2 (véase el ejercicio 9.1), los dos estimadores de σ^2, para determinar cuál es más eficaz. Suponga que son estimadores que se encuentran usando $X_1, X_2, \ldots, X_n$, las variables aleatorias independientes de $n(x; \mu, \sigma)$. ¿Cuál es el estimador más eficaz considerando sólo la varianza de los estimadores? [*Sugerencia:* Utilice el teorema 8.4 y la sección 6.8 donde aprendimos que la varianza de χ_v^2 es $2v$.]

9.24 Considere el ejercicio 9.23. Utilice el ECM que se estudió en el ejercicio 9.21 para determinar qué estimador es más eficaz. De hecho, escriba

$$\frac{\text{ECM}(S^2)}{\text{ECM}(S'^2)}.$$

9.25 Considere el ejercicio 9.12. Calcule un intervalo de predicción de 95% para el contenido de azúcar de la siguiente porción sencilla del cereal Alpha-Bits.

9.26 Considere el ejercicio 9.16. Calcule el intervalo de predicción de 95% para el siguiente número observado de palabras por minuto tecleado por un miembro del secretariado escolar.

9.27 Considere el ejercicio 9.18. Calcule un intervalo de predicción de 95% en una nueva medición observada del tiempo de secado de la pintura látex.

9.28 Considere la situación del ejercicio 9.13. Aunque la estimación del diámetro medio sea importante, no es ni con mucho tan importante como intentar "determinar" la ubicación de la mayoría de la distribución de los diámetros. Para tal fin, encuentre los límites de tolerancia de 95% que contengan 95% de los diámetros.

9.29 En un estudio realizado en el Departamento de Zoología del Virginia Tech, se recolectaron 15 "muestras" de agua de una determinada estación en el río James, con la finalidad de conocer la cantidad de ortofósforo en el río. La concentración del químico se mide en miligramos por litro. Supongamos que la media en la estación no es tan importante como los extremos superiores de la distribución del químico en la estación. El interés se centra en saber si las concentraciones en estos extremos son demasiado elevadas. Las lecturas de las 15 muestras de agua dieron una media muestral de 3.84 miligramos por litro y una desviación estándar de 3.07 miligramos por litro. Suponga que las lecturas son una muestra aleatoria de una distribución normal. Calcule un intervalo de predicción (límite de predicción superior de 95%) y un límite de tolerancia (un límite de tolerancia superior de 95% que excede 95% de la población de valor). Interprete ambos; esto es, diga qué nos comunican acerca de los extremos superiores de la distribución de ortofósforo en la estación de muestreo.

9.30 Un determinado tipo de hilo se somete a estudio para conocer sus propiedades de resistencia a la tensión. Se probaron 50 piezas en condiciones similares y los resultados mostraron una resistencia a la tensión promedio de 78.3 kilogramos y una desviación estándar de 5.6 kilogramos. Suponiendo una distribución normal de la resistencia a la tensión, dé un intervalo de predicción inferior de 95% en un único valor observado de resistencia a la tensión. Además, determine un límite inferior de tolerancia de 95% que sea excedido por 99% de los valores de resistencia a la tensión.

9.31 Remítase al ejercicio 9.30. ¿Por qué las cantidades solicitadas en el ejercicio parecen ser más importantes para el fabricante del hilo que, por ejemplo, un intervalo de confianza en la resistencia media a la tensión?

9.32 Remítase una vez más al ejercicio 9.30. Suponga que las especificaciones de un comprador del hilo son que la resistencia a la tensión del material debe ser, por lo menos, de 62 kilogramos. El fabricante está satisfecho si, cuando mucho, el 5% de las piezas producidas tienen una resistencia a la tensión menor de

62 kilogramos. ¿Hay alguna razón para preocuparse? Esta vez utilice un límite de tolerancia unilateral de 99% que sea excedido por 95% de los valores de resistencia a la tensión.

9.33 Considere las medidas del tiempo de secado del ejercicio 9.18. Suponga que las 15 observaciones en el conjunto de datos también incluyen un 16o. valor de 6.9 horas. En el contexto de las 15 observaciones originales, ¿el decimosexto es un valor extremo? Demuestre su trabajo.

9.34 Considere los datos del ejercicio 9.15. Suponga que el fabricante de los alfileres insiste en que la dureza del producto será tan baja o más que el valor de 44.0 sólo un 5% de las veces. ¿Cuál es su reacción al respecto? Realice el cálculo de un límite de tolerancia para determinar su veredicto.

9.8 Dos muestras: Estimación de la diferencia entre dos medias

Si tenemos dos poblaciones con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2, respectivamente, un estimador puntual de la diferencia entre μ_1 y μ_2 está dado por el estadístico $\bar{X}_1 - \bar{X}_2$. Por lo tanto, para obtener una estimación puntual de $\mu_1 - \mu_2$, seleccionaremos dos muestras aleatorias independientes, una de cada población, de tamaños n_1 y n_2, y calculamos la diferencia $\bar{x}_1 - \bar{x}_2$, de las medias muestrales. Evidentemente, debemos considerar las distribuciones muestrales de $\bar{X}_1 - \bar{X}_2$.

De acuerdo con el teorema 8.3, podemos esperar que la distribución muestral de $\bar{X}_1 - \bar{X}_2$ esté distribuida de forma aproximadamente normal con media $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ y desviación estándar $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Por lo tanto, podemos asegurar con una probabilidad de $1-\alpha$ que la variable normal estándar

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

caerá entre $-z_{\alpha/2}$ y $z_{\alpha/2}$. Con referencia una vez más a la figura 9.2, escribimos

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Al sustituir para Z, establecemos de manera equivalente que

$$P\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha,$$

que conduce al siguiente intervalo de confianza de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$.

Intervalo de confianza para $\mu_1 - \mu_2$; con σ_1^2 y σ_2^2 conocidas

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes de tamaños n_1 y n_2 de poblaciones con varianzas conocidas σ_1^2 y σ_2^2, respectivamente, un intervalo de confianza de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ está dado por

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

donde $z_{\alpha/2}$ es el valor z que deja un área de $\alpha/2$ a la derecha.

El grado de confianza es exacto cuando las muestras se seleccionan de poblaciones normales. Para poblaciones no normales, el teorema del límite central permite una buena aproximación para muestras de tamaños razonables.

Las condiciones experimentales y la unidad experimental

Para el caso de la estimación de un intervalo de confianza sobre la diferencia entre dos medias, necesitamos considerar las condiciones experimentales del proceso de recolección de datos. Se supone que tenemos dos muestras aleatorias independientes de distribuciones con medias μ_1 y μ_2, respectivamente. Es importante que las condiciones experimentales simulen este "ideal" descrito por las suposiciones tan cerca como sea posible. Muy a menudo el experimentador debería planear la estrategia del experimento en consecuencia. Para casi cualquier estudio de este tipo, hay una llamada *unidad experimental*, que es la parte del experimento que produce el error experimental y que es responsable de la varianza de la población que denominamos σ^2. En un estudio médico, la unidad experimental es el paciente o el sujeto. En un experimento de agricultura, puede ser una superficie de tierra. En un experimento químico, puede ser una cantidad de materias primas. Resulta importante que las diferencias entre tales unidades tengan un impacto mínimo sobre los resultados. El experimentador tendrá un grado de seguridad de que las unidades experimentales no sesgarán los resultados, si las condiciones que definen a las dos poblaciones se *asignan al azar* a las unidades experimentales. De nuevo nos concentraremos en la aleatoriedad en los siguientes capítulos que tratan de la prueba de hipótesis.

Ejemplo 9.9: Se lleva a cabo un experimento donde se comparan dos tipos de motores, A y B. Se mide el rendimiento de combustible en millas por galón. Se realizan 50 experimentos con el motor tipo A y 75 con el motor tipo B. La gasolina que se utiliza y las demás condiciones se mantienen constantes. El rendimiento promedio de gasolina para el motor A es de 36 millas por galón, y el promedio para el motor B es de 42 millas por galón. Encuentre un intervalo de confianza de 96% sobre $\mu_B - \mu_A$, donde μ_A y μ_B son el rendimiento de combustible medio poblacional para los motores A y B, respectivamente. Suponga que las desviaciones estándar poblacionales son 6 y 8 para los motores A y B, respectivamente.

Solución: La estimación puntual de $\mu_B - \mu_A$ es $\bar{x}_B - \bar{x}_A = 42 - 36 = 6$. Usando $\alpha = 0.04$, encontramos $z_{0.02} = 2.05$ de la tabla A.3. De aquí, con la sustitución en la fórmula anterior, el intervalo de confianza de 96% es

$$6 - 2.05\sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_B - \mu_A < 6 + 2.05\sqrt{\frac{64}{75} + \frac{36}{50}},$$

o simplemente $3.43 < \mu_B - \mu_A < 8.57$.

Este procedimiento para estimar la diferencia entre dos medias se aplica si se conocen σ_1^2 y σ_2^2. Si las varianzas no se conocen y las dos distribuciones implicadas son aproximadamente normales, la distribución t resulta implicada como en el caso de una sola muestra. Si no se está dispuesto a suponer normalidad, muestras grandes (digamos mayores que 30) permitirán usar s_1 y s_2 en vez de σ_1 y σ_2, respectivamente, con el fundamento de que $s_1 \approx \sigma_1$, y $s_2 \approx \sigma_2$. De nuevo, por supuesto, el intervalo de confianza es aproximado.

Varianzas desconocidas

Considere el caso donde se desconocen σ_1^2 y σ_2^2. Si $\sigma_1^2 = \sigma_2^2 = \sigma^2$, obtenemos una variable normal estándar de la forma

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}}.$$

De acuerdo con el teorema 8.4, las dos variables aleatorias

$$\frac{(n_1-1)S_1^2}{\sigma^2} \quad \text{y} \quad \frac{(n_2-1)S_2^2}{\sigma^2}$$

tienen distribuciones chi cuadrada con n_1-1 y n_2-1 grados de libertad, respectivamente. Además, son variables chi cuadradas independientes, ya que las muestras aleatorias se seleccionaron de forma independiente. En consecuencia, su suma

$$V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v=n_1+n_2-2$ grados de libertad.

Como se puede mostrar que las expresiones anteriores para Z y V son independientes, del teorema 8.5 se sigue que el estadístico

$$T=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma^2[(1/n_1)+(1/n_2)]}}\Big/\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}$$

tiene la distribución t con $v_1=n_1+n_2-2$ grados de libertad.

Se puede obtener una estimación puntual de la varianza común desconocida σ^2 al reunir las varianzas muestrales. Denotemos al estimador de unión con S_p^2 y escribimos, entonces,

Estimado de unión de la varianza

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

Al sustituir S_p^2 en el estadístico T, obtenemos la forma menos incómoda:

$$T=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{(1/n_1)+(1/n_2)}}.$$

Usando el estadístico T, tenemos

$$P(-t_{\alpha/2}<T<t_{\alpha/2})=1-\alpha,$$

donde $t_{\alpha/2}$ es el valor t con n_1+n_2-2 grados de libertad, por arriba del cual encontramos un área de $\alpha/2$. Al sustituir para T en la desigualdad, escribimos

$$P\left[-t_{\alpha/2}<\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{(1/n_1)+(1/n_2)}}<t_{\alpha/2}\right]=1-\alpha.$$

Después de llevar a cabo las manipulaciones matemáticas de costumbre, se calculan la diferencia de las medias muestrales $\bar{x}_1-\bar{x}_2$ y la varianza unida, y se obtiene el siguiente intervalo de confianza de $(1-\alpha)100\%$ para $\mu_1-\mu_2$.

Se ve con facilidad que el valor de s_p^2 es un promedio ponderado de las dos varianzas muestrales s_1^2 y s_2^2, donde los pesos son los grados de libertad.

Intervalo de confianza para $\mu_1 - \mu_2$; con $\sigma_1^2 = \sigma_2^2$ pero desconocidas

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes con tamaños n_1 y n_2, respectivamente, de poblaciones aproximadamente normales con varianzas iguales pero desconocidas, un intervalo de confianza de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ está dado por

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

donde s_p es la estimación de unión de la desviación estándar poblacional y $t_{\alpha/2}$ es el valor t con $v = n_1 + n_2 - 2$ grados de libertad, que deja un área de $\alpha/2$ a la derecha.

Ejemplo 9.10: En el artículo "Macroinvertebrate Community Structure as an Indicator of Acid Mine Pollution", publicado en el *Journal of Environmental Pollution*, se ofrece un reporte sobre una investigación realizada en Cane Creek, Alabama, para determinar la relación entre parámetros fisioquímicos seleccionados y diversas mediciones de la estructura de la comunidad de macroinvertebrados. Una faceta de la investigación fue una evaluación de la efectividad de un índice numérico de la diversidad de especies, para indicar la degradación del agua debida al desagüe ácido de una mina. Conceptualmente, un índice alto de la diversidad de especies macroinvertebradas debería indicar un sistema acuático no contaminado; mientras que un índice de diversidad baja indicaría un sistema acuático contaminado.

Se eligieron 2 estaciones de muestreo independientes para dicho estudio: una que se localiza corriente abajo del punto de descarga ácida de la mina y la otra ubicada corriente arriba. Para 12 muestras mensuales reunidas en la estación corriente abajo, el índice de diversidad de especies tuvo un valor medio $\bar{x}_1 = 3.11$ y una desviación estándar $s_1 = 0.771$; mientras que 10 muestras reunidas mensualmente en la estación corriente arriba tuvieron un valor medio del índice $\bar{x}_2 = 2.04$ y una desviación estándar $s_2 = 0.448$. Encuentre un intervalo de confianza de 90% para la diferencia entre las medias poblacionales para los dos sitios, suponiendo que las poblaciones están distribuidas de forma aproximadamente normal con varianzas iguales.

Solución: Representemos con μ_1 y μ_2 las medias poblacionales, respectivamente, para los índices de diversidad de especies en las estaciones corriente abajo y corriente arriba. Deseamos encontrar un intervalo de confianza de 90% para $\mu_1 - \mu_2$. Nuestra estimación puntual de $\mu_1 - \mu_2$ es

$$\bar{x}_1 - \bar{x}_2 = 3.11 - 2.04 = 1.07.$$

La estimación de la unión, s_p^2, de la varianza común, σ^2, es

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} = \frac{(11)(0.771^2) + (9)(0.448^2)}{12 + 10 - 2} = 0.417.$$

Al tomar la raíz cuadrada, obtenemos $s_p = 0.646$. Usando $\alpha = 0.1$, encontramos en la tabla A.4 que $t_{0.05} = 1.725$ para $v = n_1 + n_2 - 2 = 20$ grados de libertad. Por lo tanto, el intervalo de confianza de 90% para $\mu_1 - \mu_2$ es

$$1.07 - (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}} < \mu_1 - \mu_2 < 1.07 + (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}}.$$

que se simplifica $0.593 < \mu_1 - \mu_2 < 1.547$.

Interpretación del intervalo de confianza

Para el caso de un solo parámetro, el intervalo de confianza simplemente produce límites de error sobre el parámetro. Los valores contenidos en el intervalo se deberían ver como valores razonables dados los datos experimentales. En el caso de una diferencia entre dos medias, la interpretación se puede extender a una de comparación de las dos medias. Por ejemplo, si tenemos gran confianza en que una diferencia $\mu_1 - \mu_2$ es positiva, realmente inferiremos que $\mu_1 > \mu_2$ con poco riesgo de incurrir en un error. De esta forma, en el ejemplo 9.10 tenemos una confianza de 90% de que el intervalo de 0.593 a 1.547 contiene la diferencia de las medias poblacionales, para valores del índice de diversidad de especies en las dos estaciones. El hecho de que ambos límites de confianza sean positivos indica que, en promedio, el índice para la estación que se localiza corriente abajo del punto de descarga es mayor que el índice para la estación que se localiza corriente arriba.

Tamaños iguales de muestras

El procedimiento para construir intervalos de confianza para $\mu_1 - \mu_2$ con $\sigma_1 = \sigma_2 = \sigma$ desconocidas requiere la suposición de que las poblaciones son normales. Desviaciones ligeras de la suposición de varianzas iguales o de normalidad no alteran seriamente el grado de confianza de nuestro intervalo. (En el capítulo 10 se estudia un procedimiento para probar la igualdad de dos varianzas poblacionales desconocidas con base en la información que proporcionan las varianzas muestrales.) Si las varianzas poblacionales son considerablemente diferentes, aún obtenemos resultados razonables cuando las poblaciones son normales, dado que $n_1 = n_2$. Por lo tanto, en un experimento planeado se debería hacer un esfuerzo para igualar el tamaño de las muestras.

Varianzas distintas

Consideremos ahora el problema de encontrar una estimación por intervalos de $\mu_1 - \mu_2$ cuando no es probable que las varianzas poblacionales desconocidas sean iguales. El estadístico que se utiliza con mayor frecuencia en este caso es

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}},$$

que tiene aproximadamente una distribución t con v grados de libertad, donde

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}.$$

Como v rara vez es un entero, lo *redondeamos* (hacia abajo) al número entero más cercano.

Con el estadístico T', escribimos

$$P(-t_{\alpha/2} < T' < t_{\alpha/2}) \approx 1 - \alpha,$$

donde $t_{\alpha/2}$ es el valor de la distribución t con v grados de libertad, arriba del cual encontramos un área de $\alpha/2$. Al sustituir para T' en la desigualdad, y al seguir los pasos exactos como antes, establecemos el resultado final.

Intervalo de confianza para $\mu_1 - \mu_2$; $\sigma_1^2 \neq \sigma_2^2$ y desconocidas

Si $\bar{x}_1$ y s_1^2 y $\bar{x}_2$ y s_2^2 son las medias y varianzas de muestras aleatorias independientes de tamaños n_1 y n_2, respectivamente, de poblaciones aproximadamente normales con varianzas desconocidas y diferentes, un intervalo de confianza aproximado del $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ está dado por

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

donde $t_{\alpha/2}$ es el valor t con

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1-1)] + [(s_2^2/n_2)^2/(n_2-1)]}$$

grados de libertad, que deja un área $\alpha/2$ a la derecha.

Observe que el valor v anterior incluye variables aleatorias y, por ello, representa una *estimación* de los grados de libertad. En las aplicaciones dicha estimación no será un número entero, de manera que el analista lo debe redondear al siguiente entero para tener la confianza que se desea.

Antes de ilustrar el intervalo de confianza anterior con un ejemplo, deberíamos señalar que todos los intervalos de confianza sobre $\mu_1 - \mu_2$ son de la misma forma general, como los de una sola media; a saber, se pueden escribir como

$$\text{estimación puntual} \pm t_{\alpha/2}\ \widehat{\text{e.e.}}(\text{estimación puntual})$$

o

$$\text{estimación puntual} \pm z_{\alpha/2}\ \text{e.e.}(\text{estimación puntual}).$$

Por ejemplo, en el caso donde $\sigma_1 = \sigma_2 = \sigma$, el error estándar estimado de $\bar{x}_1 - \bar{x}_2$ es $s_p\sqrt{1/n_1 + 1/n_2}$. Para el caso donde $\sigma_1^2 \neq \sigma_2^2$,

$$\widehat{\text{e.e.}}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Ejemplo 9.11: El Departamento de Zoología del Instituto Politécnico y Universidad Estatal de Virginia llevó a cabo un estudio para estimar la diferencia en la cantidad de ortofósforo químico medido en dos estaciones diferentes del río James. El ortofósforo se mide en miligramos por litro. Se reunieron 15 muestras de la estación 1 y 12 muestras de la estación 2. Las 15 muestras de la estación 1 tuvieron un contenido promedio de ortofósforo de 3.84 miligramos por litro y una desviación estándar de 3.07 miligramos por litro; en tanto que las 12 muestras de la estación 2 tuvieron un contenido promedio de 1.49 miligramos por litro y una desviación estándar de 0.80 miligramos por litro. Encuentre un intervalo de confianza de 95% para la diferencia en el contenido promedio real de ortofósforo en estas dos estaciones. Suponga que las observaciones provienen de poblaciones normales con varianzas diferentes.

Solución: Para la estación 1 tenemos $\bar{x}_1 = 3.84$, $s_1 = 3.07$ y $n_1 = 15$. Para la estación 2, $\bar{x}_2 = 1.49$, $s_2 = 0.80$ y $n_2 = 12$. Deseamos encontrar un intervalo de confianza de 95% para

$\mu_1 - \mu_2$. Como las varianzas poblacionales se suponen diferentes, sólo podemos encontrar un intervalo de confianza de 95% aproximado basado en la distribución t con v grados de libertad, donde

$$v = \frac{(3.07^2/15 + 0.80^2/12)^2}{[(3.07^2/15)^2/14] + [(0.80^2/12)^2/11]} = 16.3 \approx 16.$$

Nuestra estimación puntual de $\mu_1 - \mu_2$ es

$$\bar{x}_1 - \bar{x}_2 = 3.84 - 1.49 = 2.35.$$

Al usar $\alpha = 0.05$, en la tabla A.4 encontramos que $t_{0.025} = 2.120$ para $v = 16$ grados de libertad. Por lo tanto, el intervalo de confianza de 95% para $\mu_1 - \mu_2$ es

$$2.35 - 2.120\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}} < \mu_1 - \mu_2 < 2.35 + 2.120\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}},$$

que se simplifica a $0.60 < \mu_1 - \mu_2 < 4.10$. Por ello, tenemos una confianza de 95% de que el intervalo de 0.60 a 4.10 miligramos por litro contiene la diferencia de los contenidos promedio reales de ortofósforo para estos dos lugares.

9.9 Observaciones pareadas

En esta sección consideraremos los procedimientos de estimación para la diferencia de dos medias cuando las muestras no son independientes y las varianzas de las dos poblaciones no son necesariamente iguales. La situación que se considera aquí tiene que ver con una situación experimental muy especial; a saber, *las observaciones pareadas*. A diferencia de la situación que se describió antes, las condiciones de las dos poblaciones no se asignan de forma aleatoria a las unidades experimentales. Más bien, cada unidad experimental homogénea recibe ambas condiciones poblacionales; como resultado, cada unidad experimental tiene un par de observaciones, una para cada población. Por ejemplo, si realizamos una prueba de una nueva dieta con 15 individuos, los pesos antes y después de seguir la dieta forman la información de nuestras dos muestras. Estas dos poblaciones son "antes" y "después", y la unidad experimental es el individuo. Evidentemente las observaciones en un par tienen algo en común. Para determinar si la dieta es efectiva, consideramos las diferencias d_1, $d_2, \ldots, d_n$ en las observaciones pareadas. Estas diferencias son los valores de una muestra aleatoria D_1, $D_2, \ldots, D_n$ de una población de diferencias, que supondremos distribuidas normalmente, con media $\mu_D = \mu_1 - \mu_2$ y varianza σ_D^2. Estimamos σ_D^2, mediante s_d^2, la varianza de las diferencias que constituyen nuestra muestra. El estimador puntual de μ_D está dado por $\bar{D}$.

¿Cuándo debería hacerse el pareo?

Parear observaciones en un experimento es una estrategia que se puede emplear en muchos campos de aplicación. Se expondrá al lector a tal concepto en el material relativo a la prueba de hipótesis en el capítulo 10 y en los temas de diseño experimental en los capítulos 13 y 15. Al seleccionar unidades experimentales relativamente homogéneas (dentro de las unidades) y permitir que cada unidad experimente

ambas condiciones poblacionales, se reduce la "varianza del error experimental" efectiva (en este caso σ_D^2). El lector puede visualizar que el i-ésimo par consiste en la medición

$$D_i = X_{1i} - X_{2i}.$$

Como las dos observaciones se toman de la unidad experimental de la muestra, no son independientes y, de hecho,

$$Var(D_i) = Var(X_{1i} - X_{2i}) = \sigma_1^2 + \sigma_2^2 - 2Cov(X_{1i}, X_{2i}).$$

Entonces, de manera intuitiva, se espera que σ_D^2 debería reducirse gracias a la similitud en la naturaleza de los "errores" de las dos observaciones dentro de una unidad experimental, y esto se logra a través de la expresión anterior. En realidad se espera que si la unidad es homogénea, la covarianza será positiva. Como resultado, la ganancia en calidad del intervalo de confianza sobre el no pareado será mayor cuando haya homogeneidad dentro de las unidades, y diferencias grandes conforme se vaya de una unidad a otra. Se debería tener en cuenta que el desempeño del intervalo de confianza dependerá del error estándar de $\bar{D}$ que es, por supuesto, $\sigma_D/\sqrt{n}$, donde n es el número de pares. Como indicamos antes, la intención de parear es reducir σ_D.

Evaluación entre reducir la varianza y perder grados de libertad

Al comparar la situación del intervalo de confianza pareado contra la del sin parear, se vuelve evidente que hay un "intercambio" implicado. Aunque en realidad el hecho de parear debería reducir la varianza y, por ello, reducir el error estándar de la estimación puntual, los grados de libertad se reducen al reducir el problema a uno de una sola muestra. Como resultado, el punto $t_{\alpha/2}$ unido al error estándar se ajusta en consecuencia. Por lo tanto, el pareamiento podría resultar contraproducente. Éste en realidad sería el caso si se experimenta sólo una reducción modesta en la varianza (a través de σ_D^2) mediante el pareamiento.

Otra ilustración del pareamiento implicaría la elección de n pares de sujetos, donde cada par tiene una característica similar, como CI, edad, raza, etcétera; entonces, para cada par se selecciona un miembro al azar para obtener un valor de X_1, mientras que el otro miembro da el valor de X_2. En este caso, X_1 y X_2 pueden representar las calificaciones que obtienen dos individuos de CI igual, cuando uno de los individuos se asigna al azar a un grupo que usa el sistema de clases convencional, mientras que el otro individuo se asigna a un grupo que utiliza materiales programados.

Se puede establecer un intervalo de confianza de $(1 - \alpha)100\%$ para μ_D al escribir

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

donde $T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}}$ y $t_{\alpha/2}$, como antes, es un valor de la distribución t con $n - 1$ grados de libertad.

Es ahora un procedimiento de rutina reemplazar T por su definición, en la desigualdad anterior y desarrollar los pasos matemáticos que conduzcan al siguiente intervalo de confianza de $(1 - \alpha)100\%$ para $\mu_1 - \mu_2 = \mu_D$.

Tabla 9.1: Datos para el ejemplo 9.12

Veterano	Niveles de TCDD en plasma	Niveles de TCDD en tejido adiposo	d_i	Veterano	Niveles de TCDD en plasma	Niveles de TCDD en tejido adiposo	d_i
1	2.5	4.9	−2.4	11	6.9	7.0	−0.1
2	3.1	5.9	−2.8	12	3.3	2.9	0.4
3	2.1	4.4	−2.3	13	4.6	4.6	0.0
4	3.5	6.9	−3.4	14	1.6	1.4	0.2
5	3.1	7.0	−3.9	15	7.2	7.7	−0.5
6	1.8	4.2	−2.4	16	1.8	1.1	0.7
7	6.0	10.0	−4.0	17	20.0	11.0	9.0
8	3.0	5.5	−2.5	18	2.0	2.5	−0.5
9	36.0	41.0	−5.0	19	2.5	2.3	0.2
10	4.7	4.4	0.3	20	4.1	2.5	1.6

Fuente: Schecter, A. *et al.*, "Partitioning of 2, 3, 7, 8-chlorinated dibenzo-p-dioxins and dibenzofurans between adipose tissue and plasma lipid of 20 Massachusetts Vietnam veterans", *Chemosphere*, vol. 20, núms. 7-9, 1990, pp. 954-955 (tablas I y II).

Intervalo de confianza para $\mu_D = \mu_1 - \mu_2$; para observaciones pareadas

Si $\bar{d}$ y s_d son la media y la desviación estándar, respectivamente, de las diferencias distribuidas normalmente de n pares aleatorios de mediciones, un intervalo de confianza de $(1-\alpha)100\%$ para $\mu_D = \mu_1 - \mu_2$ es

$$\bar{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad que deja un área de $\alpha/2$ a la derecha.

Ejemplo 9.12: Un estudio publicado en *Chemosphere* reporta los niveles de la dioxina TCDD de 20 veteranos de Vietnam residentes en Massachusetts, quienes posiblemente se expusieron al agente naranja. La cantidad en los niveles de TCDD en plasma y en tejido adiposo se presentan en la tabla 9.1.

Encuentre un intervalo de confianza de 95% para $\mu_1 - \mu_2$, donde μ_1 y μ_2 representen las medias reales de TCDD en plasma y en tejido adiposo, respectivamente. Suponga que la distribución de las diferencias es aproximadamente normal.

Solución: Deseamos encontrar un intervalo de confianza de 95% para $\mu_1 - \mu_2$. Como las observaciones están pareadas, $\mu_1 - \mu_2 = \mu_D$. La estimación puntual de μ_D es $\bar{d} = -0.87$. La desviación estándar s_d de las diferencias muestrales es

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} = \sqrt{\frac{168.4220}{19}} = 2.9773.$$

Usando $\alpha = 0.05$, en la tabla A.4 encontramos que $t_{0.025} = 2.093$ para $v = n - 1 = 19$ grados de libertad. Por lo tanto, el intervalo de confianza de 95% es

$$-0.8700 - (2.093)\left(\frac{2.9773}{\sqrt{20}}\right) < \mu_D < -0.8700 + (2.093)\left(\frac{2.9773}{\sqrt{20}}\right),$$

o simplemente $-2.2634 < \mu_D < 0.5234$, del cual concluimos que no hay diferencia significativa entre el nivel medio de TCDD en plasma y el nivel medio de TCDD en tejido adiposo.

Ejercicios

9.35 Una muestra aleatoria de tamaño $n_1 = 25$ que se toma de una población normal con una desviación estándar $\sigma_1 = 5$ tiene una media $\bar{x}_1 = 80$. Una segunda muestra aleatoria de tamaño $n_2 = 36$, que se toma de una población normal diferente con una desviación estándar $\sigma_2 = 3$, tiene una media $\bar{x}_2 = 75$. Encuentre un intervalo de confianza de 94% para $\mu_1 - \mu_2$.

9.36 Se comparan las resistencias de dos clases de hilo. Cincuenta piezas de cada clase de hilo se prueban bajo condiciones similares. La marca A tiene una resistencia a la tensión promedio de 78.3 kilogramos con una desviación estándar de 5.6 kilogramos; en tanto que la marca B tiene una resistencia a la tensión promedio de 87.2 kilogramos con una desviación estándar de 6.3 kilogramos. Construya un intervalo de confianza de 95% para la diferencia de las medias poblacionales.

9.37 Se realiza un estudio para determinar si cierto tratamiento metálico tiene algún efecto sobre la cantidad de metal que se elimina en una operación de decapado. Una muestra aleatoria de 100 piezas se sumerge en un baño por 24 horas sin el tratamiento, lo que da un promedio de 12.2 milímetros de metal eliminados y una desviación estándar muestral de 1.1 milímetros. Una segunda muestra de 200 piezas se somete al tratamiento, seguida de 24 horas de inmersión en el baño, lo que da como resultado una eliminación promedio de 9.1 milímetros de metal, con una desviación estándar muestral de 0.9 milímetros. Calcule una estimación del intervalo de confianza de 98% para la diferencia entre las medias de las poblaciones. ¿El tratamiento parece reducir la cantidad media del metal eliminado?

9.38 En un proceso químico por lotes, se comparan los efectos de dos catalizadores sobre la potencia de la reacción del proceso. Se preparó una muestra de 12 lotes con el uso del catalizador 1 y se obtuvo una muestra de 10 lotes con el catalizador 2. Los 12 lotes para los que se utilizó el catalizador 1 dieron un rendimiento promedio de 85 con una desviación estándar muestral de 4; en tanto que para la segunda muestra el promedio fue de 81 con una desviación estándar muestral de 5. Encuentre un intervalo de confianza de 90% para la diferencia entre las medias poblacionales, suponiendo que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales.

9.39 Los estudiantes pueden elegir entre un curso de física de tres semestres-hora sin laboratorio y un curso de cuatro semestres-hora con laboratorio. El examen final escrito es el mismo para cada sección. Si 12 estudiantes de la sección con laboratorio tiene una calificación promedio en el examen de 84 con una desviación estándar de 4, y 18 estudiantes de la sección sin laboratorio tienen una calificación promedio de 77 con una desviación estándar de 6, encuentre un intervalo de confianza de 99% para la diferencia entre las calificaciones promedio para ambos cursos. Suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales.

9.40 En un estudio que se lleva a cabo en el Instituto Politécnico y Universidad Estatal de Virginia sobre el desarrollo de ectomycorrhizal, una relación simbiótica entre las raíces de los árboles y un hongo, en la cual se transfieren minerales del hongo a los árboles y azúcares de los árboles a los hongos, se plantan en un invernadero 20 robles rojos con el hongo *Pisolithus tinctorus*. Todos los árboles se plantan en el mismo tipo de suelo y reciben la misma cantidad de luz solar y agua. La mitad no recibe nitrógeno en el momento de plantarlos para servir como control y la otra mitad recibe 368 ppm de nitrógeno en forma de $NaN0_3$. Los pesos de los tallos, que se registran en gramos, al final de 140 días se registran como sigue:

Sin nitrógeno	**Con nitrógeno**
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

Construya un intervalo de confianza de 95% para la diferencia en los pesos medios de los tallos entre los que no recibieron nitrógeno y los que recibieron 368 ppm de nitrógeno. Suponga que las poblaciones están distribuidas normalmente con varianzas iguales.

9.41 Los siguientes datos, registrados en días, representan el tiempo de recuperación para pacientes que se tratan al azar con uno de dos medicamentos para curar infecciones graves de la vejiga:

Medicamento 1	**Medicamento 2**
$n_1 = 14$	$n_2 = 16$
$\bar{x}_1 = 17$	$\bar{x}_2 = 19$
$s_1^2 = 1.5$	$s_2^2 = 1.8$

Encuentre un intervalo de confianza de 99% para la diferencia $\mu_2 - \mu_1$ en el tiempo medio de recuperación para los dos medicamentos. Suponga poblaciones normales con varianzas iguales.

9.42 Un experimento publicado en *Popular Science* compara las economías en combustible para dos tipos de camiones compactos a diesel equipados de forma similar. Supongamos que se utilizaron 12 camiones Volkswagen y 10 Toyota en pruebas de velocidad constante de 90 kilómetros por hora. Si los 12 camiones Volkswagen promedian 16 kilómetros por litro con una desviación estándar de 1.0 kilómetro por litro y los 10 Toyota promedian 11 kilómetros por litro con una desviación estándar de 0.8 kilómetros por litro, construya un intervalo de confianza de 90% para la diferencia entre los kilómetros promedio por litro de estos dos camiones compactos. Suponga que las distancias por litro para cada modelo de camión están distribuidas de forma aproximadamente normal con varianzas iguales.

9.43 Una compañía de taxis trata de decidir si comprar neumáticos de la marca A o de la B para su flotilla de taxis. Para estimar la diferencia de las dos marcas, se lleva a cabo un experimento utilizando 12 neumáticos de cada marca. Los neumáticos se utilizan hasta que se desgastan. Los resultados son

$$\begin{aligned} \text{Marca A: } \bar{x}_1 &= 36{,}300 \text{ kilómetros},\\ s_1 &= 5{,}000 \text{ kilómetros}.\\ \text{Marca B: } \bar{x}_2 &= 38{,}100 \text{ kilómetros},\\ s_2 &= 6{,}100 \text{ kilómetros}. \end{aligned}$$

Calcule un intervalo de confianza de 95% para $\mu_A - \mu_B$, suponiendo que las poblaciones se distribuyen de forma aproximadamente normal. Puede no suponer que las varianzas son iguales.

9.44 Con referencia al ejercicio 9.43, encuentre un intervalo de confianza de 99% para $\mu_1 - \mu_2$, si se asigna al azar un neumático de cada compañía a las ruedas traseras de 8 taxis y se registran las siguientes distancias, en kilómetros:

Taxi	**Marca *A***	**Marca *B***
1	34,400	36,700
2	45,500	46,800
3	36,700	37,700
4	32,000	31,100
5	48,400	47,800
6	32,800	36,400
7	38,100	38,900
8	30,100	31,500

Suponga que las diferencias de las distancias se distribuyen de forma aproximadamente normal.

9.45 El gobierno otorga fondos para los departamentos de agricultura de 9 universidades para probar las capacidades de rendimiento de dos nuevas variedades de trigo. Cada variedad se siembra en parcelas de área igual en cada universidad y el rendimiento, en kilogramos por parcela, se registra como sigue:

	Universidad								
Variedad	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	38	23	35	41	44	29	37	31	38
2	45	25	31	38	50	33	36	40	43

Encuentre un intervalo de confianza de 95% para la diferencia media entre los rendimientos de las dos variedades, suponiendo que las diferencias de rendimiento se distribuyen de forma aproximadamente normal. Explique por qué en este problema se necesita el pareamiento.

9.46 Los siguientes datos representan los tiempos de duración de las películas que producen dos compañías cinematográficas.

Compañía	**Tiempo (minutos)**						
I	103	94	110	87	98		
II	97	82	123	92	175	88	118

Calcule un intervalo de confianza de 90% para la diferencia entre los tiempos de duración promedio de las películas que producen las dos compañías. Suponga que las diferencias del tiempo de duración se distribuyen de forma aproximadamente normal con varianzas distintas.

9.47 A continuación se listan 10 de las 431 compañías estudiadas en la revista *Fortune* (marzo de 1997). Se listan las utilidades totales para los 10 años anteriores a 1996 y también para 1996. Encuentre un intervalo de confianza de 95% para el cambio promedio en el porcentaje de utilidad de los inversionistas.

	Utilidad total para los inversionistas	
Compañía	**1986–1996**	**1996**
Coca-Cola	29.8%	43.3%
Mirage Resorts	27.9%	25.4%
Merck	22.1%	24.0%
Microsoft	44.5%	88.3%
Johnson & Johnson	22.2%	18.1%
Intel	43.8%	131.2%
Pfizer	21.7%	34.0%
Procter & Gamble	21.9%	32.1%
Berkshire Hathaway	28.3%	6.2%
S&P 500	11.8%	20.3%

9.48 Una compañía automotriz considera dos tipos de baterías para sus vehículos. Se emplea la información muestral de la vida de las baterías. Se utilizan 20 baterías del tipo A y 20 baterías del tipo B. El extracto de los estadísticos es $\bar{x}_A = 32.91$, $\bar{x}_B = 30.47$, $s_A = 1.57$ y $s_B = 1.74$. Suponga que los datos de cada batería se distribuyen normalmente y que $\sigma_A = \sigma_B$.

a) Encuentre un intervalo de confianza de 95% para $\mu_A - \mu_B$.

b) A partir del inciso *a*) obtenga algunas conclusiones que ayuden a decidir si se debería adoptar A o B.

9.49 Se considera usar dos marcas diferentes de pintura látex. El tiempo de secado en horas se mide en especímenes de muestras del uso de las dos pinturas. Se seleccionan 15 especímenes de cada una y los tiempos de secado son los siguientes:

Pintura *A*					**Pintura *B***				
3.5	2.7	3.9	4.2	3.6	4.7	3.9	4.5	5.5	4.0
2.7	3.3	5.2	4.2	2.9	5.3	4.3	6.0	5.2	3.7
4.4	5.2	4.0	4.1	3.4	5.5	6.2	5.1	5.4	4.8

Suponga que el tiempo de secado se distribuye normalmente con $\sigma_A = \sigma_B$. Encuentre un intervalo de confianza en $\mu_B - \mu_A$, donde μ_A y μ_B sean los tiempos medios de secado.

9.50 Dos niveles (alto y bajo) de dosis de insulina se suministran a dos grupos de ratas diabéticas para verificar la capacidad de fijación de la insulina. Se obtuvieron los siguientes datos.

Dosis baja:	$n_1 = 8$	$\bar{x}_1 = 1.98$	$s_1 = .51$
Dosis alta:	$n_2 = 13$	$\bar{x}_2 = 1.30$	$s_2 = 0.35$

Suponga que ambas varianzas son iguales. Determine un intervalo de confianza de 95% para la diferencia de la capacidad promedio real para fijar la insulina entre las dos muestras.

9.10 Una sola muestra: Estimación de una proporción

Un estimador puntual de la proporción p en un experimento binomial está dado por el estadístico $\hat{P} = X/n$, donde X representa el número de éxitos en n pruebas. Por lo tanto, la proporción de la muestra $\hat{p} = x/n$ se utilizará como el estimador puntual del parámetro p.

Si no se espera que la proporción p desconocida esté demasiado cerca de cero o de 1, podemos establecer un intervalo de confianza para p al considerar la distribución muestral de $\hat{P}$. Al designar un fracaso en cada prueba binomial con el valor 0 y un éxito con el valor 1, el número de éxitos, x, se puede interpretar como la suma de n valores que consisten sólo de ceros y unos, y $\hat{p}$ es sólo la media muestral de estos n valores. De aquí, por el teorema del límite central, para n suficientemente grande, $\hat{P}$ está distribuida de forma aproximadamente normal con media

$$\mu_{\hat{P}} = E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

y varianza

$$\sigma^2_{\hat{P}} = \sigma^2_{X/n} = \frac{\sigma^2_X}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}.$$

Por lo tanto, podemos asegurar que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

$$Z = \frac{\hat{P} - p}{\sqrt{pq/n}},$$

y $z_{\alpha/2}$ es el valor de la curva normal estándar sobre la cual encontramos un área de $\alpha/2$. Al sustituir para Z, escribimos

$$P\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Al multiplicar cada término de la desigualdad por $\sqrt{pq/n}$, y después restar $\hat{P}$ y multiplicar por -1, obtenemos

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{pq}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right) = 1 - \alpha.$$

Es difícil manipular las desigualdades de manera que se obtenga un intervalo aleatorio cuyos puntos extremos sean independientes de p, el parámetro desconocido. Cuando n es grande, se introducen errores muy pequeños al sustituir la estimación puntual $\hat{p} = x/n$ por p bajo el signo del radical. Entonces podemos escribir

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) \approx 1 - \alpha.$$

Para nuestra muestra aleatoria particular de tamaño n, se calcula la proporción muestral $\hat{p} = x/n$ y se obtiene el siguiente intervalo de confianza de$(1 - \alpha)100\%$ aproximado para p.

Intervalo de confianza de p de una muestra grande

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, y $\hat{q} = 1 - \hat{p}$, un intervalo de confianza aproximado de $(1 - \alpha)100\%$ para el parámetro binomial p esté dado por

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}},$$

donde $z_{\alpha/2}$ es el valor z que deja un área de $\alpha/2$ a la derecha.

Cuando n es pequeña y la proporción desconocida p se considera cercana a 0 o a 1, el procedimiento del intervalo de confianza que se establece aquí no es confiable y, por lo tanto, no se debería emplear. Para estar seguro, se requiere que tanto $n\hat{p}$ como $n\hat{q}$ sean mayores que o iguales a 5. El método para encontrar un intervalo de confianza para el parámetro binomial p también se aplica cuando la distribución binomial se utiliza para aproximar la distribución hipergeométrica; es decir, cuando n es pequeña en relación con N, como se ilustra en el ejemplo 9.13.

Ejemplo 9.13: En una muestra aleatoria de $n = 500$ familias que tienen televisores en la ciudad de Hamilton, Canadá, se encuentra que $x = 340$ están suscritas a HBO. Encuentre un intervalo de confianza de 95% para la proporción real de familias en esta ciudad que están suscritas a HBO.

Solución: La estimación puntual de p es $\hat{p} = 340/500 = 0.68$. Con la tabla A.3, encontramos que $z_{0.025} = 1.96$. Por lo tanto, el intervalo de confianza de 95 % para p es

$$0.68 - 1.96\sqrt{\frac{(0.68)(0.32)}{500}} < p < 0.68 + 1.96\sqrt{\frac{(0.68)(0.32)}{500}},$$

que se simplifica a $0.64 < p < 0.72$.

Si p es el valor central de un intervalo de confianza de $(1 - \alpha)100\%$, entonces $\hat{p}$ estima p sin error. La mayoría de las veces, sin embargo, $\hat{p}$ no será exactamente igual a p y la estimación puntual será errónea. El tamaño de este error será la diferencia

positiva que separa a p y $\hat{p}$, y podemos tener una confianza de $(1-\alpha)100\%$ de que tal diferencia no excederá $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$. Podemos ver esto fácilmente si dibujamos un diagrama de un intervalo de confianza típico como en la figura 9.6.

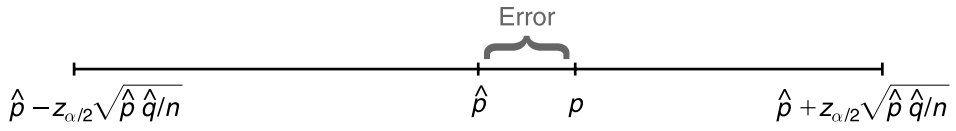


Figura 9.6: Error en la estimación de p por $\hat{p}$.

Teorema 9.3: Si $\hat{p}$ se utiliza como una estimación de p, podemos tener una confianza de $(1-\alpha)100\%$ de que el error no excederá $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$.

En el ejemplo 9.10 tenemos una confianza de 95% de que la proporción de la muestra $\hat{p} = 0.68$ difiere de la proporción real p en una cantidad que no excede 0.04.

Selección del tamaño de la muestra

Determinemos ahora qué tan grande se requiere que sea una muestra, para asegurar que el error al estimar p sea menor que una cantidad específica e. Por el teorema 7.3, esto significa que debemos elegir n de manera que $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} = e$.

Teorema 9.4: Si $\hat{p}$ se utiliza como estimación de p, podemos tener una confianza de $(1-\alpha)100\%$ de que el error será menor que una cantidad específica e cuando el tamaño de la muestra sea aproximadamente

$$n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}.$$

El teorema 9.4 es algo engañoso, pues debemos utilizar $\hat{p}$ para determinar el tamaño n de la muestra; pero $\hat{p}$ se calcula a partir de la muestra. Si se puede hacer una estimación cruda de p sin tomar una muestra, podríamos usar este valor para determinar n. A falta de tal estimación, podríamos tomar una muestra preliminar de tamaño $n \geq 30$ para proporcionar una estimación de p. Después, usando el teorema 9.4 podríamos determinar de forma aproximada cuántas observaciones se necesitan para brindar el grado de precisión que se desea. Observe que los valores fraccionarios de n se redondean (hacia arriba) al siguiente número entero.

Ejemplo 91:4 ¿Qué tan grande se requiere que sea una muestra en el ejemplo 9.13 si queremos tener 95% de confianza de que nuestra estimación de p esté dentro de 0.02?

Solución: Tratemos a las 500 familias como una muestra preliminar que proporciona una estimación $\hat{p} = 0.68$. Entonces, por el teorema 9.4,

$$n = \frac{(1.96)^2(0.68)(0.32)}{(0.02)^2} = 2089.8 \approx 2090.$$

Por lo tanto, si basamos nuestra estimación de p sobre una muestra aleatoria de tamaño 2090, podemos tener una confianza de 95% de que nuestra proporción muestral no diferirá de la proporción real en más de 0.02.

De cuando en cuando será poco práctico obtener una estimación de p que se utilice para determinar el tamaño muestral para un grado específico de confianza. Si esto sucede, se establece un límite superior para n al notar que $\hat{p}\hat{q} = \hat{p}(1 - \hat{p})$ que debe ser a lo más igual a $1/4$, ya que $\hat{p}$ debe estar entre 0 y 1. Este hecho se verifica al completar cuadrados. Por lo tanto,

$$\hat{p}(1 - \hat{p}) = -(\hat{p}^2 - \hat{p}) = \frac{1}{4} - (\hat{p}^2 - \hat{p} + \frac{1}{4}) = \frac{1}{4} - \left(\hat{p} - \frac{1}{2}\right)^2,$$

que siempre es menor que $1/4$ excepto cuando $\hat{p} = 1/2$ y entonces $\hat{p}\hat{q} = 1/4$. Entonces, si sustituimos $\hat{p} = 1/2$ en la fórmula para n del teorema 9.4, cuando, de hecho, p realmente difiere de $1/2$, entonces n se hará más grande de lo necesario para el grado de confianza específico y, como resultado, se incrementará nuestro grado de confianza.

Teorema 9.5: Si $\hat{p}$ se utiliza como estimación de p, podemos tener una confianza de **al menos** $(1 - \alpha)100\%$ de que el error no excederá una cantidad específica e cuando el tamaño de la muestra sea

$$n = \frac{z^2_{\alpha/2}}{4e^2}.$$

Ejemplo 9.15: ¿Qué tan grande se requiere que sea la muestra en el ejemplo 9.13, si queremos tener una confianza de al menos 95% de que nuestra estimación de p esté dentro de 0.02?

Solución: A diferencia del ejemplo 9.14, supondremos ahora que no se tomó una muestra preliminar para tener una estimación de p. En consecuencia, podemos tener al menos una confianza de 95% de que nuestra proporción de la muestra no diferirá de la proporción real en más de 0.02, si elegimos una muestra de tamaño

$$n = \frac{(1.96)^2}{(4)(0.02)^2} = 2401.$$

Al comparar los resultados de los ejemplos 9.14 y 9.15, la información con respecto a p, proporcionada por una muestra preliminar o quizás a partir de la experiencia pasada, nos permite elegir una muestra más pequeña, a la vez que mantenemos nuestro grado de precisión requerido.

9.11 Dos muestras: Estimación de la diferencia entre dos proporciones

Considere el problema donde deseamos estimar la diferencia entre dos parámetros binomiales p_1 y p_2. Por ejemplo, podríamos hacer que p_1 sea la proporción de fumadores con cáncer pulmonar y p_2 la proporción de no fumadores con cáncer pulmonar. Nuestro problema, entonces, consiste en estimar la diferencia entre estas dos proporciones. Primero, seleccionamos muestras aleatorias independientes de tamaños n_1 y n_2 a partir de las dos poblaciones binomiales con medias n_1p_1 y n_2p_2 y varianzas

$n_1p_1q_1$ y $n_2p_2q_2$, respectivamente, después determinamos los números x_1 y x_2 de personas con cáncer pulmonar en cada muestra, y formamos las proporciones $\hat{p} = x_1/n$ y $\hat{p} = x_2/n$. Un estimador puntual de la diferencia entre las dos proporciones, $p_1 - p_2$ está dado por el estadístico $\hat{P}_1 - \hat{P}_2$. Por lo tanto, la diferencia de las proporciones muestrales, $\hat{p}_1 - \hat{p}_2$, se utilizará como la estimación puntual de $p_1 - p_2$.

Se puede establecer un intervalo de confianza para $p_1 - p_2$ al considerar la distribución muestral de $\hat{P}_1 - \hat{P}_2$. De la sección 9.10 sabemos que $\hat{P}_1$ y $\hat{P}_2$ están distribuidos cada uno de forma aproximadamente normal, con medias p_1 y p_2, y varianzas p_1q_1/n_1 y p_2q_2/n_2, respectivamente. Al elegir muestras independientes de las dos poblaciones, las variables $\hat{P}_1$ y $\hat{P}_2$ serán independientes y, por ello, por la propiedad reproductiva de la distribución normal que se estableció en el teorema 7.11, concluimos que $\hat{P}_1 - \hat{P}_2$ está distribuida de forma aproximadamente normal con media

$$\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$$

y varianza

$$\sigma^2_{\hat{P}_1 - \hat{P}_2} = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}.$$

Por lo tanto, podemos asegurar que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1q_1/n_1 + p_2q_2/n_2}},$$

$z_{\alpha/2}$ es un valor de la curva normal estándar sobre la cual encontramos un área de $\alpha/2$. Al sustituir para Z, escribimos

$$P\left[-z_{\alpha/2} < \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1q_1/n_1 + p_2q_2/n_2}} < z_{\alpha/2}\right] = 1 - \alpha.$$

Después de realizar las manipulaciones matemáticas usuales, reemplazamos p_1, p_2, q_1 y q_2 bajo el signo del radical por sus estimaciones $\hat{p}_1 = x_1/n_1$, $\hat{p}_2 = x_2/n_2$, $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$, dado que $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$ y $n_2\hat{q}_2$ son todas mayores que o iguales a 5, y se obtiene el siguiente intervalo de confianza de $(1 - \alpha)100\%$ aproximado para $p_1 - p_2$.

Intervalo de confianza de $p_1 - p_2$ de una muestra grande

Si $\hat{p}_1$ y $\hat{p}_2$ son las proporciones de éxitos en muestras aleatorias de tamaño n_1 y n_2, respectivamente, $\hat{q}_1 = 1 - \hat{p}_1$, y $\hat{q}_2 = 1 - \hat{p}_2$, un intervalo de confianza aproximado de $(1 - \alpha)100\%$ para la diferencia de dos parámetros binomiales $p_1 - p_2$, está dado por

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \quad &< \quad p_1 - p_2 \\ &< \quad (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},\end{aligned}$$

donde $z_{\alpha/2}$ es el valor z que deja un área de $\alpha/2$ a la derecha.

Ejemplo 9.16: Se considera cierto cambio en un proceso de fabricación de partes componentes. Se toman muestras del procedimiento actual y del nuevo, para determinar si el nuevo tiene como resultado una mejoría. Si se encuentra que 75 de 1500 artículos del procedimiento actual son defectuosos y 80 de 2000 artículos del procedimiento nuevo también lo son, encuentre un intervalo de confianza de 90% para la diferencia real en la fracción de defectuosos entre el proceso actual y el nuevo.

Solución: Sean p_1 y p_2 las proporciones reales de defectuosos para los procedimientos actual y nuevo, respectivamente. De aquí, $\hat{p}_1 = 75/1500 = 0.05$ y $\hat{p}_2 = 80/2000 = 0.04$, y la estimación puntual de $p_1 - p_2$ es

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04 = 0.01.$$

Con la tabla A.3, encontramos $z_{0.05} = 1.645$. Por lo tanto, al sustituir en esta fórmula

$$1.645\sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}} = 0.0117,$$

obtenemos el intervalo de confianza de 90% que se simplifica $-0.0017 < p_1 - p_2 < 0.0217$. Como el intervalo contiene el valor 0, no hay razón para creer que el nuevo procedimiento resultó en una disminución significativa en la proporción de artículos defectuosos, comparado con el método actual.

Hasta aquí, todos los intervalos de confianza presentados son de la forma

$$\text{estimación puntual} \pm K \text{ e.e.(estimación puntual)},$$

donde K es una constante (ya sea t o el punto porcentual normal). Éste es el caso cuando el parámetro es una media, diferencia entre medias, proporción o diferencia entre proporciones, debido a la simetría de las distribuciones t y Z. Sin embargo, ello no se extiende a las varianzas ni a las razones de varianzas que se examinarán en las secciones 9.12 y 9.13.

Ejercicios

9.51 *a*) Se selecciona una muestra aleatoria de 200 votantes y se encuentra que 114 apoyan un juicio de anexión. Encuentre el intervalo de confianza de 96% para la fracción de la población votante que favorece el juicio.

b) ¿Qué podemos asegurar con 96% de confianza acerca de la posible magnitud de nuestro error, si estimamos que la fracción de votantes que favorecen el juicio de anexión es 0.57?

9.52 Un fabricante de reproductores de discos compactos utiliza un conjunto de pruebas amplias para evaluar la función eléctrica de su producto. Todos los reproductores de discos compactos deben pasar todas las pruebas antes de venderse. Una muestra aleatoria de 500 reproductores tiene como resultado 15 que fallan en una o más de las pruebas. Encuentre un intervalo de confianza de 90% para la proporción de los reproductores de discos compactos de la población que pasan todas las pruebas.

9.53 En una muestra aleatoria de 1000 viviendas en cierta ciudad, se encuentra que 228 se calientan con petróleo. Encuentre el intervalo de confianza de 99% para la proporción de viviendas en esta ciudad que se calientan con petróleo.

9.54 Calcule un intervalo de confianza de 98% para la proporción de artículos defectuosos en un proceso cuando se encuentra que una muestra de tamaño 100 da como resultado 8 defectuosos.

9.55 Se considera un nuevo sistema de lanzamiento de cohetes para el despliegue de cohetes pequeños de corto alcance. El sistema existente tiene $p = 0.8$ como la probabilidad de lanzamiento exitoso. Se realiza una muestra de 40 lanzamientos experimentales con el nuevo sistema y 34 resultan exitosos.

a) Construya un intervalo de confianza de 95% para p.

b) ¿Concluiría que es mejor el nuevo sistema?

9.56 Un genetista se interesa en la proporción de hombres africanos que tienen cierto trastorno sanguíneo menor. En una muestra aleatoria de 100 hombres africanos, se encuentra que 24 lo padecen.

a) Calcule un intervalo de confianza de 99% para la proporción de hombres africanos que tienen este trastorno sanguíneo.

b) ¿Qué se puede asegurar con 99% de confianza acerca de la posible magnitud de nuestro error, si estimamos que la proporción de hombres africanos con dicho trastorno sanguíneo es 0.24?

9.57 *a*) De acuerdo con un reporte del *Roanoke Times & World-News*, aproximadamente 2/3 de los 1600 adultos encuestados vía telefónica dijeron que piensan que el programa del trasbordador espacial es una buena inversión para el país. Encuentre un intervalo de confianza de 95% para la proporción de adultos estadounidenses que piensan que el programa del trasbordador espacial es una buena inversión para el país.

b) ¿Qué podemos asegurar con una confianza de 95% acerca de la posible magnitud de nuestro error, si estimamos que la proporción de adultos estadounidenses que piensan que el programa del trasbordador espacial es una buena inversión de 2/3?

9.58 En el artículo del periódico al que se hace referencia en el ejercicio 9.57, 32% de los 1600 adultos encuestados dijeron que el programa espacial estadounidense debería enfatizar la exploración científica. ¿Qué tan grande se necesita que sea una muestra de adultos en la encuesta si se desea tener una confianza de 95% de que el porcentaje estimado esté dentro de 2% del porcentaje real?

9.59 ¿Qué tan grande se requiere que sea la muestra en el ejercicio 9.51 si deseamos tener una confianza de 96% de que nuestra proporción de la muestra estará dentro del 0.02 de la fracción real de la población votante?

9.60 ¿Qué tan grande se requiere que sea la muestra en el ejercicio 9.53, si deseamos tener una confianza de 99% de que nuestra proporción de la muestra estará dentro del 0.05 de la proporción real de casas en esta ciudad que se calientan con petróleo?

9.61 ¿Qué tan grande se necesita la muestra en el ejercicio 9.54, si deseamos tener una confianza de 98% de que nuestra proporción de la muestra esté dentro del 0.05 de la proporción real de defectuosos?

9.62 Se lleva a cabo un estudio para estimar el porcentaje de ciudadanos de una ciudad que están a favor de tener su agua fluorada. ¿Qué tan grande se requiere que sea la muestra si se desea tener al menos una confianza de 95% de que nuestra estimación esté dentro del 1% del porcentaje real?

9.63 La conjetura de un miembro del profesorado del departamento de microbiología de la Escuela de Odontología de la Universidad de Washington, en St. Louis, afirma que un par de tasas diarias de té verde o negro proporcionan suficiente flúor para evitar caries en los dientes. ¿Qué tan grande se requiere que sea la muestra para estimar el porcentaje de habitantes de cierta ciudad que están a favor de tener su agua fluorada, si se desea tener al menos el 99% de confianza de que la estimación está dentro del 1% del porcentaje real?

9.64 Se lleva a cabo un estudio para estimar la proporción de residentes de cierta ciudad y sus suburbios que están a favor de la construcción de una planta de energía nuclear. ¿Qué tan grande se requiere que sea la muestra, si se desea tener al menos 95% de confianza de que la estimación está dentro del 0.04 de la proporción real de residentes de esta ciudad y sus suburbios, que están a favor de la construcción de la planta de energía nuclear?

9.65 Cierto genetista se interesa en la proporción de hombres y mujeres en la población que tienen cierto trastorno sanguíneo menor. En una muestra aleatoria de 1000 hombres se encuentra que 250 lo padecen; mientras que 275 de 1000 mujeres examinadas parecen tener el trastorno. Calcule un intervalo de confianza de 95% para la diferencia entre la proporción de hombres y mujeres que padecen el trastorno sanguíneo.

9.66 Se encuestan 10 escuelas de ingeniería en Estados Unidos. La muestra contiene 250 ingenieros eléctricos, donde 80 son mujeres; y 175 ingenieros químicos, donde 40 son mujeres. Calcule un intervalo de confianza de 90% para la diferencia entre la proporción de mujeres en estos dos campos de la ingeniería. ¿Hay una diferencia significativa entre las dos proporciones?

9.67 Se lleva a cabo una prueba clínica para determinar si cierto tipo de inoculación tiene un efecto sobre la incidencia de cierta enfermedad. Una muestra de 1000 ratas se mantiene en un ambiente controlado durante un periodo de un año y a 500 de éstas se les inoculó. Del grupo al que no se le dio el fármaco, hubo 120 incidencias de la enfermedad; mientras que 98 del grupo inoculado la contrajeron. Si p_1 es la probabilidad de incidencia de la enfermedad en las ratas no inoculadas y p_2 es la probabilidad de incidencia después de recibir el fármaco, calcule un intervalo de confianza de 90% para $p_1 - p_2$.

9.68 En un estudio, *Germination and Emergence of Broccoli*, que lleva a cabo el Departamento de Horticultura del Instituto Politécnico y Universidad Estatal

de Virginia, un investigador encuentra que a 5 °C, 10 semillas de 20 germinaron; en tanto que a 15 °C, 15 semillas de 20 lo hicieron. Calcule un intervalo de confianza de 95% para la diferencia entre la proporción de germinación en las dos diferentes temperaturas, y decida si hay una diferencia significativa.

9.69 Una encuesta a 1000 estudiantes concluye que 274 eligen al equipo profesional de béisbol A como su equipo favorito. En 1991, se realizó la misma encuesta con 760 estudiantes. Concluyó que 240 de ellos también eligieron al equipo A como su favorito. Calcule un intervalo de confianza de 95% para la diferencia entre la proporción de estudiantes que favorecen al equipo A entre las dos encuestas. ¿Hay una diferencia significativa?

9.70 De acuerdo con *USA Today* (17 de marzo de 1997), las mujeres constituían 33.7% del equipo de redacción en las estaciones locales de televisión en 1990, y 36.2% en 1994. Suponga que se contrataron 20 nuevos empleados para el equipo de redacción.

a) Estime el número que habrían sido mujeres en cada año, respectivamente.

b) Calcule un intervalo de confianza de 95%, para saber si hay evidencia de que la proporción de mujeres contratadas para el equipo de redacción en 1994 fue mayor que la proporción contratada en 1990.

9.12 Una sola muestra: Estimación de la varianza

Si se extrae una muestra de tamaño n de una población normal con varianza σ^2 y se calcula la varianza muestral s^2, obtenemos un valor del estadístico S^2 . Esta varianza muestral calculada se usará como estimación puntual de σ^2. Por ello, el estadístico S^2 se llama estimador de σ^2.

Se puede establecer una estimación por intervalos de σ^2 utilizando el estadístico

$$X^2 = \frac{(n-1)S^2}{\sigma^2}.$$

De acuerdo con el teorema 8.4, el estadístico X^2 tiene una distribución chi cuadrada con $n - 1$ grados de libertad, cuando las muestras se eligen de una población normal. Podemos escribir (véase la figura 9.7)

$$P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha,$$

donde $\chi^2_{1-\alpha/2}$ y $\chi^2_{\alpha/2}$ son valores de la distribución chi cuadrada con $n - 1$ grados de libertad, que dejan áreas de $1 - \alpha/2$ y $\alpha/2$, respectivamente, a la derecha. Al sustituir para X^2, escribimos

$$P\left[\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right] = 1 - \alpha.$$

Al dividir cada término de la desigualdad entre $(n - 1)S^2$ y, después, invertir cada término (lo que cambia el sentido de las desigualdades), obtenemos

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1 - \alpha.$$

Para nuestra muestra aleatoria particular de tamaño n, se calcula la varianza muestral s^2 y se obtiene el siguiente intervalo de confianza de $(1 - \alpha)100\%$ para σ^2.

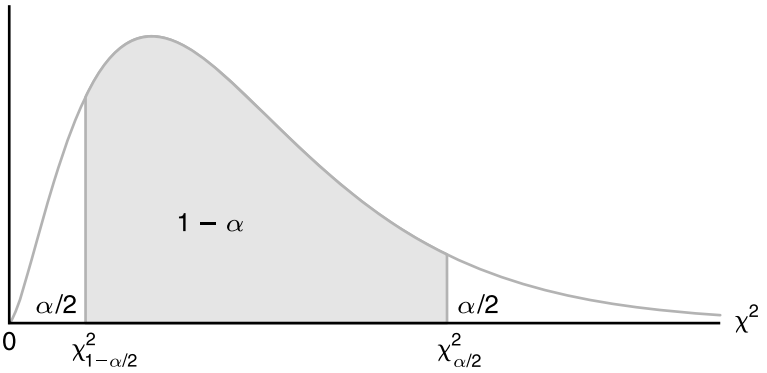


Figura 9.7: $P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha$.

Intervalo de confianza para σ^2

Si s^2 es la varianza de una muestra aleatoria de tamaño n de una población normal, un intervalo de confianza de $(1 - \alpha)100\%$ para σ^2 es

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

donde $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ son valores χ^2 con $v = n - 1$ grados de libertad, que dejan áreas de $\alpha/2$ y $1 - \alpha/2$, respectivamente, a la derecha.

Un intervalo de confianza de $(1 - \alpha)100\%$ para σ se obtiene al tomar la raíz cuadrada de cada extremo del intervalo para σ^2.

Ejemplo 9.17: Los siguientes son los pesos, en decagramos, de 10 paquetes de semillas para césped distribuidas por cierta compañía: 46.4, 46.1, 45.8, 47.0, 46.1. 45.9, 45.8, 46.9, 45.2 y 46.0. Encuentre un intervalo de confianza de 95% para la varianza de todos los paquetes de semillas para césped que distribuye esta compañía. Suponga una población normal.

Solución: Primero, encontramos

$$\begin{aligned} s^2 &= \frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)} \\ &= \frac{(10)(21,273.12) - (461.2)^2}{(10)(9)} = 0.286. \end{aligned}$$

Para obtener un intervalo de confianza de 95%, elegimos $\alpha = 0.05$. Después, usando la tabla A.5 con $v = 9$ grados de libertad, encontramos $\chi^2_{0.025} = 19.023$ y $\chi^2_{0.075} = 2.700$. Por lo tanto, el intervalo de confianza de 95% para σ^2 es

$$\frac{(9)(0.286)}{19.023} < \sigma^2 < \frac{(9)(0.286)}{2.700},$$

o simplemente $0.135 < \sigma^2 < 0.953$.

9.13 Dos muestras: Estimación de la razón de dos varianzas

Una estimación puntual de la razón de dos varianzas poblacionales σ_1^2/σ_2^2 está dada por la razón s_1^2/s_2^2 de las varianzas muestrales. De aquí, el estadístico S_1^2/S_2^2 se denomina estimador de σ_1^2/σ_2^2.

Si $\sigma^2{}_1$ y $\sigma^2{}_2$ son las varianzas de poblaciones normales, podemos establecer una estimación por intervalos de σ_1^2/σ_2^2 usando el estadístico

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

De acuerdo con el teorema 8.8, la variable aleatoria F tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Por lo tanto, podemos escribir (véase la figura 9.8)

$$P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha,$$

donde $f_{1-\alpha/2}(v_1,\, v_2)$ y $f_{\alpha/2}(v_1,\, v_2)$ son los valores de la distribución F con v_1 y v_2 grados de libertad, que dejan áreas de $1 - \alpha/2$ y $\alpha/2$, respectivamente, a la derecha. Al sustituir para F, escribimos

$$P\left[f_{1-\alpha/2}(v_1, v_2) < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(v_1, v_2)\right] = 1 - \alpha.$$

Al multiplicar cada término en la desigualdad por S_2^2/S_1^2, y después invertir cada término (de nuevo para cambiar el sentido de las desigualdades), obtenemos

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha/2}(v_1, v_2)}\right] = 1 - \alpha.$$

Los resultados del teorema 8.7 nos permiten reemplazar la cantidad $f_{1-\alpha/2}(v_1,\, v_2)$ por $1/f_{\alpha/2}(v_1,\, v_2)$. Por lo tanto,

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(v_2, v_1)\right] = 1 - \alpha.$$

Para cualesquiera dos muestras aleatorias independientes de tamaño n_1 y n_2 que se seleccionan de dos poblaciones normales, la razón de las varianzas muestrales s_1^2/s_2^2 se calcula y se obtiene el siguiente intervalo de confianza de $(1 - \alpha)100\%$ para σ_1^2/σ_2^2.

Intervalo de confianza para σ_1^2/σ_2^2

Si s_1^2 y s_2^2 son las varianzas de muestras independientes de tamaño n_1 y n_2, respectivamente, de poblaciones normales, entonces un intervalo de confianza de $(1 - \alpha)100\%$ para σ_1^2/σ_2^2 es

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(v_2, v_1),$$

donde $f_{\alpha/2}\,(v_1,\, v_2)$ es un valor f con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad que deja un área de $\alpha/2$ a la derecha, y $f_{\alpha/2}\,(v_2,\, v_1)$ es un valor f similar con $v_2 = n_2 - 1$ y $v_1 = n_1 - 1$ grados de libertad.

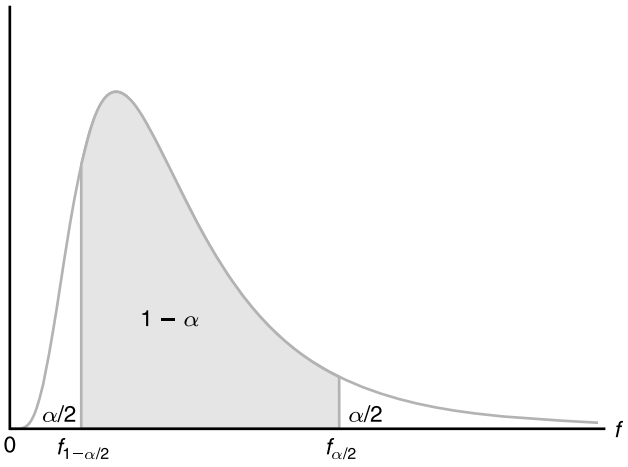


Figura 9.8: $P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha$.

Como en la sección 9.12, se obtiene un intervalo de confianza de $(1 - \alpha)100\%$ para σ_1/σ_2 al tomar la raíz cuadrada de cada extremo del intervalo para σ_1^2/σ_2^2.

Ejemplo 9.18: En el ejemplo 9.11 de la página 293 se construyó un intervalo de confianza para la diferencia en el contenido medio de ortofósforo, que se mide en miligramos por litro, en dos estaciones sobre el río James, suponiendo que las varianzas normales de la población son diferentes. Justifique esta suposición mediante la construcción de un intervalo de confianza de 98% para σ_1^2/σ_2^2 y para σ_1/σ_2 donde σ_1^2 y σ_2^2 son las varianzas poblacionales del contenido de ortofósforo en la estación 1 y en la estación 2, respectivamente.

Solución: Del ejemplo 9.11, tenemos $n_1 = 15$, $n_2 = 12$, $s_1 = 3.07$ y $s_2 = 0.80$. Para un intervalo de confianza de 98%, $\alpha = 0.02$. Al interpolar en la tabla A.6, encontramos $f_{0.01}(14,11) \approx 4.30$ y $f_{0.01}(11,14) \approx 3.87$. Por lo tanto, el intervalo de confianza de 98% para σ_1^2/σ_2^2 es

$$\frac{3.07^2}{0.80^2}\left(\frac{1}{4.30}\right) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{3.07^2}{0.80^2}(3.87),$$

que se simplifica a $3.425 < \frac{\sigma_1^2}{\sigma_2^2} < 56.991$. Al calcular las raíces cuadradas de los límites de confianza, encontramos que un intervalo de confianza de 98% para σ_1/σ_2 es

$$1.851 < \frac{\sigma_1}{\sigma_2} < 7.549.$$

Como este intervalo no permite la posibilidad de que σ_1/σ_2 sea igual a 1, es correcto suponer que $\sigma_1 \neq \sigma_2$ o $\sigma_1^2 \neq \sigma_2^2$ en el ejemplo 9.11.

Ejercicios

9.71 Un fabricante de baterías para automóvil afirma que sus baterías durarán, en promedio, 3 años con una varianza de 1 año. Si 5 de estas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años, construya un intervalo de confianza de 95% para σ^2 y decida si es válida la afirmación del fabricante de que $\sigma^2 = 1$. Suponga que la población de duraciones de las baterías se distribuye de forma aproximadamente normal.

9.72 Una muestra aleatoria de 20 estudiantes obtuvo una media de $\bar{x} = 72$ y una varianza de $s^2 = 16$ en un examen universitario de colocación en matemáticas. Suponga que las calificaciones se distribuyen normalmente y construya un intervalo de confianza de 98% para σ^2.

9.73 Construya un intervalo de confianza de 95% para σ^2 en el ejercicio 9.12 de la página 286.

9.74 Construya un intervalo de confianza de 99% para σ^2 en el ejercicio 9.13 de la página 286.

9.75 Construya un intervalo de confianza de 99% para σ en el ejercicio 9.14 de la página 286.

9.76 Construya un intervalo de confianza de 90% para σ en el ejercicio 9.15 de la página 286.

9.77 Construya un intervalo de confianza de 98% para σ_1/σ_2 en el ejercicio 9.42 de la página 298, donde σ_1 y σ_2 son, respectivamente, las desviaciones estándar para las distancias que se obtienen por litro de combustible en los camiones compactos Volkswagen y Toyota.

9.78 Construya un intervalo de confianza de 90% para σ_1^2/σ_2^2 en el ejercicio 9.43 de la página 298. ¿Estamos justificados al suponer que $\sigma_1^2 \neq \sigma_2^2$ cuando construimos nuestro intervalo de confianza para $\mu_1 - \mu_2$?

9.79 Construya un intervalo de confianza de 90% para σ_1^2/σ_2^2 en el ejercicio 9.46 de la página 298. ¿Deberíamos suponer $\sigma_1^2 = \sigma_2^2$ al construir nuestro intervalo de confianza para $\mu_I - \mu_{II}$?

9.80 Construya un intervalo de confianza de 95% para σ_A^2/σ_B^2 en el ejercicio 9.49 de la página 299. ¿Debería utilizarse la suposición de la varianza igual?

9.14 Estimación de la probabilidad máxima (opcional)

A menudo los estimadores de parámetros son los que recurren a la intuición. El estimador $\bar{X}$ ciertamente parece razonable como estimador de una media poblacional μ. La virtud de S^2 como estimador de σ^2 se destaca en el estudio de estimadores insesgados de la sección 9.3. El estimador para un parámetro binomial p es simplemente una proporción de la muestra que, desde luego, es un *promedio* y recurre al sentido común. Sin embargo, hay muchas situaciones en las cuales no es del todo evidente cuál debería ser el estimador adecuado. Como resultado, el estudiante de estadística tiene mucho por aprender con respecto a las diferentes filosofías que producen diversos métodos de estimación. En esta sección estudiaremos el **método de probabilidad máxima**.

La estimación por probabilidad máxima representa una de las aproximaciones a la estimación más importantes en toda la inferencia estadística. No haremos un desarrollo completo del método. En cambio, intentaremos comunicar la filosofía de la probabilidad máxima y la ilustraremos con ejemplos que la relacionan con otros problemas de estimación que se examinan en este capítulo.

Función de probabilidad

Como el nombre lo indica, el método de probabilidad máxima es aquel para el que se maximiza la *función de probabilidad*, la cual se ilustra mejor con un ejemplo de una distribución discreta y un solo parámetro. Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes tomadas de una distribución de probabilidad representada por $f(\boldsymbol{x}, \theta)$,

donde θ es un solo parámetro de la distribución. Ahora bien,

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \theta) &= f(x_1, x_2, \ldots, x_n; \theta) \\ &= f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) \end{aligned}$$

es la *distribución conjunta de las variables aleatorias*. Esto a menudo se denomina función de probabilidad. Observe que la variable de la función de probabilidad es θ, no $\boldsymbol{x}$. Sean $x_1, x_2, \ldots, x_n$ valores observados en una muestra. En el caso de una variable aleatoria discreta la interpretación es muy clara. La cantidad $L(x_1, x_2, \ldots, x_n; \theta)$, *la probabilidad de la muestra*, es la siguiente probabilidad conjunta:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | \theta),$$

que es la probabilidad de obtener los valores muestrales $x_1, x_2, \ldots, x_n$. Para el caso discreto el estimador de probabilidad máxima es el que tiene como resultado un valor máximo para esta probabilidad conjunta, o que maximiza la probabilidad de la muestra.

Considere un ejemplo ficticio donde se inspeccionan tres artículos que salen de una línea de ensamble. Los artículos se clasifican como defectuosos o no defectuosos, de manera que se aplica el proceso de Bernoulli. La inspección de los tres artículos tiene como resultado dos artículos no defectuosos seguidos por uno defectuoso. Es de interés estimar p, la proporción de no defectuosos en el proceso. La probabilidad de la muestra para esta ilustración está dada por

$$p \cdot p \cdot q = p^2 q = p^2 - p^3,$$

donde $q = 1 - p$. La estimación de probabilidad máxima daría una estimación de p para la que se maximiza la probabilidad. Resulta claro que si diferenciamos la probabilidad con respecto a p, hacemos la derivada igual a cero y la resolvemos, obtenemos el valor

$$\hat{p} = \frac{2}{3}.$$

Entonces, por supuesto, en esta situación $\hat{p} = 2/3$ es la proporción muestral de defectuosos y, por ello, un estimador razonable de la probabilidad de un defectuoso. El lector debería intentar comprender que la filosofía de la estimación de probabilidad máxima proviene de la noción de que el estimador razonable de un parámetro que se basa en información muestral *es el valor del parámetro que produce la mayor probabilidad de obtener la muestra*. Ésta es, de hecho, la interpretación para el caso discreto, pues se trata de la probabilidad de observar de manera conjunta los valores en la muestra.

Así, mientras que la interpretación de la función de probabilidad como una probabilidad conjunta se reduce al caso discreto, la noción de probabilidad máxima se extiende a la estimación de parámetros de una distribución continua. Presentamos ahora una definición formal de la estimación de probabilidad máxima.

Definición 9.3:

Dadas las observaciones independientes $x_1, x_2, \ldots, x_n$ de una función de densidad de probabilidad (caso continuo) o de una función de masa de probabilidad (caso discreto) $f(\boldsymbol{x}, \theta)$, el estimador de probabilidad máxima $\hat{\theta}$ es el que maximiza la función de probabilidad

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta).$$

Muy a menudo conviene trabajar con el logaritmo natural de la función de probabilidad para encontrar el máximo de ésta. Considere el siguiente ejemplo acerca del parámetro μ de una distribución de Poisson.

Ejemplo 9.19: Considere una distribución de Poisson con función de masa de probabilidad

$$f(x|\mu) = \frac{e^{-\mu}\mu^x}{x!} \quad x = 0, 1, 2, \ldots.$$

Suponga que se toma una muestra aleatoria $x_1, x_2, \ldots, x_n$ de la distribución. ¿Cuál es la estimación de probabilidad máxima de μ?

Solución: La función de probabilidad es

$$L(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} f(x_i|\mu) = \frac{e^{-n\mu}\mu^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

Considere ahora

$$\ln L(x_1, x_2, \ldots, x_n; \mu) = -n\mu + \sum_{i=1}^{n} x_i \ln \mu - \ln \prod_{i=1}^{n} x_i!$$

$$\frac{\partial \ln L(x_1, x_2, \ldots, x_n; \mu)}{\partial \mu} = -n + \sum_{i=1}^{n} \frac{x_i}{\mu}.$$

Al resolver para $\hat{\mu}$, el estimador de probabilidad máxima implica hacer la derivada igual a cero y resolver para el parámetro. De esta forma,

$$\hat{\mu} = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x}.$$

Como μ es la media de la distribución de Poisson (capítulo 5), el promedio muestral en realidad parecería ser un estimador razonable.

El siguiente ejemplo presenta el uso del método de probabilidad máxima para encontrar estimaciones de dos parámetros. Simplemente encontramos los valores de los parámetros que maximizan (de forma conjunta) la función de probabilidad.

Ejemplo 9.20: Considere una muestra aleatoria $x_1, x_2, \ldots, x_n$ de una distribución normal $N(\mu, \sigma)$. Encuentre los estimadores de probabilidad máxima para μ y σ^2.

Solución: La función de probabilidad para la distribución normal es

$$L(x_1, x_2, \ldots, x_n; \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right].$$

Al tomar logaritmos da

$$\ln L(x_1, x_2, \ldots, x_n; \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2.$$

Por lo tanto,

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma^2}\right)$$

y

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Al igualar la derivada a cero, obtenemos

$$\sum_{i=1}^{n} x_i - n\mu = 0 \quad \text{y} \quad n\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2.$$

De manera que el estimador de probabilidad máxima está dado por

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

este resultado es satisfactorio, ya que $\bar{x}$ juega un papel importante en este capítulo como estimador puntual de μ. Por otro lado, el estimador de probabilidad máxima de σ^2 es

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Al verificar la matriz derivada parcial de segundo orden se confirma que las soluciones que resultan en el máximo de la función de probabilidad.

Resulta interesante notar la distinción entre el estimador de probabilidad máxima de σ^2 y el estimador insesgado S^2 que se desarrolló al principio de este capítulo. El numerador es idéntico, por supuesto, y el denominador son los "grados de libertad" $n - 1$ para el estimador insesgado, y n para el estimador de probabilidad máxima. Los estimadores de probabilidad máxima no necesariamente gozan de la propiedad de estar insesgados. Sin embargo, los estimadores de probabilidad máxima tienen importantes propiedades asintóticas.

Ejemplo 9.21: Suponga que se utilizan 15 ratas en un estudio biomédico, donde a los roedores se les inyectan células cancerosas y se les suministra un fármaco contra el cáncer diseñado para aumentar su tasa de supervivencia. Los tiempos de supervivencia, en meses, son 14, 17, 27, 18, 12, 8, 22, 13, 19 y 12. Suponga que se aplica la distribución exponencial. Dé una estimación de probabilidad máxima de la supervivencia media.

Solución: Del capítulo 6 sabemos que la función de densidad de probabilidad para la variable aleatoria exponencial X es

$$f(x, \beta) = \begin{cases} \frac{1}{\beta}e^{-x \triangleleft \beta}, & x > 0, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Por ello, la probabilidad logarítmica de los datos dados ($n = 10$) es

$$\ln L(x_1, x_2, \ldots, x_{10}; \beta) \quad = -10 \ln \beta - \tfrac{1}{\beta}\sum_{i=1}^{10} x_i.$$

Al hacer

$$\frac{\partial \ln L}{\partial \beta} = -\frac{10}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{10} x_i = 0$$

implica que

$$\hat{\beta} = \frac{1}{10}\sum_{i=1}^{10} x_i = \bar{x} = 16.2.$$

La segunda derivada de la probabilidad logarítmica evaluada en el valor $\hat{\beta}$ anterior da como resultado un valor negativo. Como resultado el estimador del parámetro β, la media de la población, es el promedio muestral $\bar{x}$.

El siguiente ejemplo ilustra el estimador de probabilidad máxima para una distribución que no se incluye en los capítulos anteriores.

Ejemplo 9.22: Se sabe que una muestra de 12, 11.2, 13.5, 12.3, 13.8 y 11.9 se tomó de una población con la función de densidad

$$f(x;\theta) = \begin{cases} \frac{\theta}{x^{\theta+1}}, & x > 1, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

donde $\theta > 0$. Encuentre la estimación de probabilidad máxima de θ.

Solución: La función de probabilidad de n observaciones de esta población se escribe como

$$L(x_1, x_2, \ldots, x_{10}; \theta) = \prod_{i=1}^{n} \frac{\theta}{\boldsymbol{x}_i^{\theta+1}} = \frac{\theta^n}{(\prod_{i=1}^{n} x_i)^{\theta+1}},$$

lo cual implica que

$$\text{en } L(x_1, x_2, \ldots, x_{10}; \theta) = n\ln(\theta) - (\theta+1)\sum_{i=1}^{n} \ln(x_i).$$

Hacer $0 = \frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \ln(x_i)$ da como resultado

$$\begin{aligned} \hat{\theta} &= \frac{n}{\sum_{i=1}^{n} \ln(x_i)} \\ &= \frac{6}{\ln(12) + \ln(11.2) + \ln(13.5) + \ln(12.3) + \ln(13.8) + \ln(11.9)} = 0.3970. \end{aligned}$$

Como la segunda derivada de L es $-n/\theta^2$, que siempre es negativa, la función de probabilidad alcanza su valor máximo en $\hat{\theta}$.

Comentarios adicionales con respecto a la estimación por probabilidad máxima

Una discusión completa de las propiedades de la estimación por probabilidad máxima está fuera del alcance de este libro y, por lo general, es un tema principal de un curso teórico sobre inferencia estadística. El método de probabilidad máxima permite al analista utilizar el conocimiento de la distribución para determinar un estimador adecuado. *El método de probabilidad máxima no se puede aplicar sin el conocimiento de la distribución subyacente.* En el ejemplo 9.20 aprendimos que el estimador de probabilidad máxima no necesariamente está insesgado. El estimador de probabilidad máxima está insesgado *asintóticamente* o *en el límite*; es decir, la magnitud del sesgo se aproxima a cero conforme la muestra se hace más grande. Al principio de este capítulo examinamos la noción de eficacia, que se vincula con la propiedad de varianza de un estimador. Los estimadores de probabilidad máxima tienen propiedades de varianza deseables en el límite. El lector debería consultar la obra de Lehmann para más detalles.

Ejercicios

9.81 Suponga que hay n pruebas $x_1, x_2, \ldots, x_n$ de un proceso de Bernoulli con parámetro p, la probabilidad de un éxito. Es decir, la probabilidad de r éxitos está dada por $\binom{n}{r}p^r(1-p)^{n-r}$. Determine el estimador de probabilidad máxima para el parámetro p.

9.82 Considere una muestra de $x_1, x_2, \ldots, x_n$ observaciones de una distribución de Weibull con parámetros α y β y función de densidad

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1}e^{-\alpha x^\beta}, & x > 0, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

para $\alpha, \beta > 0$.

a) Escriba la función de probabilidad.

b) Escriba las ecuaciones que al resolverse dan los estimadores de probabilidad máxima de α y β.

9.83 Considere la distribución logarítmica normal con la función de densidad dada en la sección 6.9. Suponga que tenemos una muestra $x_1, x_2, \ldots, x_n$ de una distribución logarítmica normal.

a) Escriba la función de probabilidad.

b) Desarrolle los estimadores de probabilidad máxima de μ y σ^2.

9.84 Considere las observaciones $x_1, x_2, \ldots, x_n$ de la distribución gamma que se discutió en la sección 6.6.

a) Escriba la función de probabilidad.

b) Escriba un conjunto de ecuaciones que cuando se resuelvan den los estimadores de probabilidad máxima de α y β.

9.85 Considere un experimento hipotético donde un hombre con un hongo utiliza un medicamento fungicida y se cura. Considérelo, entonces, como una muestra de uno de una distribución de Bernoulli con función de probabilidad

$$f(x) = p^x q^{1-x}, \quad x = 0, 1,$$

donde p es la probabilidad de un éxito (curación) y $q = 1 - p$. Ahora, por supuesto, la información muestral da $x = 1$. Escriba un desarrollo que demuestre que $\hat{p} = 1.0$ es el estimador de probabilidad máxima de la probabilidad de curación.

9.86 Considere la observación X de la distribución binomial negativa dada en la sección 5.5. Encuentre el estimador de probabilidad máxima para p, con k desconocida.

Ejercicios de repaso

9.87 Considere dos estimadores de σ^2 en una muestra $x_1, x_2, \ldots, x_n$, que se extrae de una distribución normal con media μ y varianza σ^2. Los estimadores son el estimador insesgado $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, y el estimador de probabilidad máxima $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Discuta las propiedades de la varianza de estos dos estimadores.

9.88 Se afirma que una nueva dieta reducirá en 4.5 kilogramos el peso de un individuo, en promedio, en

un lapso de 2 semanas. Los pesos de 7 mujeres que siguieron esta dieta se registraron antes y después de un periodo de 2 semanas.

Mujer	Peso antes	Peso después
1	58.5	60.0
2	60.3	54.9
3	61.7	58.1
4	69.0	62.1
5	64.0	58.5
6	62.6	59.9
7	56.7	54.4

Pruebe la afirmación del fabricante calculando un intervalo de confianza de 95% para la diferencia media en el peso. Suponga que las diferencias de los pesos se distribuyen de forma aproximadamente normal.

9.89 De acuerdo con el *Roanoke Times* (16 de marzo de 1997), la cadena McDonald's vendió 42.1% del mercado de hamburguesas. Una muestra aleatoria de 75 hamburguesas vendidas tiene como resultado que 28 de ellas fueron vendidas por McDonald's. Utilice el material de la sección 9.9 para determinar si esta información apoya la afirmación del *Roanoke Times*.

9.90 Se llevó a cabo un estudio en el Instituto Politécnico y Universidad Estatal de Virginia para determinar si el fuego se puede utilizar como una herramienta de control viable, para aumentar la cantidad de forraje disponible para los venados, durante los meses críticos a finales del invierno y principios de la primavera. El calcio es un elemento que requieren las plantas y los animales. La cantidad que la planta toma y almacena está estrechamente correlacionada con la cantidad presente en el suelo. Se formuló la hipótesis de que el fuego puede cambiar los niveles de calcio presentes en el suelo y afectar así la cantidad disponible para los venados. Se seleccionó una extensión grande de tierra en el Fishburn Forest para efectuar un incendio controlado. Se tomaron muestras de suelo de 12 parcelas de igual área justo antes de la quema, y se analizaron para verificar el contenido de calcio. Los niveles de calcio después de la quema se analizaron en las mismas parcelas. Tales valores, en kilogramos por parcela, se presentan en la siguiente tabla:

	Nivel de calcio (kg/parcela)	
Parcela	Antes de la quema	Después de la quema
1	50	9
2	50	18
3	82	45
4	64	18
5	82	18
6	73	9
7	77	32
8	54	9
9	23	18
10	45	9
11	36	9
12	54	9

Construya un intervalo de confianza de 95% para la diferencia media en el nivel de calcio presente en el suelo antes y después del incendio controlado. Suponga que la distribución de las diferencias en los niveles de calcio es aproximadamente normal.

9.91 Un gimnasio con spa afirma que un nuevo programa de ejercicios reducirá la talla de la cintura de una persona en 2 centímetros, en promedio, durante un periodo de 5 días. Las tallas de cintura de 6 hombres que participaron en este programa de ejercicio se registraron, antes y después del periodo de 5 días, en la siguiente tabla:

Hombre	Talla de cintura antes	Talla de cintura después
1	90.4	91.7
2	95.5	93.9
3	98.7	97.4
4	115.9	112.8
5	104.0	101.3
6	85.6	84.0

Mediante el cálculo de un intervalo de confianza de 95% para la reducción media de la talla de cintura, determine si la afirmación del gimnasio con spa es válida. Suponga que la distribución de las diferencias de tallas de cintura antes y después del programa es aproximadamente normal.

9.92 El Departamento de Ingeniería Civil del Instituto Politécnico y Universidad Estatal de Virginia comparó una técnica de ensayo modificada (M-5 hr) para recuperar coliformes fecales en residuos líquidos (charcos) de agua de lluvia, en un área urbana con la técnica del número más probable (NMP). Se colectaron un total de 12 muestras de tales residuos y se analizan con las dos técnicas. Los conteos de coliformes fecales por 100 mililitros se registraron en la siguiente tabla:

Muestra	Conteo NMP	Conteo con M-5 hr
1	2300	2010
2	1200	930
3	450	400
4	210	436
5	270	4100
6	450	2090
7	154	219
8	179	169
9	192	194
10	230	174
11	340	274
12	194	183

Construya un intervalo de confianza de 90% para diferencia en los conteos medios de coliformes fecales entre las técnicas M-5 hr y NMP. Suponga que las diferencias de conteos se distribuyen de forma aproximadamente normal.

9.93 Se lleva a cabo un experimento para determinar si el acabado superficial tiene un efecto sobre el límite de fatiga del acero. Una teoría existente indica que el pulido aumenta el límite de fatiga medio (flexión inversa). Desde un punto de vista práctico, el pulido no debería tener efecto alguno en la desviación estándar del límite de fatiga, el cual se sabe que es de 4000 psi,

gracias a la realización de diversos experimentos de límite de fatiga. El experimento se realiza sobre acero al carbón al 0.4% usando especímenes sin y con pulido suave. Los datos son los siguientes:

Límite de fatiga (psi) para:	
Acero al carbón 0.4% pulido	**Acero al carbón 0.4% sin pulir**
85,500	82,600
91,900	82,400
89,400	81,700
84,000	79,500
89,900	79,400
78,700	69,800
87,500	79,900
83,100	83,400

Encuentre un intervalo de confianza de 95% para la diferencia entre las medias poblacionales para los dos métodos. Suponga que las poblaciones se distribuyen de forma aproximadamente normal.

9.94 Un antropólogo se interesa en la proporción de individuos de dos tribus indias con doble remolino de cabello en la zona occipital de la cabeza. Suponga que se toman muestras independientes de cada una de las dos tribus, y se encuentra que 24 de 100 individuos de la tribu A y 36 de 120 individuos de la tribu B poseen tal característica. Construya un intervalo de confianza de 95% para la diferencia $p_B - p_A$ entre las proporciones de estas dos tribus con remolinos de cabello en la zona occipital de la cabeza.

9.95 Un fabricante de planchas eléctricas produce estos artículos en dos plantas. Ambas plantas tienen al mismo proveedor de partes pequeñas. Se puede tener un ahorro al comprar termostatos para la planta B de un proveedor local. Se compra un solo lote del proveedor local y se desea probar si estos nuevos termostatos son tan precisos como los anteriores. Los termostatos se prueban en planchas a 550 °F, y las temperaturas reales se redondean al siguiente 0.1 °F con un termopar. Los datos son los siguientes:

Proveedor nuevo (°F)					
530.3	559.3	549.4	544.0	551.7	566.3
549.9	556.9	536.7	558.8	538.8	543.3
559.1	555.0	538.6	551.1	565.4	554.9
550.0	554.9	554.7	536.1	569.1	
Proveedor anterior (°F)					
559.7	534.7	554.8	545.0	544.6	538.0
550.7	563.1	551.1	553.8	538.8	564.6
554.5	553.0	538.4	548.3	552.9	535.1
555.0	544.8	558.4	548.7	560.3	

Encuentre un intervalo de confianza de 95% para σ_1^2/σ_2^2 y para σ_1/σ_2, donde σ_1^2 y σ_2^2 son las varianzas poblacionales de las lecturas de los termostatos del proveedor nuevo y del anterior, respectivamente.

9.96 Se afirma que la resistencia del alambre A es mayor que la del alambre B. Un experimento sobre los alambres muestra los siguientes resultados (en ohms):

Alambre *A*	**Alambre *B***
0.140	0.135
0.138	0.140
0.143	0.136
0.142	0.142
0.144	0.138
0.137	0.140

Suponiendo varianzas iguales, ¿qué conclusiones extrae? Justifique su respuesta.

9.97 Una forma alternativa de estimación se lleva a cabo a través del método de momentos. El método implica igualar la media y la varianza poblacionales a las correspondientes media muestral $\bar{x}$ y varianza muestral s^2, y resolver para el parámetro; el resultado son los **estimadores del momento**. En el caso de un solo parámetro, únicamente se utilizan las medias. Dé un argumento de que en el caso de la distribución de Poisson el estimador de probabilidad máxima y los estimadores del momento son iguales:

9.98 Especifique los estimadores del momento para μ y σ^2 para la distribución normal.

9.99 Especifique los estimadores del momento para μ y σ^2 para la distribución logarítmica normal.

9.100 Especifique los estimadores del momento para α y β para la distribución gamma.

9.101 Se realizó una encuesta con la finalidad de comparar los sueldos de administradores de plantas químicas empleados en dos áreas del país: las regiones norte y centro-occidente. Se eligieron muestras aleatorias independientes de 300 gerentes de planta para cada una de las dos regiones. A tales gerentes se les preguntó el monto de su sueldo anual. Los resultados fueron.

Norte	**Centro-Occidente**
$\bar{x}_1 = \$102{,}300$	$\bar{x}_2 = \$98{,}500$
$s_1 = \$5{,}700$	$s_2 = \$3{,}800$

a) Construya un intervalo de confianza de 99% en $\mu_1 - \mu_2$, la diferencia en los dos sueldos medios.
b) ¿Cuál es la suposición que usted hizo en el inciso *a*) acerca de la distribución de los sueldos anuales para las dos regiones? ¿Es necesaria la suposición de normalidad? ¿Por qué?
c) ¿Qué suposición hizo acerca de las dos varianzas? ¿Es razonable la suposición de igualdad de varianzas?

9.102 Considere el ejercicio de repaso 9.101. Supongamos que los datos no se han recabado aún. Supongamos también que los estadísticos previos sugieren que

$\sigma_1 = \sigma_2 = \$4000$. Los tamaños de las muestras en el ejercicio de repaso 9.101 son suficientes para producir un intervalo de confianza de 95% en $\mu_1 - \mu_2$ que tenga un ancho de sólo \$1000? Presente el desarrollo completo.

9.103 Un sindicato específico se preocupa por el notorio ausentismo de sus miembros. El sindicato decidió verificar esto monitoreando una muestra aleatoria de sus miembros. A los líderes del sindicato siempre se les ha reclamado que, en un mes típico, 95% de sus afiliados están ausentes al menos 10 horas mensuales. Utilice los datos para responder tal reclamación. Utilice un límite de tolerancia unilateral y elija el nivel de confianza de 99%. Asegúrese de aplicar lo que ya sabe acerca del cálculo del límite de tolerancia. El número de miembros en esta muestra fue 300. El número de horas de ausentismo se registró para cada uno de los 300 miembros. Los resultados fueron $\bar{x} = 6.5$ horas y $s = 2.5$ horas.

9.104 Se seleccionó una muestra aleatoria de 30 empresas que comercializan productos inalámbricos para determinar la proporción de tales firmas que implementaron software nuevo para mejorar la productividad. Resultó que 8 de 30 habían implementado tal software. Encuentre un intervalo de confianza de 95% en p, la proporción real de tales empresas que implementó el nuevo software.

9.105 Refiérase al ejercicio de repaso 9.104. Suponga que hay interés en si la estimación puntual $\hat{p} = 8/30$ el lo suficientemente precisa porque el intervalo de confianza alrededor de p no es suficientemente estrecho. Utilizando $\hat{p}$ como nuestro estimado de p, ¿cuántas compañías necesitarían muestrearse para tener un intervalo de confianza de 95% con un ancho de sólo 0.05?

9.106 Un fabricante produce un artículo que se clasifica como "defectuosos" o "no defectuoso". Para estimar la proporción de defectuosos, se toma una muestra aleatoria de 100 artículos de la producción y se encuentran 10 defectuosos. Después de la implementación del programa de mejoramiento de la calidad, se realizó nuevamente el experimento. Se tomó una nueva muestra de 100 y esta vez únicamente 6 salieron defectuosos.

a) Dado un intervalo de confianza de 95% en $p_1 - p_2$, donde p_1 es la proporción de defectuosos de la población antes de la mejoría, y p_2 es la proporción de defectuosos después de la mejoría.

b) ¿Hay información en el intervalo de confianza que se encontró en el inciso *a*) que sugiera que $p_1 > p_2$? Explique.

9.107 Se utiliza una máquina para llenar cajas de un producto en una operación de la línea de ensamble. Mucho del interés se centra en la variabilidad del número de onzas del producto en la caja. Se sabe que la desviación estándar en el peso del producto es de 0.3 onzas. Se realizan mejoras y luego se toma una muestra aleatoria de 20 cajas y se encuentra que la varianza de la muestra es 0.045 onzas. Encuentre un intervalo de confianza de 95% en la varianza del peso del producto. Considerando el rango del intervalo de confianza, ¿parecería que el mejoramiento en el proceso incrementó la calidad en cuanto a variabilidad se refiere. Suponga normalidad en la distribución del peso del producto.

9.108 Un grupo de consumidores está interesado en comparar los costos de operación para dos diferentes tipos de motor para automóvil. El grupo es capaz de encontrar 15 propietarios cuyos automóviles tienen motor tipo A y 15 que tienen motor tipo B. Los 30 propietarios compraron sus automóviles en aproximadamente el mismo tiempo y todos llevan buenos registros por cierto periodo de 12 meses. Además, se encontró que los propietarios recorrieron aproximadamente el mismo número de millas. Los estadísticos de costo son $\hat{y}_A = \$87.00/1{,}000$ millas, $\hat{y}_B = \$75.00/1{,}000$ millas, $s_A = \$5.99$, y $s_B = \$4.85$. Calcule un intervalo de confianza de 95% para estimar $\mu_A - \mu_B$, la diferencia en el costo medio de operación. Suponga normalidad e igual varianza.

9.109 Considere el estadístico S_p^2, el estimado de unión de σ^2. El estimador se examinó en la sección 9.8 y se utiliza cuando se está dispuesto a suponer que $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Demuestre que el estimador está insesgado para σ^2 (es decir, demuestre que $E(S_p^2) = \sigma^2$). Puede utilizar los resultados de cualquier teorema o ejemplo del capítulo 9.

9.110 Un grupo de investigadores del factor humano están interesados en la reacción de los pilotos de avión ante un estímulo con cierta disposición de la cabina del avión. Se realizó un experimento de simulación en un laboratorio y se utilizaron 15 pilotos con un tiempo de reacción promedio de 3.2 segundos y una desviación estándar muestral de 0.6 segundos. Resulta de interés caracterizar los extremos (es decir, el escenario del peor de los casos). Para tal objetivo, responda lo siguiente:

a) Determine un importante límite de confianza específico de 99% unilateral en el tiempo medio de reacción. ¿Qué suposición, si la hubiera, debería hacer acerca de la distribución del tiempo de reacción?

b) Determine un intervalo de predicción de 99% unilateral y dé una interpretación de lo que significa. ¿Debería usted hacer alguna suposición sobre la distribución del tiempo de reacción para calcular este límite?

c) Calcule un límite de tolerancia unilateral con 99% de confianza que implique 95% del tiempo de reacción. De nuevo, si las hubiera, dé una interpretación y una suposición de la distribución. [Nota: Los valores del límite de tolerancia unilateral también se incluyen en la tabla A.7.]

9.111 Cierto proveedor fabrica un tipo de estera de goma que vende a las compañías automotrices. En la

aplicación, las piezas del material deben tener ciertas características de dureza. Ocasionalmente, se detectan las esteras defectuosas y se rechazan. El proveedor afirma que la proporción de defectuosas es 0.05. El desafío llegó de un cliente que compró el producto. De manera que se realizó un experimento donde se probaron 400 esteras y se encontraron 17 defectuosas.

a) Calcule un intervalo de confianza bilateral de 95% en la proporción de defectuosos.

b) Calcule un intervalo de confianza unilateral de 95% adecuado en la proporción de defectuosos.

c) Interprete los intervalos de ambos incisos y comente acerca de la afirmación hecha por el proveedor.

9.15 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

El concepto de un *intervalo de confianza de muestra grande* en una población a menudo confunde a los estudiantes que se inician en esta materia, lo cual tiene como base la prescripción que incluso cuando se desconoce σ y no se está convencido de que la distribución que se muestrea sea normal, entonces un intervalo de confianza en μ puede calcularse de

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

En la práctica, esto se usa frecuentemente cuando la muestra es demasiado pequeña. El origen de este intervalo de *muestra grande* es, por supuesto, el teorema del límite central (TLC), con el cual no se requiere la normalidad. Aquí, el TLC requiere una σ, de la cual s sea sólo una estimación. Entonces, el requerimiento es que n sea, por lo menos, tan grande como 30 y la distribución subyacente esté cercana a la simetría, en cuyo caso el intervalo sigue siendo una aproximación.

Hay casos en que la aplicación práctica del material del capítulo 9 debe supervisarse en el contexto de ese capítulo. Un ejemplo importante es el uso de la distribución t, para el intervalo de confianza sobre μ cuando se desconoce σ. En sentido estricto, el uso de la distribución t requiere que la muestra de la distribución sea normal. Sin embargo, se sabe bien que cualquier aplicación de la distribución t es razonablemente insensible (es decir, **robusta**) a la suposición de normalidad. Esto representa una de esas situaciones afortunadas que ocurren en el campo de la estadística, donde no se mantiene el supuesto básico e incluso "¡todo resulta correcto!" Sin embargo, una población de la que se toma no puede desviarse sustancialmente de la normal. Entonces, a las gráficas de probabilidad normal estudiadas en el capítulo 8 y a las pruebas de bondad del ajuste que se presentan en el capítulo 10 se les requerirá con frecuencia que indaguen algún sentido de "cercanía a la normalidad". Esta idea de "robusteza de la normalidad" vuelve a presentarse en el capítulo 10.

Por experiencia, sabemos que uno de los más graves "usos incorrectos de la estadística" en la práctica surge de la confusión en la distinción entre la interpretación de los tipos de intervalos estadísticos. Por eso, en este capítulo ofrecemos una subsección donde se examinan las diferencias entre los tres tipos de intervalos. Es muy probable que, en la práctica, **el intervalo de confianza utilice en exceso**, es decir, se emplea cuando en realidad no hay interés en la media. Más bien habría algunas preguntas del tipo: "¿a dónde va a caer la siguiente observación?" O, a menudo, y más importante: "¿dónde está la mayoría de la distribución?" Éstas son preguntas fundamentales que no pueden responderse calculando un intervalo en la media. Un intervalo de confianza generalmente se emplea incorrectamente como un intervalo tal que la probabilidad de que el parámetro caiga en este intervalo es,

digamos, 95%, lo cual es una interpretación correcta del **intervalo posterior bayesiano**. (Para mayores referencias sobre la inferencia bayesiana véase el capítulo 18). El intervalo de confianza tan sólo sugiere que si el experimento o los datos se observan una y otra vez, aproximadamente el 95% de tales intervalos contendrá el parámetro real. Cualquier estudiante que se inicie en la estadística práctica debería tener muy claras las diferencias entre estos intervalos estadísticos.

Otra aplicación incorrecta potencial grave de la estadística es el uso de la distribución χ^2 para un intervalo de confianza de una sola varianza. De nuevo se supone normalidad en la distribución a partir de la cual se toma la muestra. A diferencia del uso de la distribución t, el uso de la prueba χ^2 para esta aplicación **no es robusta para la suposición de normalidad** (es decir, la distribución muestral $\frac{(n-1)S^2}{\sigma^2}$ se desvía bastante de χ^2 si la distribución subyacente no es normal). Entonces, el uso correcto de la prueba de bondad del ajuste (capítulo 10) y/o las gráficas de probabilidad normal puede ser muy importante aquí. Más información sobre esta cuestión general en particular se verá en los siguientes capítulos.

Capítulo 10

Pruebas de hipótesis de una y dos muestras

10.1 Hipótesis estadísticas: Conceptos generales

A menudo, el problema al que se enfrentan el científico o el ingeniero no es tanto la estimación de un parámetro poblacional, como vimos en el capítulo 9, sino más bien la formación de un procedimiento de decisión que se base en los datos, el cual ofrezca una conclusión acerca de algún sistema científico. Por ejemplo, un investigador médico puede decidir, sobre la base de evidencia experimental, si en los seres humanos beber café incrementa el riesgo de padecer cáncer; un ingeniero quizá tenga que decidir sobre la base de datos muestrales si hay una diferencia entre la precisión de dos tipos de medidores; o tal vez un sociólogo desee reunir los datos apropiados que le permitan decidir si el tipo sanguíneo de un individuo y el color de los ojos son variables independientes. En cada uno de estos casos, el científico o el ingeniero *postulan* o *conjeturan* algo acerca de un sistema. Además, cada uno debe incluir el uso de datos experimentales y la toma de decisiones basadas en ellos. De manera formal, en cada caso, la conjetura se puede poner en forma de hipótesis estadística. Los procedimientos que conducen a la aceptación o al rechazo de hipótesis estadísticas como éstas comprenden un área importante de la inferencia estadística: Primero, definamos con precisión lo que entendemos por **hipótesis estadística**.

Definición 10.1: Una **hipótesis estadísticas** es una aseveración o conjetura con respecto a una o más poblaciones.

La verdad o falsedad de una hipótesis estadística nunca se sabe con absoluta certidumbre, a menos que examinemos toda la población, lo cual, por supuesto, sería poco práctico en la mayoría de las situaciones. En cambio, tomamos una muestra aleatoria de la población de interés, y utilizamos los datos contenidos en esta muestra para proporcionar evidencia que apoye o no la hipótesis. La evidencia de la muestra que sea inconsistente con la hipótesis que se establece conduce al rechazo de ésta.

El papel de la probabilidad en la prueba de hipótesis

Debería quedar claro al lector que un procedimiento de decisión debe hacerse con la noción de la *probabilidad de una conclusión errónea*. Por ejemplo, suponga que la hipótesis que postuló el ingeniero es que la fracción p de defectuosos en cierto proceso es 0.10. El experimento es la observación de una muestra aleatoria del producto en cuestión. Suponga que se prueban 100 artículos y se encuentran 12 defectuosos. Es razonable concluir que esta evidencia no rechaza la condición $p = 0.10$, y por ello puede conducir a la aceptación de la hipótesis. Sin embargo, tampoco rechaza $p = 0.12$ o quizá incluso $p = 0.15$. Como resultado, el lector se debe acostumbrar a comprender que **el rechazo de una hipótesis simplemente implica que la evidencia de la muestra la refuta**. Por otro lado, **el rechazo significa que hay una pequeña probabilidad de obtener la información muestral observada cuando, de hecho, la hipótesis es verdadera**. Por ejemplo, en nuestra hipótesis de la proporción de defectuosos, una muestra de 100 que revela 20 artículos defectuosos es ciertamente evidencia de rechazo. ¿Por qué? Si, en realidad, $p = 0.10$, la probabilidad de obtener 20 o más defectuosos es aproximadamente 0.002. Con el pequeño riesgo resultante de una conclusión errónea, parecería seguro **rechazar la hipótesis** de que $p = 0.10$. En otras palabras, el rechazo de una hipótesis tiende a casi "descartar" la hipótesis. Por otro lado, es muy importante enfatizar que la aceptación o, más bien, la falla al rechazo no excluye otras posibilidades. Como resultado, *el analista de los datos establece una conclusión firme cuando se rechaza una hipótesis.*

El planteamiento formal de una hipótesis a menudo está influido por la estructura de la probabilidad de una conclusión errónea. Si el científico se interesa en apoyar con fuerza una opinión, desea llegar a la opinión en la forma del rechazo de una hipótesis. Si el investigador médico desea mostrar evidencia sólida a favor de la opinión de que beber café aumenta el riesgo de contraer cáncer, la hipótesis a probar debería tener la forma "no hay aumento en el riesgo de padecer cáncer como consecuencia de beber café". Como resultado, la opinión se alcanza mediante un rechazo. De manera similar, para apoyar la afirmación de que un tipo de medidores es más preciso que otro, el ingeniero prueba la hipótesis de que no hay diferencia en la precisión de los dos tipos de medidor.

Lo anterior implica que cuando el análisis de datos formaliza la evidencia experimental con base en la prueba de hipótesis, es muy importante la **declaración** o el **establecimiento formal** de la hipótesis.

Hipótesis nula e hipótesis alternativa

La estructura de la prueba de hipótesis se formulará usando el término **hipótesis nula**, el cual se refiere a cualquier hipótesis que deseamos probar y se denota con H_0. El rechazo de H_0 conduce a la aceptación de una **hipótesis alternativa**, que se denota con H_1. La comprensión de las diferentes funciones que desempeñan la hipótesis nula (H_0) y la hipótesis alternativa (H_1) es fundamental para entender los principios de la prueba de hipótesis. La hipótesis alternativa H_1, por lo general, representa la *pregunta que debe responderse*, *la teoría que debe probarse* y, por ello, su especificación es muy importante. La hipótesis nula H_0 anula o se opone a H_1 y a menudo es el complemento lógico para H_1. Conforme el lector vaya aprendiendo más sobre la prueba de hipótesis, debería notar que el analista llega a una de las siguientes dos conclusiones:

rechace H_0: a favor de H_1 debido a evidencia suficiente en los datos.

no rechace H_0: debido a evidencia insuficiente en los datos.

Observe que las conclusiones no implican una "aceptación" formal y literal de H_0. El enunciado de H_0 a menudo representa el "status quo" contrario a una nueva idea, conjetura, etcétera, enunciada en H_1; en tanto que no rechazar H_0 representa la conclusión adecuada. En nuestro ejemplo binomial, la cuestión práctica podría surgir a partir de un interés en que la probabilidad histórica de defectuosos de 0.10 ya no es real. De hecho, la conjetura podría ser que excede 0.10. De manera que enunciaríamos

$$H_0\text{: } p = 0.10,$$
$$H_1\text{: } p > 0.10.$$

Ahora, 12 artículos defectuosos de cada 100 no rechazan $p = 0.10$, por lo que la conclusión es "no rechace H_0". Sin embargo, si los datos producen 20 artículos defectuosos de cada 100, la conclusión sería "rechace H_0" a favor de H_1: $p > 0.10$.

Aunque las aplicaciones de la prueba de hipótesis son bastante abundantes en trabajos científicos y de ingeniería, quizás el mejor ejemplo para un principiante sea la dificultad que se encuentra en el veredicto de un jurado. Las hipótesis nula y alternativa son

H_0: el acusado es inocente,

H_1: el acusado es culpable.

La acusación proviene de una sospecha de culpabilidad. La hipótesis H_0 (status quo) se establece en oposición a H_1 y se mantiene a menos que se apoye H_1 con evidencia "más allá de una duda razonable". Sin embargo, en este caso "no rechace H_0" no implica inocencia, sino tan sólo que la evidencia fue insuficiente para lograr una condena. De manera que el jurado no necesariamente *acepta* H_0 sino que *no rechaza* H_0.

10.2 Prueba de una hipótesis estadística

Para ilustrar los conceptos que se utilizan al probar una hipótesis estadística acerca de una población, considere el siguiente ejemplo. Se sabe que cierto tipo de vacuna contra el resfriado tan sólo es efectiva en 25% después de un periodo de dos años. Para determinar si una vacuna nueva, y algo más cara, es superior al dar protección contra el mismo virus durante un periodo más largo, suponga que se elige a 20 personas al azar y se inoculan. En un estudio real de este tipo, los participantes que reciben la nueva vacuna pueden llegar a varios miles. El número 20 se utiliza aquí sólo para demostrar los pasos básicos para realizar una prueba estadística. Si más de 8 de quienes reciben la nueva vacuna superan el lapso de 2 años sin contraer el virus, la nueva vacuna se considerará superior a la que se usa en la actualidad. El requisito de que el número exceda de 8 es algo arbitrario, aunque parece razonable, ya que representa una ganancia modesta sobre las 5 personas que se esperaría que recibieran protección si las 20 personas se inocularon con la vacuna ya en uso. En esencia probamos la hipótesis nula de que —después de un periodo de 2 años— la nueva vacuna es igualmente eficaz que la que, por lo general, se utiliza ahora. La hipótesis alternativa es que la nueva vacuna es de hecho superior, lo cual es equivalente a

probar la hipótesis de que el parámetro binomial para la probabilidad de un éxito sobre una prueba dada es $p = 1/4$ contra la alternativa de que $p = 1/4$. Esto, por lo general, se escribe como:

$$H_0\text{: } p = 0.25,$$
$$H_1\text{: } p > 0.25.$$

El estadístico de prueba

El **estadístico de prueba** sobre el cual se basa nuestra decisión es X, el número de individuos en nuestro grupo de prueba que reciben protección de la nueva vacuna durante un periodo de al menos 2 años. Los valores posibles de X, de 0 a 20, se dividen en dos grupos: los números menores que o iguales a 8 y aquellos mayores que 8. Todos los posibles valores mayores que 8 constituyen la **región crítica**. El último número que observamos al pasar a la región crítica se llama **valor crítico**. En nuestro caso el valor crítico es el número 8. Por lo tanto, si $x > 8$, rechazamos H_0 a favor de la hipótesis alternativa H_1. Si $x \leq 8$, no rechazamos H_0. Este criterio de aceptación se ilustra en la figura 10.1.

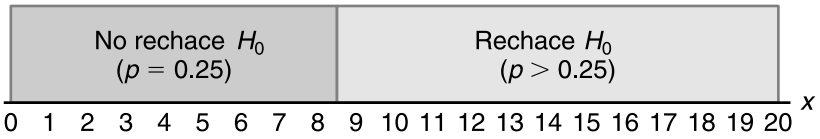


Figura 10.1: Criterio de decisión para probar $p = 0.25$ contra $p > 0.25$.

El procedimiento de decisión recién descrito podría conducir a cualquiera de dos conclusiones erróneas. Por ejemplo, la nueva vacuna puede no ser mejor que la que se usa actualmente y, para este grupo específico de individuos seleccionado de forma aleatoria, más de 8 pasan el periodo de 2 años sin contraer el virus. Cometeríamos un error al rechazar H_0 a favor de H_1 cuando, de hecho, H_0 es verdadera. Tal error se llama **error tipo I**.

Definición 10.2: El rechazo de la hipótesis nula cuando es verdadera se llama **error tipo I**.

Una segunda clase de error se comete si 8 o menos del grupo superan exitosamente el periodo de 2 años y concluimos que la nueva vacuna no es mejor cuando en realidad sí lo es. En este caso aceptamos H_0 cuando de hecho es falsa. Éste se llama **error tipo II**.

Definición 10.3: No rechazar la hipótesis nula cuando es falsa se llama **error tipo II**.

Al probar cualquier hipótesis estadística, hay cuatro situaciones posibles que determinan si nuestra decisión es correcta o errónea. Estas cuatro situaciones se resumen en la tabla 10.1.

La probabilidad de cometer un error tipo I, también llamada **nivel de significancia**, se denota con la letra griega α. En nuestro caso, un error tipo I ocurrirá

Tabla 10.1: Situaciones posibles al probar una hipótesis estadística

	H_0 es verdadera	H_0 es falsa
No rechace H_0	Decisión correcta	Error tipo II
Rechace H_0	Error tipo I	Decisión correcta

cuando más de ocho individuos superen el periodo de 2 años sin contraer el virus, al usar la nueva vacuna que en realidad equivale a la que está en uso. Por lo tanto, si X es el número de individuos que permanecen libres del virus por al menos dos años,

$$\alpha = P(\text{error tipo I}) = P\left(X > 8 \text{ cuando } p = \frac{1}{4}\right) = \sum_{x=9}^{20} b\left(x; 20, \frac{1}{4}\right)$$
$$= 1 - \sum_{x=0}^{8} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.9591 = 0.0409.$$

Decimos que la hipótesis nula, $p = 1/4$, se prueba al nivel de significancia $\alpha = 0.0409$. Algunas veces el nivel de significancia se llama **tamaño de la prueba**. Una región crítica de tamaño 0.0409 es muy pequeña y, por lo tanto, es poco probable que se cometa un error de tipo I. En consecuencia, sería poco probable que más de 8 individuos permanecieran inmunes a un virus por un periodo de dos años mediante el uso de una vacuna nueva, que en esencia es equivalente a la que ahora existe en el mercado.

La probabilidad de un error tipo II

La probabilidad de cometer un error tipo II, que se denota con β, es imposible de calcular a menos que tengamos una hipótesis alternativa específica. Si probamos la hipótesis nula $p = 1/4$ contra la hipótesis alternativa $p = 1/2$, entonces seremos capaces de calcular la probabilidad de no rechazar H_0 cuando es falsa. Simplemente encontramos la probabilidad de obtener 8 o menos en el grupo que supera el periodo de 2 años cuando $p = 1/2$. En este caso,

$$\beta = P(\text{error tipo II}) = P\left(X \leq 8 \text{ cuando } p = \frac{1}{2}\right)$$
$$= \sum_{x=0}^{8} b\left(x; 20, \frac{1}{2}\right) = 0.2517.$$

Ésta es una probabilidad más bien alta, que indica un procedimiento de prueba donde es muy probable que rechacemos la nueva vacuna cuando, de hecho, es superior a la que está en uso. Idealmente, preferiríamos utilizar un procedimiento de prueba en el cual sean pequeñas las probabilidades de los errores tipo I y tipo II.

Es posible que el director del programa de prueba esté dispuesto a cometer un error tipo II, si la vacuna más cara no es significativamente superior. De hecho, la única ocasión en la que desea estar prevenido contra un error tipo II es cuando el

valor real de p es al menos 0.7. Si $p = 0.7$, este procedimiento de prueba da

$$\begin{aligned}\beta = P(\text{error tipo II}) &= P(X \leq 8 \text{ cuando } p = 0.7)\\ &= \sum_{x=0}^{8} b(x; 20, 0.7) = 0.0051.\end{aligned}$$

Con una probabilidad tan pequeña de cometer un error tipo II, es bastante improbable que se rechace la nueva vacuna cuando tiene una efectividad de 70% después de un periodo de 2 años. Conforme la hipótesis alternativa se aproxima a la unidad, el valor de β tiende a cero.

El papel de α, β y el tamaño de la muestra

Supongamos que el director del programa de prueba no está dispuesto a cometer un error tipo II cuando la hipótesis alternativa $p = 1/2$ es verdadera, aun cuando se encuentre que la probabilidad de tal error es $\beta = 0.2517$. Una reducción de β siempre es posible al aumentar el tamaño de la región crítica. Por ejemplo, considere qué sucede a los valores de α y β cuando cambiamos nuestro valor crítico a 7, de manera que todos los valores mayores que 7 caigan en la región crítica, y aquellos menores que o iguales a 7 caigan en la región de no rechazo. Así, al probar $p = 1/4$ contra la hipótesis alternativa $p = 1/2$, encontramos que

$$\alpha = \sum_{x=8}^{20} b\left(x; 20, \frac{1}{4}\right) = 1 - \sum_{x=0}^{7} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.8982 = 0.1018,$$

y

$$\beta = \sum_{x=0}^{7} b\left(x; 20, \frac{1}{2}\right) = 0.1316.$$

Al adoptar un nuevo procedimiento de decisión, reducimos la probabilidad de cometer un error tipo II a costa de aumentar la probabilidad de cometer un error tipo I. Para un tamaño muestral fijo, una disminución en la probabilidad de un error, por lo general, tendrá como resultado un incremento en la probabilidad del otro error. Por fortuna, **la probabilidad de cometer ambos tipos de errores se puede reducir al aumentar el tamaño de la muestra**. Considere el mismo problema usando una muestra aleatoria de 100 individuos. Si más de 36 del grupo superan el periodo de 2 años, rechazamos la hipótesis nula $p = 1/4$ y aceptamos la hipótesis alternativa $p > 1/4$. El valor crítico ahora es 36. Todos los valores posibles por arriba de 36 constituyen la región crítica y todos los valores posibles menores que o iguales a 36 caen en la región de aceptación.

Para determinar la probabilidad de cometer un error tipo I, utilizaremos la aproximación de la curva normal con

$$\mu = np = (100)\left(\frac{1}{4}\right) = 25, \quad \text{y} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/4)(3/4)} = 4.33.$$

Con referencia a la figura 10.2, necesitamos el área bajo la curva normal a la derecha de $x = 36.5$. El valor z correspondiente es

$$z = \frac{36.5 - 25}{4.33} = 2.66.$$

De la tabla A.3 encontramos que

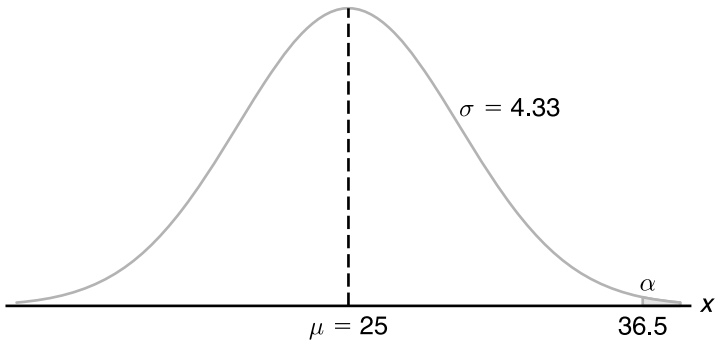


Figura 10.2: Probabilidad de un error tipo I.

$$\alpha = P(\text{Error tipo I}) = P\left(X > 36 \text{ cuando } p = \frac{1}{4}\right) \approx P(Z > 2.66)$$
$$= 1 - P(Z < 2.66) = 1 - 0.9961 = 0.0039.$$

Si H_0 es falsa y el verdadero valor de H_1 es $p = 1/2$, determinamos la probabilidad de un error tipo II usando la aproximación a la curva normal con

$$\mu = np = (100)(1/2) = 50 \quad \text{y} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/2)(1/2)} = 5.$$

La probabilidad de caer en la región de aceptación cuando H_0 es verdadera está dada por el área de la región sombreada a la izquierda de $x = 36.5$ en la figura 10.3. El valor z que corresponde a $x = 36.5$ es

$$z = \frac{36.5 - 50}{5} = -2.7.$$

Por lo tanto,

$$\beta = P(\text{Error tipo II}) = P\left(X \leq 36 \text{ cuando } p = \frac{1}{2}\right) \approx P(Z < -2.7) = 0.0035.$$

Evidentemente, los errores tipo I y tipo II rara vez ocurren si el experimento consiste en 100 individuos.

La ilustración anterior destaca la estrategia del científico en la prueba de hipótesis. Después de que se establecen las hipótesis nula y alternativa, es importante considerar la sensibilidad del procedimiento de prueba. Con esto queremos decir que debería haber una determinación, para una α fija, de un valor razonable para la probabilidad de aceptar de manera errónea H_0 (es decir, el valor de β) cuando la situación real representa alguna *desviación importante de* H_0. Por lo general, se puede determinar el valor del tamaño de la muestra para el que hay un equilibrio razonable entre α y el valor de β que se calcula de esta manera. El problema de la vacuna es un ejemplo.

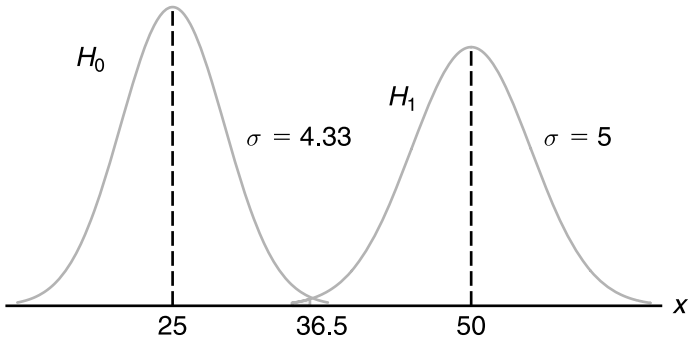


Figura 10.3: Probabilidad de un error tipo II.

Ilustración con una variable aleatoria continua

Los conceptos que aquí se discuten para una población discreta se pueden aplicar igualmente bien a variables aleatorias continuas. Considere la hipótesis nula de que el peso promedio de estudiantes hombres en cierta universidad es 68 kilogramos, contra la hipótesis alternativa de que es diferente de 68. Es decir, deseamos probar

$$H_0\text{: } \mu = 68,$$
$$H_1\text{: } \mu \neq 68.$$

La hipótesis alternativa nos permite la posibilidad de que $\mu < 68$ o $\mu > 68$.

Una media muestral que caiga cerca del valor hipotético de 68 se consideraría como evidencia en favor de H_0. Por otro lado, una media muestral considerablemente menor que o mayor que 68 sería una evidencia de inconsistencia de H_0 y, por lo tanto, favorecería a H_1. La media muestral es el estadístico de prueba en este caso. Una región crítica para el estadístico de prueba se puede elegir de manera arbitraria como los dos intervalos $\bar{x} < 67$ y $\bar{x} > 69$. La región de aceptación será entonces el intervalo $67 \leq \bar{x} \leq 69$. Este criterio de decisión se ilustra en la figura 10.4. Utilicemos

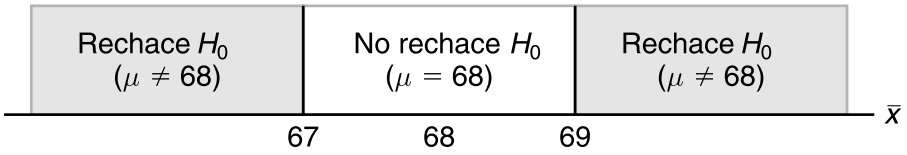


Figura 10.4: Región crítica (sombreada).

ahora el criterio de decisión de la figura 10.4 para calcular las probabilidades de cometer errores tipo I y tipo II, cuando se prueba la hipótesis nula $\mu = 68$ kilogramos contra la alternativa $\mu \neq 68$ kilogramos.

Suponga que la desviación estándar de la población de pesos es $\sigma = 3.6$. Para muestras grandes podemos sustituir s por σ si no se dispone de ninguna otra estimación de σ. Nuestro estadístico de decisión, que se basa en una muestra aleatoria de tamaño $n = 36$, será $\bar{X}$, el estimador más eficaz de μ. Del teorema del límite

central, sabemos que la distribución muestral de $\bar{X}$ es aproximadamente normal con desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 3.6/6 = 0.6$.

La probabilidad de cometer un error tipo I, o el nivel de significancia de nuestra prueba, es igual a la suma de las áreas sombreadas en cada cola de la distribución en la figura 10.5. Por lo tanto,

$$\alpha = P(\bar{X} < 67 \text{ cuando } \mu = 68) + P(\bar{X} > 69 \text{ cuando } \mu = 68).$$

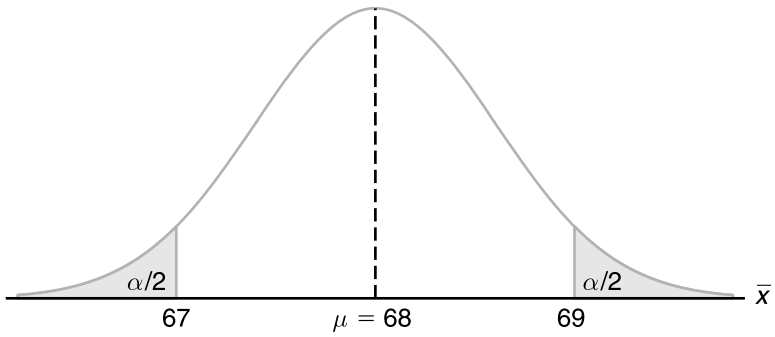


Figura 10.5: Región crítica para probar $\mu = 68$ contra $\mu \neq 68$.

Los valores z correspondientes a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando H_0 es verdadera son

$$z_1 = \frac{67 - 68}{0.6} = -1.67 \quad \text{y} \quad z_2 = \frac{69 - 68}{0.6} = 1.67.$$

Por lo tanto,

$$\alpha = P(Z < -1.67) + P(Z > 1.67) = 2P(Z < -1.67) = 0.0950.$$

De esta manera, 9.5% de todas las muestras de tamaño 36 nos conducirían a rechazar $\mu = 68$ kilogramos cuando, de hecho, ésta es verdadera. Para reducir α, tenemos que elegir entre aumentar el tamaño de la muestra o ampliar la región de aceptación. Suponga que aumentamos el tamaño de la muestra a $n = 64$. Entonces $\sigma_{\bar{X}} = 3.6/8 = 0.45$. Entonces,

$$z_1 = \frac{67 - 68}{0.45} = -2.22 \quad \text{y} \quad z_2 = \frac{69 - 68}{0.45} = 2.22.$$

De aquí,

$$\alpha = P(Z < -2.22) + P(Z > 2.22) = 2P(Z < -2.22) = 0.0264.$$

La reducción en α no es suficiente por sí misma para garantizar un buen procedimiento de prueba. Debemos evaluar β para varias hipótesis alternativas. Si es importante rechazar H_0 cuando la media real sea algún valor $\mu \geq 70$ o $\mu \leq 66$, entonces, la probabilidad de cometer un error tipo II se debería calcular y examinar para las alternativas $\mu = 66$ y $\mu = 70$. Debido a la simetría, sólo es necesario

considerar la probabilidad de no rechazar la hipótesis nula $\mu = 68$ cuando la alternativa $\mu = 70$ es verdadera. Resultará un error tipo II cuando la media muestral $\bar{x}$ caiga entre 67 y 69 cuando H_1 sea verdadera. Por lo tanto, con referencia a la figura 10.6, encontramos que

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ cuando } \mu = 70).$$

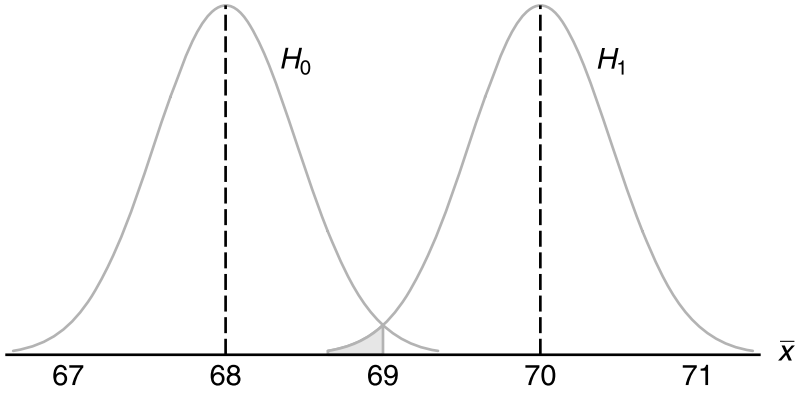


Figura 10.6: Probabilidad del error tipo II para probar $\mu = 68$ contra $\mu = 70$.

Los valores z que corresponden a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando H_1 es verdadera son

$$z_1 = \frac{67 - 70}{0.45} = -6.67 \quad \text{y} \quad z_2 = \frac{69 - 70}{0.45} = -2.22.$$

Por lo tanto,

$$\begin{aligned}\beta &= P(-6.67 < Z < -2.22) = P(Z < -2.22) - P(Z < -6.67) \\ &= 0.0132 - 0.0000 = 0.0132.\end{aligned}$$

Si el valor real de μ es la alternativa $\mu = 66$, el valor de β nuevamente será 0.0132. Para todos los valores posibles de $\mu < 66$ o $\mu > 70$, el valor de β será incluso más pequeño cuando $n = 64$ y, en consecuencia, habría poca oportunidad de aceptar H_0 cuando sea falsa.

La probabilidad de cometer un error tipo II aumenta rápidamente cuando el valor real de μ se aproxima al valor hipotético, pero no es igual a éste. Desde luego, por lo general, ésta es la situación en que no nos importa cometer un error tipo II. Por ejemplo, si la hipótesis alternativa $\mu = 68.5$ es verdadera, podemos cometer un error tipo II al concluir que la respuesta verdadera es $\mu = 68$. La probabilidad de cometer tal error será alta cuando $n = 64$. Con referencia a la figura 10.7, tenemos

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ cuando } \mu = 68.5).$$

Los valores z correspondientes a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando $\mu = 68.5$ son

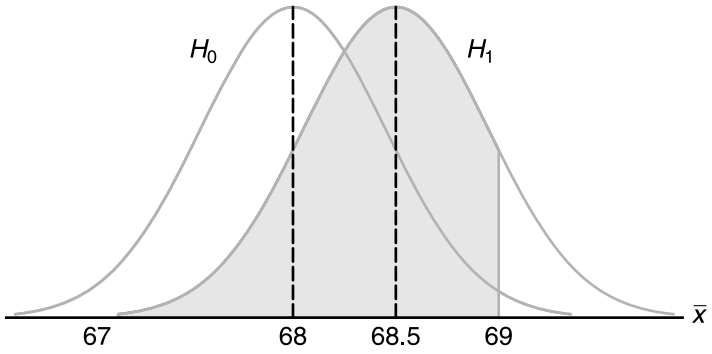


Figura 10.7: Error tipo II para probar $\mu = 68$ contra $\mu = 68.5$.

$$z_1 = \frac{67 - 68.5}{0.45} = -3.33 \quad \text{y} \quad z_2 = \frac{69 - 68.5}{0.45} = 1.11.$$

Por lo tanto,

$$\begin{aligned}\beta = P(-3.33 < Z < 1.11) &= P(Z < 1.11) - P(Z < -3.33) \\ &= 0.8665 - 0.0004 = 0.8661.\end{aligned}$$

Los ejemplos anteriores ilustran las siguientes propiedades importantes:

Propiedades importantes de una prueba de hipótesis

1. Los errores tipo I y tipo II están relacionados. Por lo general, una disminución en la probabilidad de uno tiene como resultado un incremento en la probabilidad del otro.
2. El tamaño de la región crítica y, por lo tanto, la probabilidad de cometer un error tipo I, siempre se puede reducir al ajustar el(los) valor(es) crítico(s).
3. Un aumento en el tamaño muestral n reducirá a α y β de forma simultánea.
4. Si la hipótesis nula es falsa, β es un máximo cuando el valor real de un parámetro se aproxima al valor hipotético. Cuanto más grande sea la distancia entre el valor real y el valor hipotético, β será menor.

Un concepto muy importante que se relaciona con las probabilidades del error es la noción de potencia de una prueba.

Definición 10.4: La **potencia** de una prueba es la probabilidad de rechazar H_0 dado que una alternativa específica es verdadera.

La potencia de una prueba se puede calcular como $1 - \beta$. A menudo **diferentes tipos de pruebas se comparan al contrastar propiedades de potencia**. Considere la ilustración anterior en la que probamos H_0: $\mu = 68$ y H_1: $\mu \neq 68$. Como antes, suponga que nos interesamos en evaluar la sensibilidad de la prueba. La prueba está determinada por la regla de que aceptamos H_0 si $67 \leq \bar{x} \leq 69$. Buscamos la capacidad de la prueba para rechazar H_0 de manera adecuada cuando en realidad $\mu = 68.5$. Vimos que la probabilidad de un error tipo II está dada por $\beta = 0.8661$. De esta manera, la **potencia** de la prueba es $1 - 0.8661 = 0.1339$. En cierto sentido, la potencia es una medida más sucinta de cuán sensible es la prueba para "detectar

diferencias" entre una media de 68 y 68.5. En este caso, si μ es realmente 68.5, la prueba como se describe *rechazará de forma adecuada* H_0 *sólo 13.39% de las veces.* Como resultado, la prueba no sería buena si es importante que el analista tenga una oportunidad razonable de distinguir realmente entre una media de 68.0 (que especifica H_0) y una media de 68.5. De lo anterior, resulta claro que para producir una potencia deseable (digamos, mayor que 0.8), se debe aumentar α o aumentar el tamaño de la muestra.

En las secciones anteriores de este capítulo, la mayoría del texto sobre prueba de hipótesis gira alrededor de los principios y las definiciones. En las secciones que siguen seremos más específicos y clasificaremos las hipótesis en categorías. También estudiaremos pruebas de hipótesis sobre varios parámetros de interés. Comenzamos estableciendo la distinción entre hipótesis unilaterales y bilaterales.

10.3 Pruebas de una y dos colas

Una prueba de cualquier hipótesis estadística, donde la alternativa es **unilateral**, como

$$H_0\colon \theta = \theta_0,$$
$$H_1\colon \theta > \theta_0,$$

o quizás

$$H_0\colon \theta = \theta_0,$$
$$H_1\colon \theta < \theta_0,$$

se denomina **prueba de una sola cola**. En la sección 10.2 se hizo referencia al **estadístico de prueba** para una hipótesis. Por lo general, la región crítica para la hipótesis alternativa $\theta > \theta_0$ yace en la cola derecha de la distribución del estadístico de prueba; en tanto que la región crítica para la hipótesis alternativa $\theta < \theta_0$ yace por completo en la cola izquierda. En cierto sentido, el símbolo de desigualdad apunta en la dirección donde se encuentra la región crítica. En el experimento de la vacuna de la sección 10.2 se utiliza una prueba de una sola cola para probar la hipótesis $p = 1/4$, contra la alternativa unilateral $p > 1/4$ para la distribución binomial. La región crítica de una sola cola, por lo general, es bastante evidente. Para una mejor comprensión, el lector debería visualizar el comportamiento del estadístico de prueba y observar la notoria *señal* que produciría evidencia que apoye la hipótesis alternativa.

Una prueba de cualquier hipótesis alternativa donde la alternativa sea **bilateral**, como

$$H_0\colon \theta = \theta_0,$$
$$H_1\colon \theta \neq \theta_0,$$

se llama **prueba de dos colas**, ya que la región crítica se divide en dos partes, que a menudo tienen probabilidades iguales que se colocan en cada cola de la distribución del estadístico de prueba. La hipótesis alternativa $\theta \neq \theta_0$ establece que ya sea que $\theta < \theta_0$ o que $\theta > \theta_0$. Una prueba de dos colas se utilizó para probar la hipótesis nula $\mu = 68$ kilogramos, contra la alternativa bilateral $\mu \neq 68$ kilogramos, para la población continua de los pesos de estudiantes en la sección 10.2.

¿Cómo se eligen las hipótesis nula y alternativa?

La hipótesis nula, H_0, con frecuencia se establecerá usando el *signo de igualdad*. De esta manera se observa cómo se controla la probabilidad de cometer un error tipo I. No obstante, hay situaciones en que la aplicación sugiere que "no rechazar H_0" implica que el parámetro θ podría ser cualquier valor definido por el complemento natural de la hipótesis alternativa. Por ejemplo, en el caso de la vacuna, donde la hipótesis alternativa es H_1: $p > 1/4$, es bastante posible que un no rechazo de H_0 no pueda descartar un valor de p menor que $1/4$. Claramente, en el caso de las pruebas de una cola, la declaración de la alternativa es la consideración más importante.

Si se establece una prueba de una cola o de dos colas, dependerá de la conclusión que se obtenga si se rechaza H_0. La posición de la región crítica puede determinarse sólo después de que se establece H_1. Por ejemplo, al probar una medicina nueva, se establece la hipótesis de que no es mejor que las medicinas similares que actualmente hay en el mercado, y se prueba ésta contra la hipótesis alternativa de que la medicina nueva es superior. Tal hipótesis alternativa tendrá como resultado una prueba de una sola cola con la región crítica en la cola derecha. No obstante, si deseamos comparar una nueva técnica de enseñanza con el procedimiento convencional del salón de clases, la hipótesis alternativa debe permitir que la nueva aproximación sea inferior o superior al procedimiento convencional. Por lo tanto, la prueba será de dos colas con la región crítica dividida en partes iguales, de manera que caiga en los extremos de las colas izquierda y derecha de la distribución de nuestro estadístico.

Ejemplo 10.1: Un fabricante de cierta marca de cereal de arroz afirma que el contenido promedio de grasa saturada no excede de 1.5 gramos. Establezca las hipótesis nula y alternativa a utilizar para probar esta afirmación y determinar dónde se localiza la región crítica.

Solución: La afirmación del fabricante se debería rechazar sólo si μ es mayor que 1.5 miligramos y no se debería rechazar si μ es menor o igual que 1.5 miligramos. Entonces, probamos

$$H_0\text{: } \mu = 1.5,$$
$$H_1\text{: } \mu > 1.5,$$

de manera que el no rechazo de H_0 no descarta valores que 1.5 miligramos. Como tenemos una prueba de una cola, el símbolo mayor que indica que la región crítica yace por completo en la cola derecha de la distribución de nuestro estadístico de prueba $\bar{X}$.

Ejemplo 10.2: Un agente de bienes raíces afirma que 60% de todas las viviendas privadas que se construyen actualmente son casas con tres dormitorios. Para probar esta afirmación, se inspecciona una muestra grande de viviendas nuevas. La proporción de tales casas con tres dormitorios se registra y se utiliza como estadístico de prueba. Establezca las hipótesis nula y alternativa a utilizarse en esta prueba y determine la posición de la región crítica.

Solución: Si el estadístico de prueba fuera considerablemente mayor o menor que $p = 0.6$, rechazaríamos la afirmación del agente, por lo que deberíamos establecer la hipótesis

$$H_0\text{: } p = 0.6,$$
$$H_1\text{: } p \neq 0.6.$$

La hipótesis alternativa implica una prueba de dos colas con la región crítica dividida por igual en ambas colas de la distribución de $\hat{P}$, nuestro estadístico de prueba.

10.4 Uso de valores *P* para la toma de decisiones en la prueba de hipótesis

Al probar hipótesis en que el estadístico de prueba es discreto, la región crítica se puede elegir de manera arbitraria y determinar su tamaño. Si α es demasiado grande, se puede reducir al realizar un ajuste en el valor crítico. Quizá sea necesario aumentar el tamaño de la muestra para compensar la disminución que ocurre de manera automática en la potencia de la prueba.

Por generaciones enteras de análisis estadístico, se ha vuelto costumbre elegir una α de 0.05 o 0.01 y, en consecuencia, seleccionar la región crítica. Entonces, desde luego, el rechazo o el no rechazo estrictos de H_0 dependerá de esa región crítica. Por ejemplo, si la prueba es de dos colas y α se fija al nivel de significancia de 0.05 y el estadístico de prueba implica, digamos, la distribución normal estándar, entonces se observa un valor z de los datos y la región crítica es

$$z > 1.96 \qquad \text{o} \qquad z < -1.96,$$

donde el valor 1.96 se encuentra como $z_{0.025}$ en la tabla A.3. Un valor de z en la región crítica sugiere el planteamiento: "El valor del estadístico de prueba es significativo." Podemos traducir esto al lenguaje del usuario. Por ejemplo, si la hipótesis está dada por

$$H_0\text{: } \mu = 10,$$
$$H_1\text{: } \mu \neq 10,$$

se puede decir: "La media difiere de manera significativa del valor 10."

Preselección del nivel de significancia

Esta preselección de un nivel de significancia α tiene sus raíces en la filosofía de que se debería controlar el riesgo máximo de cometer un error tipo I. Sin embargo, este enfoque no explica los valores del estadístico de prueba que están "cercanos" a la región crítica. Suponga que, por ejemplo, en la ilustración con H_0: $\mu = 10$, contra H_1: $\mu \neq 10$, se observa un valor $z = 1.87$. Estrictamente hablando, con $\alpha = 0.05$ el valor no es significativo; pero el riesgo de cometer un error tipo I si se rechaza H_0 en este caso difícilmente se podría considerar severo. De hecho, en un escenario de dos colas el riesgo se cuantifica como

$$P = 2P(Z > 1.87 \text{ cuando } \mu = 10) = 2(0.0307) = 0.0614.$$

Como resultado, 0.0614 es la probabilidad de obtener un valor z tan grande o mayor (en magnitud) que 1.87 cuando, de hecho, $\mu = 10$. Aunque esta evidencia contra H_0 no es tan fuerte como la que resultaría de un rechazo en un nivel $\alpha = 0.05$, se trata de información importante para el usuario. De hecho, el uso continuo de $\alpha = 0.05$ o 0.01 tan sólo es un resultado de lo que los estándares han establecido por generaciones. **En la estadística aplicada, los usuarios han adoptado de forma extensa la aproximación del valor *P***. La aproximación se diseña para dar al usuario una alternativa (en términos de una probabilidad) a la simple conclusión de "rechazo" o "no rechazo". El cálculo del valor P también da al usuario información

importante cuando el valor z cae *por completo dentro de la región crítica ordinaria.* Por ejemplo, si z es 2.73, resulta informativo para el usuario observar que

$$P = 2(0.0032) = 0.0064$$

y de esta forma el valor z es significativo a un nivel considerablemente menor que 0.05. Es importante saber que bajo la condición de H_0, un valor de $z = 2.73$ es un evento demasiado raro. A saber, un valor al menos tan grande en magnitud sólo ocurriría 64 veces en 10,000 experimentos.

Demostración gráfica de un valor *P*

Una manera muy simple de explicar gráficamente un valor P es considerar dos muestras distintas. Suponga que se consideran dos materiales para cubrir un tipo específico de metal, con la finalidad de prevenir la corrosión. Se obtienen especímenes y se cubre un grupo con el material 1 y otro grupo con el material 2. Los tamaños muestrales son $n_1 = n_2 = 10$ para cada muestra, y la corrosión se midió en porcentaje del área superficial afectada. La hipótesis es que las muestras provienen de distribuciones comunes con media $\mu = 10$. Supongamos que la varianza poblacional es 1.0. Entonces, probamos que

$$H_0\text{: } \mu_1 = \mu_2 = 10.$$

Representemos con la figura 10.8 una gráfica de puntos de los datos. Los datos se colocan en

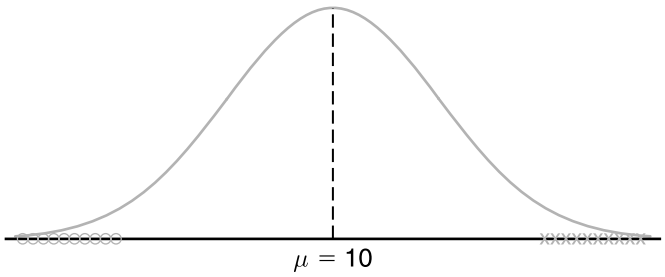


Figura 10.8: Datos probablemente generados de poblaciones que tienen dos medias diferentes.

la distribución que establece la hipótesis nula. Supongamos que los datos "×" se refieren al material 1; y los datos "o", al material 2. Parece claro ahora que los datos en verdad rechazan la hipótesis nula. Pero, ¿cómo se podría resumir esto en un número? **El valor *P* se puede ver simplemente como la probabilidad de obtener este conjunto de datos *dado* que las muestras provienen de la misma distribución**. Es claro que esta probabilidad es bastante pequeña, digamos, ¡0.00000001! De esta manera, el pequeño valor P evidentemente rechaza H_0, y la conclusión es que las medias poblacionales son significativamente diferentes.

La aproximación del valor P como ayuda en la toma de decisiones es bastante natural, ya que casi todos los paquetes computacionales que ofrecen el cálculo de pruebas de hipótesis dan valores P junto con valores del estadístico de prueba adecuado. La siguiente es una definición formal de un valor P.

Definición 10.5: Un valor P es el nivel (de significancia) más bajo donde es significativo el valor observado del estadístico de prueba.

¿En qué difiere el uso de valores P de la prueba de hipótesis clásica?

En este momento resulta tentador resumir los procedimentos que se asocian con la prueba de, digamos, H_0: $\theta = \theta_0$. No obstante, el estudiante que es novato en esta área deberá tener en cuenta que hay diferencias de enfoque y filosofía entre el enfoque clásico de α fija que tiene su punto más importante en la conclusión de "rechace H_0" o "no rechace H_0" y el enfoque del valor P. En este último se determina una α no fija y las conclusiones se obtienen con base en el tamaño del valor P según la apreciación subjetiva del ingeniero o del científico. Sin embargo, mientras que el moderno software computacional produce valores P, es importante que el lector comprenda ambos enfoques para apreciar la totalidad de los conceptos. Por lo tanto, ofrecemos una breve lista con los pasos tanto para el enfoque clásico como para el del valor P.

Aproximación a la prueba de hipótesis con probabilidad fija del error tipo I

1. Establezca las hipótesis nula y alternativa.
2. Elija un nivel de significancia α fijo.
3. Seleccione un estadístico de prueba adecuado y establezca la región crítica con base en α.
4. A partir del estadístico de prueba calculado, rechace H_0 si el estadístico de prueba está en la región crítica. De otra manera, no rechace H_0.
5. Obtenga conclusiones científicas y de ingeniería.

Prueba de significancia (aproximación al valor P)

1. Establezca las hipótesis nula y alternativa.
2. Elija un estadístico de prueba adecuado.
3. Calcule el valor P con base en los valores calculados del estadístico de prueba.
4. Utilice el juicio con base en el valor P y reconozca el sistema científico.

En las secciones de este capítulo y en los capítulos siguientes muchos ejemplos y ejercicios destacarán el enfoque del valor P para obtener conclusiones científicas.

Ejercicios

10.1 Suponga que un alergólogo desea probar la hipótesis de que al menos 30% del público es alérgico a algunos productos de queso. Explique cómo el alergólogo podría cometer

a) un error tipo I;

b) un error tipo II.

10.2 Una socióloga se interesa en la eficacia de un curso de entrenamiento diseñado para lograr que más conductores utilicen los cinturones de seguridad en los automóviles.

a) ¿Qué hipótesis prueba ella si comete un error tipo I al concluir de manera errónea que el curso de entrenamiento no es eficaz?

b) ¿Qué hipótesis prueba ella si comete un error tipo II al concluir de forma errónea que el curso de entrenamiento es eficaz?

10.3 A una empresa manufacturera grande se le acusa de discriminación en sus prácticas de contratación.

a) ¿Qué hipótesis se prueba si un jurado comete un error tipo I al encontrar culpable a dicha empresa?

b) ¿Qué hipótesis se prueba si un jurado comete un error tipo II al encontrar culpable a dicha empresa?

10.4 Se estima que la proporción de adultos que viven en una pequeña ciudad que son graduados universitarios es $p = 0.6$. Para probar esta hipótesis, se selecciona una muestra aleatoria de 15 adultos. Si el número de graduados en nuestra muestra es cualquier número de 6 a 12, aceptaremos la hipótesis nula de que $p = 0.6$; en caso contrario, concluiremos que $p \neq 0.6$.

a) Evalúe α con la suposición de que $p = 0.6$. Utilice la distribución binomial.

b) Evalúe β para las alternativas $p = 0.5$ y $p = 0.7$.

c) ¿Es éste un buen procedimiento de prueba?

10.5 Repita el ejercicio 10.4 cuando se seleccionan 200 adultos y la región de aceptación se define como $110 \leq x \leq 130$, donde x es el número de graduados universitarios en nuestra muestra. Utilice la aproximación normal.

10.6 Un fabricante de telas considera que la proporción de pedidos de materia prima que llegan tarde es $p = 0.6$. Si una muestra aleatoria de 10 pedidos muestra que 3 o menos llegan tarde, la hipótesis de que $p = 0.6$ se debería rechazar a favor de la alternativa $p < 0.6$. Utilice la distribución binomial.

a) Encuentre la probabilidad de cometer un error tipo I si la proporción verdadera es $p = 0.6$.

b) Encuentre la probabilidad de cometer un error tipo II para las alternativas $p = 0.3$, $p = 0.4$ y $p = 0.5$.

10.7 Repita el ejercicio 10.6 cuando se seleccionan 50 pedidos, y se define la región crítica como $x \leq 24$, donde x es el número de pedidos en nuestra muestra que llegan tarde. Utilice la aproximación normal.

10.8 Una tintorería afirma que un nuevo removedor de manchas quitará más de 70% de las manchas en las que se aplique. Para verificar esta afirmación, el removedor de manchas se utilizará sobre 12 manchas que se eligieron al azar. Si menos de 11 de las manchas se eliminan, no rechazaremos la hipótesis nula de que $p = 0.7$; en cualquier otro caso, concluiremos que $p = 0.7$.

a) Evalúe α, suponiendo que $p = 0.7$.

b) Evalúe β para la alternativa $p = 0.9$.

10.9 Repita el ejercicio 10.8 cuando se tratan 100 manchas y la región crítica se define como $x > 82$, donde x es el número de manchas que se eliminan.

10.10 En la publicación *Relief from Arthritis* de Thorsons Publishers, Ltd., John E. Croft afirma que más de 40% de los individuos que sufren de artritis ósea obtienen un alivio mensurable de un ingrediente producido por una especie particular de mejillón que se encuentra en la costa de Nueva Zelanda. Para demostrar tal afirmación, el extracto de mejillón se suministra a un grupo de 7 pacientes con artritis ósea. Si 3 o más de los pacientes obtienen alivio, no rechazaremos la hipótesis nula de que $p = 0.4$; de otro modo, concluiremos que $p < 0.4$.

a) Evalúe α suponiendo que $p = 0.4$.

b) Evalúe β para la alternativa $p = 0.3$.

10.11 Repita el ejercicio 10.10 cuando se administra el extracto de mejillón a 70 pacientes y la región crítica se define como $x < 24$, donde x es el número de pacientes con artritis ósea que obtienen alivio.

10.12 Se pregunta a una muestra aleatoria de 400 votantes en cierta ciudad si están a favor de un impuesto adicional de 4% sobre la venta de gasolina, para obtener los fondos que se necesitan con urgencia para la reparación de calles. Si más de 220 pero menos de 260 favorecen el impuesto a tales ventas, concluiremos que 60% de los votantes lo apoyan.

a) Encuentre la probabilidad de cometer un error tipo I si 60% de los votantes están a favor del aumento de impuestos.

b) ¿Cuál es la probabilidad de cometer un error tipo II al utilizar este procedimiento de prueba si en realidad tan sólo 48% de los votantes está a favor del impuesto adicional a la gasolina?

10.13 Suponga que, en el ejercicio 10.12, concluimos que 60% de los votantes está a favor del impuesto a la venta de gasolina, si más de 214, pero menos de 266, votantes de nuestra muestra lo favorecen. Demuestre que esta nueva región crítica tiene como resultado un valor más pequeño para α a costa de aumentar β.

10.14 Un fabricante desarrolla un nuevo sedal para pesca que, según afirma, tiene una resistencia media a la rotura de 15 kilogramos con una desviación estándar de 0.5 kilogramos. Para probar la hipótesis de que $\mu = 15$ kilogramos contra la alternativa de que $\mu < 15$ kilogramos, se prueba una muestra aleatoria de 50 sedales. La región crítica se define como $\bar{x} < 14.9$.

a) Encuentre la probabilidad de cometer un error tipo I cuando H_0 es verdadera.

b) Evalúe β para las alternativas $\mu = 14.8$ y $\mu = 14.9$ kilogramos.

10.15 En un restaurante de carnes asadas una máquina de bebidas gaseosas se ajusta de manera que la cantidad de bebida que sirva esté distribuida de forma aproximadamente normal, con una media de 200 mililitros y una desviación estándar de 15 mililitros. La máquina se verifica periódicamente tomando una muestra de 9 bebidas y calculando el contenido promedio. Si $\bar{x}$ cae en el intervalo $191 < \bar{x} < 209$, se considera que la máquina opera de forma satisfactoria; de otro modo, concluimos que $\mu \neq 200$ mililitros.

a) Encuentre la probabilidad de cometer un error tipo I cuando $\mu = 200$ mililitros.

b) Encuentre la probabilidad de cometer un error tipo II cuando $\mu = 215$ mililitros.

10.16 Repita el ejercicio 10.15 para muestras de tamaño $n = 25$. Utilice la misma región crítica.

10.17 Se desarrolla una nueva cura para cierto tipo de cemento que tiene como resultado un coeficiente de compresión de 5000 kilogramos por centímetro cuadrado y una desviación estándar de 120. Para probar la hipótesis de que $\mu = 5000$ contra la alternativa de que $\mu < 5000$, se prueba una muestra aleatoria de 50 piezas de cemento. La región crítica se define como $\bar{x} < 4970$.

a) Encuentre la probabilidad de cometer un error tipo I cuando H_0 es verdadera.

b) Evalúe β para las alternativas $\mu = 4970$ y $\mu = 4960$.

10.18 Si graficamos las probabilidades de aceptación de H_0 que corresponden a diversas alternativas para μ (incluido el valor especificado por H_0) y conectamos todos los puntos mediante una curva suave, obtenemos la **curva característica de operación** del criterio de prueba o, simplemente, curva CO. Observe que la probabilidad de aceptación de H_0 cuando es verdadera es simplemente $1 - \alpha$. Las curvas características de operación se utilizan ampliamente en aplicaciones industriales para brindar una muestra visual de los méritos del criterio de prueba. Con referencia al ejercicio 10.15, encuentre las probabilidades de aceptación de H_0 para los siguientes 9 valores de μ y grafique la curva CO: 184, 188, 192, 196, 200, 204, 208, 212 y 216.

10.5 Una sola muestra: Pruebas con respecto a una sola media (varianza conocida)

En esta sección consideramos de manera formal pruebas de hipótesis en una sola media poblacional. Muchas de las ilustraciones de las secciones anteriores incluyen pruebas sobre la media, por lo que el lector ya debería tener una idea de algunos de los detalles que aquí se señalan. Primero deberíamos describir las suposiciones en las que se basa el experimento. El modelo para la situación subyacente se centra alrededor de un experimento con $X_1, X_2, \ldots, X_n$, que representan una muestra aleatoria de una distribución con media μ y varianza $\sigma^2 > 0$. Considere primero la hipótesis

$$H_0\text{: } \mu = \mu_0,$$
$$H_1\text{: } \mu \neq \mu_0.$$

El estadístico de prueba adecuado se debería basar en la variable aleatoria $\bar{X}$. En el capítulo 8 se presentó el teorema del límite central, el cual establece en esencia que sin importar la distribución de X, la variable aleatoria $\bar{X}$ tiene una distribución aproximadamente normal con media μ y varianza σ^2/n para tamaños de muestras razonablemente grandes. De esta manera, $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$. Podemos determinar, entonces, una región crítica basada en el promedio muestral calculado, $\bar{x}$. Debería quedar claro ya al lector que habrá una región crítica de dos colas para la prueba.

Estandarización de $\bar{X}$

Es conveniente estandarizar $\bar{X}$ e incluir de manera formal la variable aleatoria **normal estándar** Z, donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Sabemos que *bajo* H_0, es decir, si $\mu = \mu_0$, entonces $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ tiene una distribución $n(X; 0, 1)$ y, por lo tanto, se puede utilizar la expresión

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

para escribir una región de aceptación adecuada. El lector debería tener en mente que, formalmente, la región crítica se diseña para controlar α, la probabilidad de cometer un error tipo I. Debería ser evidente también que se necesita una *señal* de evidencia *de dos colas* para apoyar H_1. Así, dado un valor calculado $\bar{x}$, la prueba formal implica rechazar H_0 si el *estadístico de prueba* z calculado cae en la región crítica que se describe anteriormente.

Procedimiento de prueba para una sola media

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{o} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

Si $-z_{\alpha/2} < z < z_{\alpha/2}$ no se rechaza H_0. El rechazo de H_0, desde luego, implica la aceptación de la hipótesis alternativa $\mu \neq \mu_0$. Con esta definición de la región crítica debería quedar claro que habrá la probabilidad α de rechazar H_0 (que cae en la región crítica) cuando, en realidad, $\mu = \mu_0$.

Aunque es más fácil entender la región crítica escrita en términos de z, escribimos la misma región crítica en términos del promedio calculado z. Lo siguiente se puede escribir como un procedimiento de decisión idéntico:

$$\text{rechace } H_0 \quad \text{si } \bar{x} < a \quad \text{o} \quad \bar{x} > b,$$

donde

$$a = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad b = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

De aquí, para un nivel de significancia α, los valores críticos de la variable aleatoria z y $\bar{x}$ se representan en la figura 10.9.

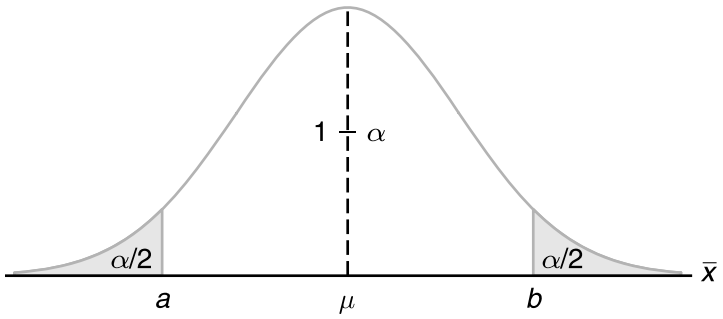


Figura 10.9: Región crítica para la hipótesis alternativa $\mu \neq \mu_0$.

Las pruebas de hipótesis unilaterales sobre la media incluyen el mismo estadístico que se describe en el caso bilateral. La diferencia, por supuesto, es que la región crítica sólo está en una cola de la distribución normal estándar. Como resultado, por ejemplo, supongamos que buscamos probar

$$H_0\colon \mu = \mu_0,$$
$$H_1\colon \mu > \mu_0.$$

La señal que favorece H_1 proviene de *valores grandes* de z. Así, el rechazo de H_0 resulta cuando se calcula $z < z_\alpha$. Evidentemente, si la alternativa es $H_1\colon \mu < \mu_0$, la

región crítica está por completo en la cola inferior, por lo que el rechazo resulta de $z < -z_\alpha$. Aunque en caso de una prueba unilateral la hipótesis nula puede escribirse como H_0: $\mu \leq \mu_0$ o H_0: $\mu \geq \mu_0$, por lo general, se escribe como H_0: $\mu = \mu_0$.

Los siguientes dos ejemplos ilustran pruebas de medias para el caso en el que se conoce σ.

Ejemplo 10.3: Una muestra aleatoria de 100 muertes registradas en Estados Unidos el año pasado mostró una vida promedio de 71.8 años. Suponiendo una desviación estándar poblacional de 8.9 años, ¿esto parece indicar que la vida media actual es mayor que 70 años? Utilice un nivel de significancia de 0.05.

Solución:

1. H_0: $\mu = 70$ años.

2. H_1: $\mu > 70$ años.

3. $\alpha = 0.05$.

4. Región crítica: $z > 1.645$, donde $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.

5. Cálculos: $\bar{x} = 71.8$ años, $\sigma = 8.9$ años y $z = \frac{71.8-70}{8.9/\sqrt{100}} = 2.02$.

6. Decisión: rechace H_0 y concluya que la vida media actual es mayor que 70 años.

En el ejemplo 10.3 el valor P que corresponde a $z = 2.02$ está dado por el área de la región sombreada en la figura 10.10.

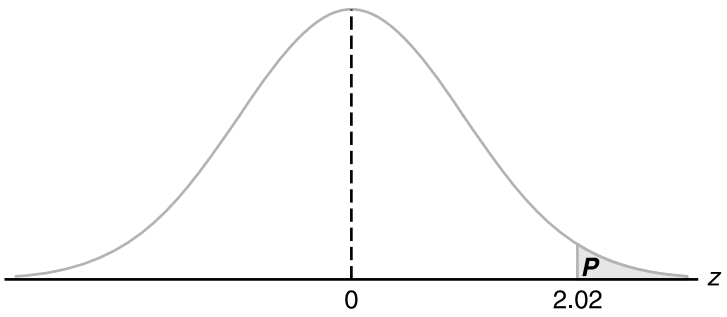


Figura 10.10: Valor P para el ejemplo 10.3.

Usando la tabla A.3, tenemos

$$P = P(Z > 2.02) = 0.0217.$$

Como resultado, la evidencia a favor de H_1 es incluso más fuerte que la sugerida por un nivel de significancia de 0.05.

Ejemplo 10.4: Un fabricante de equipo deportivo desarrolló un nuevo sedal para pesca sintético que afirma que tiene una resistencia media a la rotura de 8 kilogramos con una desviación estándar de 0.5 kilogramos. Pruebe la hipótesis de que $\mu \neq 8$ kilogramos contra la alternativa de que $\mu \neq 8$ kilogramos, si se prueba una muestra aleatoria de 50 sedales y se encuentra que tiene una resistencia media a la rotura de 7.8 kilogramos. Utilice un nivel de significancia de 0.01.

Solución: **1.** H_0: $\mu = 8$ kilogramos.

2. H_1: $\mu \neq 8$ kilogramos.

3. $\alpha = 0.01$.

4. Región crítica: $z < -2.575$ y $z > 2.575$, donde $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.

5. Cálculos: $\bar{x} = 7.8$ kilogramos, $n = 50$ y, de aquí, $z = \frac{7.8-8}{0.5/\sqrt{50}} = -2.83$.

6. Decisión: rechace H_0 y concluya que la resistencia promedio a la rotura no es igual a 8 sino que, de hecho, es menor que 8 kilogramos.

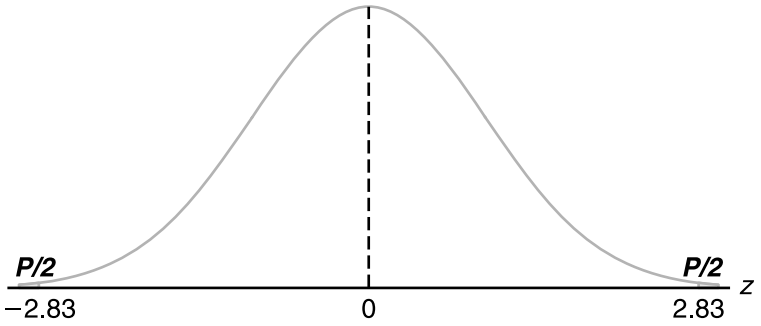


Figura 10.11: Valor P para el ejemplo 10.4.

Como la prueba en este ejemplo es de dos colas, el valor de P que se desea es dos veces el área de la región sombreada de la figura 10.11 a la izquierda de $z = -2.83$. Por lo tanto, con la tabla A.3, tenemos

$$P = P(|Z| > 2.83) = 2P(Z < -2.83) = 0.0046.$$

que nos permite rechazar la hipótesis nula de que $\mu = 8$ kilogramos en un nivel de significancia menor que 0.01.

10.6 Relación con la estimación del intervalo de confianza

El lector ya debería haberse dado cuenta de que, en este capítulo, el enfoque de la prueba de hipótesis para la inferencia estadística está relacionado muy de cerca con el enfoque del intervalo de confianza del capítulo 9. La estimación del intervalo de confianza incluye el cálculo de límites para los cuales es "razonable" que el parámetro en cuestión se encuentre dentro de ellos. Para el caso de una sola media poblacional μ con σ^2 conocida, la estructura tanto de la prueba de hipótesis como de la estimación del intervalo de confianza se basa en la variable aleatoria

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Resulta que la prueba de H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$ a un nivel de significancia α es *equivalente a calcular un intervalo de confianza de* $(1 - \alpha)100\%$ *sobre* μ *y rechazar* H_0, *si* μ_0 *no está dentro del intervalo de confianza.* Si μ_0 está dentro del

intervalo de confianza, no se rechaza la hipótesis. La equivalencia es muy intuitiva y bastante simple de ilustrar. Recuerde que con un valor observado $\bar{x}$ no rechazar H_0 a un nivel de significancia α implica que

$$-z_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2},$$

que es equivalente a

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

La equivalencia del intervalo de confianza con la prueba de hipótesis se extiende a las diferencias entre dos medias, varianzas, razones de varianzas, etcétera. Como resultado, el estudiante de estadística no debería considerar la estimación del intervalo de confianza y la prueba de hipótesis como formas separadas de la inferencia estadística. Entonces, considere el ejemplo 9.2 de la página 275. El intervalo de confianza de 95% sobre la media está dado por los límites (2.50, 2.70). Así, con la misma información muestral, no se rechazará una hipótesis bilateral sobre μ que incluya cualquier valor hipotético entre 2.50 y 2.70. A medida que regresemos a diferentes áreas de la prueba de hipótesis, se seguirá aplicando la equivalencia con la estimación del intervalo de confianza.

10.7 Una sola muestra: Pruebas sobre una sola media (varianza desconocida)

Ciertamente sospecharíamos que las pruebas sobre una media poblacional μ con σ^2 desconocida, como la estimación del intervalo de confianza, debería incluir el uso de la distribución t de Student. Estrictamente hablando, la aplicación de la t de Student tanto para los intervalos de confianza como para la prueba de hipótesis se desarrolla con las siguientes suposiciones. Las variables aleatorias X_1, $X_2, \ldots, X_n$ representan una muestra aleatoria de una distribución normal con μ y σ^2 desconocidas. Entonces, la variable aleatoria $\sqrt{n}(\bar{X} - \mu)/S$ tiene una distribución t de Student con $n - 1$ grados de libertad. La estructura de la prueba es idéntica a la del caso con σ conocida, con la excepción de que el valor σ en el estadístico de prueba se reemplaza con la estimación de S calculada, y la distribución normal estándar se reemplaza con una distribución t. Como resultado, para la hipótesis bilateral

$$H_0\text{: } \mu = \mu_0,$$
$$H_1\text{: } \mu \neq \mu_0,$$

el rechazo de H_0 en un nivel de significancia α resulta cuando un estadístico t calculado

El estadístico t para una prueba en una sola media (varianza desconocida)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

excede $t_{\alpha/2,n-1}$ o es menor que $-t_{\alpha/2,n-1}$. El lector debería recordar de los capítulos 8 y 9 que la distribución t es simétrica alrededor del valor cero. Así, esta región crítica de dos colas se aplica de forma similar a la del caso de σ conocida.

Para la hipótesis bilateral en un nivel de significancia α, se aplican las regiones críticas de dos colas. Para H_1: $\mu > \mu_0$, el rechazo resulta cuando $t > t_{\alpha, n-1}$. Para H_1: $\mu < \mu_0$, la región crítica está dada por $t > -t_{\alpha, n-1}$.

Ejemplo 10.5: El *Instituto Eléctrico Edison* publica cifras del número anual de kilowatts-hora que gastan varios aparatos electrodomésticos. Se afirma que una aspiradora gasta un promedio de 46 kilowatts-hora al año. Si una muestra aleatoria de 12 hogares que se incluye en un estudio planeado indica que las aspiradoras gastan un promedio de 42 kilowatts-hora al año con una desviación estándar de 11.9 kilowatts-hora, ¿en un nivel de significancia de 0.05 esto sugiere que las aspiradoras gastan, en promedio, menos de 46 kilowatts-hora anualmente? Suponga que la población de kilowatts-hora es normal.

Solución:

1. H_0: $\mu = 46$ kilowatts-hora.

2. H_1: $\mu < 46$ kilowatts-hora.

3. $\alpha = 0.05$.

4. Región crítica: $t < -1.796$, donde $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ con 11 grados de libertad.

5. Cálculos: $\bar{x} = 42$ kilowatts-hora, $s = 11.9$ kilowatts-hora y $n = 12$. De aquí,

$$t = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16, \qquad P = P(T < -1.16) \approx 0.135.$$

6. Decisión: no rechace H_0 y concluya que el número promedio de kilowatts-hora que gastan al año las aspiradoras domésticas no es significativamente menor que 46.

Comentario sobre la prueba T de una sola muestra

Es probable que el lector note que se mantiene la equivalencia de la prueba t de dos colas para una sola media y el cálculo de un intervalo de confianza sobre μ con σ reemplazada por s. Así, considere el ejemplo 9.5 de la página 280. En esencia, podemos ver ese cálculo como uno donde encontramos todos los valores de μ_0, el volumen medio hipotético de contenedores de ácido sulfúrico, para los que la hipótesis H_0: $\mu = \mu_0$ *no se rechazará* con $\alpha = 0.05$. Nuevamente, esto es consistente con el planteamiento: "Con base en la información muestral, son razonables los valores del volumen medio de la población entre 9.74 y 10.26 litros."

En este momento vale la pena destacar algunos comentarios con respecto a la suposición de normalidad. Indicamos que cuando se conoce σ, el teorema del límite central permite el uso de un estadístico de prueba o de un intervalo de confianza que se base en Z, la variable aleatoria normal estándar. Estrictamente hablando, por supuesto, el teorema del límite central y, por ello, el uso de la normal estándar no se aplica a menos que se conozca σ. En el capítulo 8, se estudió el desarrollo de la distribución t. En ese momento se estableció que la normalidad sobre $X_1, X_2, \ldots, X_n$ era una suposición básica. Entonces, *en sentido estricto*, las tablas de la t de Student de puntos porcentuales para pruebas o intervalos de confianza no se deberían utilizar, a menos que se sepa que la muestra proviene de una población normal. En la práctica, σ rara vez se puede suponer conocida. Sin embargo, se dispondría de una buena es-

timación a partir de experimentos anteriores. Muchos libros de estadística sugieren que es posible reemplazar con seguridad σ por s en el estadístico de prueba

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

cuando $n \geq 30$ y aún así utilizar las tablas Z para la región crítica adecuada. Aquí, la implicación es que en realidad se recurre al teorema del límite central y se cuenta con el hecho de que $s \approx \sigma$. Evidentemente, cuando se hace esto el resultado se debe ver como aproximación. De esta manera, un valor P calculado (de la distribución Z) de 0.15 puede ser 0.12 o quizá 0.17; o un intervalo de confianza calculado puede ser un intervalo de confianza de 93% en vez de un intervalo de 95% como se desea. Entonces, ¿qué sucede con las situaciones donde $n \leq 30$? El usuario no puede confiar en que s esté cercana a σ, y para tomar en cuenta la inexactitud de la estimación, el intervalo de confianza debería ser más amplio o el valor crítico de mayor magnitud. Los puntos porcentuales de la distribución t realizan esto; pero son correctos sólo cuando la muestra proviene de una distribución normal. Desde luego, se pueden utilizar las gráficas de probabilidad normal para tener alguna noción de la desvación de la normalidad en un conjunto de datos.

Para muestras pequeñas, a menudo resulta difícil detectar desviaciones de una distribución normal. (Las pruebas de la bondad del ajuste se presentan en una sección posterior de este capítulo). Para distribuciones en forma de campana de las variables aleatorias $X_1, X_2, \ldots, X_n$, el uso de la distribución t para pruebas o intervalos de confianza es probable que sea bastante bueno. Cuando haya duda, el usuario debería recurrir a los procedimientos no paramétricos que se presentan en el capítulo 16.

Resultados por computadora comentados para pruebas *T* de una sola muestra

Debería ser de interés para el lector ver resultados por computadora comentados que muestren el resultado de una prueba t de una sola muestra. Suponga que un ingeniero se interesa en probar el sesgo en un medidor de pH. Se reúnen datos de una sustancia neutra (pH = 7.0). Se toma una muestra de las mediciones y los datos son los siguientes:

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08

Entonces, es de interés probar

$$H_0\text{: } \mu = 7.0,$$
$$H_1\text{: } \mu \neq 7.0.$$

En esta ilustración utilizamos el software *MINITAB* para ilustrar el análisis del conjunto de datos anterior. Observe los componentes clave de la salida que se muestra en la figura 10.12. Desde luego, la media $\hat{y} = 7.0250$, StDev es simplemente la desviación estándar de la muestra $s = 0.440$ y SE Mean es el error estándar estimado de la media y se calcula como $s/\sqrt{n} = 0.0139$. El valor t es la razón

$$(7.0250 - 7)/0.0139 = 1.80.$$

El valor P de 0.106 sugiere resultados que no son concluyentes. No hay un rechazo sólido de H_0 (con base en una α de 0.05 o de 0.10) **ni se puede concluir con**

```
pH-meter
  7.07   7.00   7.10   6.97   7.00   7.03   7.01   7.01   6.98   7.08
MTB > Onet 'pH-meter'; SUBC>   Test 7.

One-Sample T: pH-meter Test of mu = 7 vs not = 7
Variable  N    Mean   StDev  SE Mean      95% CI          T      P
pH-meter 10 7.02500 0.04403  0.01392 (6.99350, 7.05650) 1.80 0.106
```

Figura 10.12: Salida de *MINITAB* para la prueba t de una muestra para el medidor de pH.

certeza que el medidor de pH está insesgado. Observe que el tamaño de la muestra de 10 es bastante pequeño. Un aumento en el tamaño de la muestra (quizás otro experimento) podría arreglar la situación. En la sección 10.10 aparece un análisis con respecto al tamaño de la muestra adecuado.

10.8 Dos muestras: Pruebas sobre dos medias

El lector ya llegó a comprender la relación entre pruebas e intervalos de confianza, y puede confiar por completo en los detalles que ofrece el material sobre el intervalo de confianza del capítulo 9. Las pruebas con respecto a dos medias representan un conjunto de herramientas analíticas muy importantes para el científico o el ingeniero. El procedimiento experimental es muy parecido al que se describe en la sección 9.8. Se extraen dos muestras aleatorias independientes de tamaño n_1 y n_2, respectivamente, de dos poblaciones con medias μ_1 y μ_2, y varianzas σ_1^2 y σ_2^2. Sabemos que la variable aleatoria

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

tiene una distribución normal estándar. Suponemos aquí que n_1 y n_2 son suficientemente grandes, por lo que se aplica el teorema del límite central. Por supuesto, si las dos poblaciones son normales, el estadístico anterior tiene una distribución normal estándar aun para n_1 y n_2 pequeñas. Evidentemente, si podemos suponer que $\sigma_1 = \sigma_2 = \sigma$, el estadístico anterior se reduce a

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}.$$

Los dos estadísticos anteriores sirven como base para el desarrollo de los procedimientos de prueba que incluyen dos medias. La equivalencia con el intervalo de confianza y la facilidad de la transición del caso de pruebas sobre una sola media hacen que esto sea sencillo.

La hipótesis bilateral sobre dos medias se escribe con bastante generalidad como

$$H_0\text{: } \mu_1 - \mu_2 = d_0.$$

En efecto, la alternativa puede ser bilateral o unilateral. De nuevo, la distribución que se utiliza es la distribución del estadístico de prueba bajo H_0. Se calculan los valores $\bar{x}_1$ y $\bar{x}_2$, y para σ_1 y σ_2 conocidas, el estadístico de prueba está dado por

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

con una región crítica de dos colas en el caso de una alternativa bilateral. Es decir, rechazace H_0 a favor de H_1: $\mu_1 - \mu_2 \neq d_0$ si $z > z_{\alpha/2}$ o $z < -z_{\alpha/2}$. Las regiones críticas de una cola se utilizan en el caso de alternativas unilaterales. El lector debería estudiar, como antes, el estadístico de prueba y quedar satisfecho de que para, digamos, H_1: $\mu_1 - \mu_2 > d_0$, la señal que favorece H_1 provenga de valores grandes de z. De esta manera se aplica la región crítica de la cola superior.

Varianzas desconocidas pero iguales

Las situaciones que más prevalecen que implican pruebas sobre dos medias son aquellas con varianzas desconocidas. Si el científico interesado está dispuesto a suponer que ambas distribuciones son normales y que $\sigma_1 = \sigma_2 = \sigma$, se puede utilizar la *prueba t combinada* (a menudo llamada prueba t de dos muestras). El estadístico de prueba (véase la sección 9.8) está dado por el siguiente procedimiento de prueba.

Prueba T combinada de dos muestras

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}},$$

donde

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}.$$

Se incluye la distribución t y *no se rechaza* la hipótesis bilateral cuando

$$-t_{\alpha/2,n_1+n_2-2} < t < t_{\alpha/2,n_1+n_2-2}.$$

Del material del capítulo 9 recuerde que los grados de libertad para la distribución t son un resultado de la combinación de la información de las dos muestras para estimar σ^2. Las alternativas unilaterales sugieren regiones críticas unilaterales, como era de esperarse. Por ejemplo, para H_1: $\mu_1 - \mu_2 > d_0$, rechace H_1: $\mu_1 - \mu_2 = d_0$ cuando $t > t_{\alpha,n_1+n_2-2}$.

Ejemplo 10.6: Se lleva a cabo un experimento para comparar el desgaste por abrasivos de dos diferentes materiales laminados. Se prueban 12 piezas del material 1 exponiendo cada pieza a una máquina para medir el desgaste. Diez piezas del material 2 se prueban de manera similar. En cada caso, se observa la profundidad del desgaste. Las muestras del material 1 dan un desgaste promedio (codificado) de 85 unidades con una desviación estándar muestral de 4; en tanto que las muestras del material 2 dan un promedio de 81 y una desviación estándar muestral de 5. ¿Podríamos concluir, con un nivel de significancia de 0.05, que el desgaste abrasivo del material 1 excede el del material 2 en más de 2 unidades? Suponga que las poblaciones son aproximadamente normales con varianzas iguales.

Solución: Representemos con μ_1 y μ_2 las medias poblacionales del desgaste abrasivo para el material 1 y el material 2, respectivamente.

1. H_0: $\mu_1 - \mu_2 = 2$.
2. H_1: $\mu_1 - \mu_2 > 2$.
3. $\alpha = 0.05$.
4. Región crítica: $t > 1.725$, donde $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$ con $v = 20$ grados de libertad.

5. Cálculos:

$$\bar{x}_1 = 85, \qquad s_1 = 4, \qquad n_1 = 12,$$
$$\bar{x}_2 = 81, \qquad s_2 = 5, \qquad n_2 = 10.$$

De aquí,

$$s_p = \sqrt{\frac{(11)(16) + (9)(25)}{12 + 10 - 2}} = 4.478,$$
$$t = \frac{(85 - 81) - 2}{4.478\sqrt{1/12 + 1/10}} = 1.04,$$
$$P = P(T > 1.04) \approx 0.16. \qquad \text{(Véase la tabla A.4.)}$$

6. Decisión: no rechace H_0. Somos incapaces de concluir que el desgaste abrasivo del material 1 excede el del material 2 en más de 2 unidades.

Varianzas desconocidas pero diferentes

Hay situaciones donde el analista **no** es capaz de suponer que $\sigma_1 = \sigma_2$. Del capítulo 9 recuerde que, si las poblaciones son normales, el estadístico

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

tiene una distribución t aproximada con grados de libertad aproximados

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Como resultado el procedimiento de prueba es *no rechazar* H_0 cuando

$$-t_{\alpha/2,v} < t' < t_{\alpha/2,v},$$

con v dado como antes. De nuevo, como en el caso de la prueba t combinada, las alternativas unilaterales sugieren regiones críticas unilaterales.

Observaciones pareadas

Cuando el aprendiz de estadística estudia la prueba t de dos muestras o el intervalo de confianza sobre la diferencia entre medias, se debería dar cuenta de que algunas nociones elementales que se tratan en el diseño experimental se vuelven relevantes y se deben considerar. Recuerde la discusión sobre las unidades experimentales en el capítulo 9, donde se sugirió en ese momento que la condición de las dos poblaciones (a menudo denominadas los dos tratamientos) se deberían asignar de manera aleatoria a las unidades experimentales. Esto se realiza para evitar resultados sesgados debido a las diferencias sistemáticas entre unidades experimentales. En otras palabras, en términos de la jerga en prueba de hipótesis, es importante que la diferencia significativa que se encuentra (o que no se encuentra) entre las medias se deba a las diferentes condiciones de las poblaciones, y no a las unidades experimentales en el estudio. Por ejemplo, considere el ejercicio 9.40 de la sección 9.9. Los 20 retoños

juegan el papel de unidades experimentales. Diez de ellas se tratan con nitrógeno y 10 sin nitrógeno. Puede ser muy importante que esta asignación al tratamiento "con nitrógeno" y "sin nitrógeno" sea aleatoria, para asegurar que las diferencias sistemáticas entre los retoños no interfieran con una comparación válida entre las medias.

En el ejemplo 10.6, el tiempo de medición es la elección más probable de la unidad experimental. Las 22 piezas de material se deberían medir en orden aleatorio. Necesitamos protegernos contra la posibilidad de que las mediciones del desgaste que se realicen casi al mismo tiempo tiendan a dar resultados similares. *No se esperan* **diferencias sistemáticas** (no aleatorias) **en las unidades experimentales**. Sin embargo, las asignaciones aleatorias protegen contra el problema.

Referencias a la planeación de experimentos, aleatorización, elección del tamaño de la muestra, etcétera, continuarán influyendo en gran parte del desarrollo en los capítulos 13, 14 y 15. Cualquier científico o ingeniero cuyo interés resida en el análisis de datos reales debería estudiar este material. La prueba t combinada se amplía en el capítulo 13 para cubrir más de dos medias.

La prueba de dos medias se puede llevar a cabo cuando los datos están en la forma de observaciones pareadas como se estudió en el capítulo 9. En esta estructura de pareamiento, las condiciones de las dos poblaciones (tratamientos) se asignan de forma aleatoria dentro de unidades homogéneas. El cálculo del intervalo de confianza para $\mu_1 - \mu_2$ en la situación con observaciones pareadas se basa en la variable aleatoria

$$T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}},$$

donde $\bar{D}$ y S_d son variables aleatorias que representan la media muestral y las desviaciones estándar de las diferencias de las observaciones en las unidades experimentales. Como en el caso de la *prueba t combinada*, la suposición es que las observaciones de cada población son normales. Este problema de dos muestras se reduce en esencia a un problema de una muestra utilizando las diferencias calculadas d_1, $d_2, \ldots, d_n$. De esta manera, la hipótesis se reduce a

$$H_0\colon\ \mu_D = d_0.$$

El estadístico de prueba calculado está dado entonces por

$$t = \frac{\overline{d} - d_0}{s_d/\sqrt{n}}.$$

Las regiones críticas se construyen usando la distribución t con $n - 1$ grados de libertad.

Ejemplo 10.7: En un estudio realizado en el Departamento de Silvicultura y Fauna del Instituto Politécnico y Universidad Estatal de Virginia, J. A. Wesson examinó la influencia del fármaco *succinylcholine* sobre los niveles de circulación de andrógenos en la sangre. Se obtuvieron muestras sanguíneas de venados salvajes vía la vena yugular, inmediatamente después de una inyección intramuscular de succinylcholine con dardos de un rifle de caza. Los venados se sangraron nuevamente aproximadamente 30 minutos después de la inyección y luego se liberaron. Los niveles de andrógenos al momento de la captura y 30 minutos más tarde, medidos en nanogramos por mililitro (ng/ml), para 15 venados se presentan en la tabla 10.2.

Tabla 10.2: Datos para el ejemplo 10.7

Venado	Andrógenos (ng/ml) Al momento de la inyección	Andrógenos (ng/ml) 30 minutos después de la inyección	d_i
1	2.76	7.02	4.26
2	5.18	3.10	2.08
3	2.68	5.44	2.76
4	3.05	3.99	0.94
5	4.10	5.21	1.11
6	7.05	10.26	3.21
7	6.60	13.91	7.31
8	4.79	18.53	13.74
9	7.39	7.91	0.52
10	7.30	4.85	−2.45
11	11.78	11.10	−0.68
12	3.90	3.74	−0.16
13	26.00	94.03	68.03
14	67.48	94.03	26.55
15	17.04	41.70	24.66

Suponiendo que las poblaciones de andrógenos al momento de la inyección y 30 minutos después se distribuyen normalmente, pruebe con un nivel de significancia de 0.05 si las concentraciones de andrógenos se alteran después de 30 minutos de encierro.

Solución: Sean μ_1 y μ_2 la concentración promedio de andrógenos al momento de la inyección y 30 minutos después, respectivamente. Procedemos como sigue:

1. H_0: $\mu_1 = \mu_2$ o $\mu_D = \mu_1 - \mu_2 = 0$.

2. H_1: $\mu_1 \neq \mu_2$ o $\mu_D = \mu_1 - \mu_2 \neq 0$.

3. $\alpha = 0.05$.

4. Región crítica: $t < -2.145$ y $t > 2.145$, donde $t = \frac{\overline{d} - d_0}{s_D/\sqrt{n}}$ con $v = 14$ grados de libertad.

5. Cálculos: La media muestral y la desviación estándar para las d_i son

$$\bar{d} = 9.848 \qquad \text{y} \qquad s_d = 18.474.$$

Por lo tanto,

$$t = \frac{9.848 - 0}{18.474/\sqrt{15}} = 2.06.$$

6. Aunque el estadístico t no es significativo al nivel 0.05, de la tabla A.4,

$$P = P(|T| > 2.06) \approx 0.06.$$

Como resultado, existe alguna evidencia de que hay una diferencia en los niveles medios circulantes de andrógenos.

En el caso de observaciones pareadas, es importante que no haya interacción entre los tratamientos y las unidades experimentales. Esto se discutió en el capítulo 9

en el desarrollo de intervalos de confianza. La suposición de no interacción implica que el efecto de la unidad experimental, o pareada, es el mismo para cada uno de los dos tratamientos. En el ejemplo 10.7, suponemos que el efecto en el venado es el mismo para las dos condiciones que se estudian; a saber, "en el momento de la inyección" y 30 minutos después de la inyección.

Salidas por computadora comentadas para la prueba *T* pareada

La figura 10.13 muestra una salida por computadora del SAS para una prueba t pareada usando los datos del ejemplo 10.7. Observe que la apariencia de la salida es la de una prueba t de una sola muestra y, por supuesto, esto es exactamente lo que se realizó, ya que la prueba busca determinar si $\bar{d}$ es significativamente diferente de cero.

```
                    Analysis Variable : Diff

 N            Mean       Std Error     t Value     Pr > |t|
-----------------------------------------------------------
15       9.8480000       4.7698699        2.06       0.0580
-----------------------------------------------------------
```

Figura 10.13: Salida del SAS de la prueba t pareada para los datos del ejemplo 10.7.

Resumen de los procedimientos de prueba

Dado que completamos el desarrollo formal de pruebas sobre medias poblacionales, ofrecemos la tabla 10.2 que resume el procedimiento de prueba para los casos de una sola media y de dos medias. Note el procedimiento aproximado cuando las distribuciones son normales y las varianzas se desconocen pero no se suponen iguales. Este estadístico se estudió en el capítulo 9.

10.9 Elección del tamaño de la muestra para probar medias

En la sección 10.2 demostramos cómo el analista puede explotar las relaciones entre el tamaño de la muestra, el nivel de significancia α y la potencia de la prueba para alcanzar cierto estándar de calidad. En la mayoría de las circunstancias prácticas el experimento debería planearse con una elección de un tamaño muestral que se realiza antes del proceso de recolección de datos, si es posible. Por lo general, el tamaño de la muestra se establece para lograr una buena potencia para una α fija y una alternativa específica fija. Esta alternativa fija puede estar en la forma de $\mu - \mu_0$ en el caso de una hipótesis que incluya una sola media, de o $\mu_1 - \mu_2$ en el caso de un problema que implique dos medias. Los casos específicos serán ilustrativos.

Suponga que deseamos probar la hipótesis

$$H_0\text{: } \mu = \mu_0,$$
$$H_1\text{: } \mu > \mu_0.$$

Tabla 10.3: Pruebas relacionadas con medias

H_0	Valor del estadístico de prueba	H_1	Región crítica
$\mu = \mu_0$	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$; σ conocida	$\mu < \mu_0$	$z < -z_\alpha$
		$\mu > \mu_0$	$z > z_\alpha$
		$\mu \neq \mu_0$	$z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$; $v = n-1$,	$\mu < \mu_0$	$t < -t_\alpha$
	σ desconocida	$\mu > \mu_0$	$t > t_\alpha$
		$\mu \neq \mu_0$	$t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$;	$\mu_1 - \mu_2 < d_0$	$z < -z_\alpha$
	σ_1 y σ_2 conocida	$\mu_1 - \mu_2 > d_0$	$z > z_\alpha$
		$\mu_1 - \mu_2 \neq d_0$	$z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{s_p\sqrt{1/n_1+1/n_2}}$;	$\mu_1 - \mu_2 < d_0$	$t < -t_\alpha$
	$v = n_1 + n_2 - 2$,	$\mu_1 - \mu_2 > d_0$	$t > t_\alpha$
	$\sigma_1 = \sigma_2$ pero desconocida	$\mu_1 - \mu_2 \neq d_0$	$t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
	$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$		
$\mu_1 - \mu_2 = d_0$	$t' = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$;	$\mu_1 - \mu_2 < d_0$	$t' < -t_\alpha$
	$v = \frac{(s_1^2/n_1+s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}$;	$\mu_1 - \mu_2 > d_0$	$t' > t_\alpha$
	$\sigma_1 \neq \sigma_2$ y desconocida	$\mu_1 - \mu_2 \neq d_0$	$t' < -t_{\alpha/2}$ o $t' > t_{\alpha/2}$
$\mu_D = d_0$	$t = \frac{\bar{d}-d_0}{s_d/\sqrt{n}}$; $v = n-1$	$\mu_D < d_0$	$t < -t_\alpha$
observaciones		$\mu_D > d_0$	$t > t_\alpha$
pareadas		$\mu_D \neq d_0$	$t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$

Con un nivel de significancia α cuando se conoce la varianza σ^2. Para una alternativa específica, digamos, $\mu = \mu_0 + \delta$, en la figura 10.14, se muestra que la potencia de nuestra prueba es

$$1 - \beta = P(\bar{X} > a \text{ cuando } \mu = \mu_0 + \delta).$$

Por lo tanto,

$$\begin{aligned}\beta &= P(\bar{X} < a \text{ cuando } \mu = \mu_0 + \delta) \\ &= P\left[\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} < \frac{a - (\mu_0 + \delta)}{\sigma/\sqrt{n}} \text{ cuando } \mu = \mu_0 + \delta\right].\end{aligned}$$

Bajo la hipótesis alternativa $\mu = \mu_0 + \delta$, el estadístico

$$\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}}$$

es la variable normal estándar Z. Por lo tanto,

$$\beta = P\left(Z < \frac{a - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{\sigma/\sqrt{n}}\right) = P\left(Z < z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right),$$

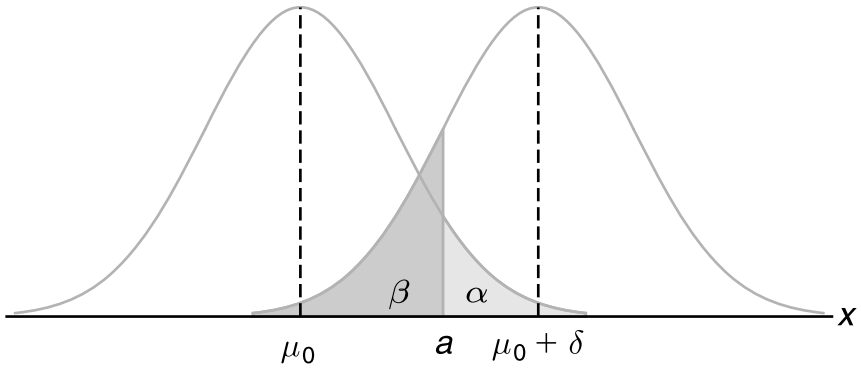


Figura 10.14: Prueba de $\mu = \mu_0$ contra $\mu = \mu_0 + \delta$.

de donde concluimos que

$$-z_\beta = z_\alpha - \frac{\delta\sqrt{n}}{\sigma},$$

y de aquí,

$$\text{Selección del tamaño de la muestra: } n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{\delta^2},$$

resultado que también es verdadero cuando la hipótesis alternativa es $\mu < \mu_0$.

En el caso de una prueba de dos colas, obtenemos la potencia $1 - \beta$ para una alternativa específica cuando

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}.$$

Ejemplo 10.8: Suponga que deseamos probar la hipótesis

$$H_0\text{: } \mu = 68 \text{ kilogramos},$$
$$H_1\text{: } \mu > 68 \text{ kilogramos},$$

para los pesos de estudiantes hombres en cierta universidad usando un nivel de significancia $\alpha = 0.05$ cuando se sabe que $\sigma = 5$. Encuentre el tamaño muestral que se requiere si la potencia de nuestra prueba debe ser 0.95 cuando la media real es 69 kilogramos.

Solución: Como $\alpha = \beta = 0.05$, tenemos $z_\alpha = z_\beta = 1.645$. Para la alternativa $\beta = 69$, tomamos $\delta = 1$ y, entonces,

$$n = \frac{(1.645 + 1.645)^2(25)}{1} = 270.6.$$

Por lo tanto, se requieren 271 observaciones si la prueba debe rechazar la hipótesis nula 95% de las veces cuando, de hecho, μ es tan grande como 69 kilogramos.

El caso de dos muestras

Se puede utilizar un procedimiento similar para determinar el tamaño de la muestra $n = n_1 = n_2$ que se requiere para una potencia específica de la prueba en que se comparan dos medias poblacionales. Por ejemplo, suponga que deseamos probar la hipótesis

$$H_0\colon \mu_1 - \mu_2 = d_0,$$
$$H_1\colon \mu_1 - \mu_2 \neq d_0,$$

cuando se conocen σ_1 y σ_2. Para una alternativa específica, digamos, $\mu_1 - \mu_2 = d_0 + \delta$, en la figura 10.15 se muestra que la potencia de nuestra prueba es

$$1 - \beta = P(|\bar{X}_1 - \bar{X}_2| > a \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta.$$

Por lo tanto,

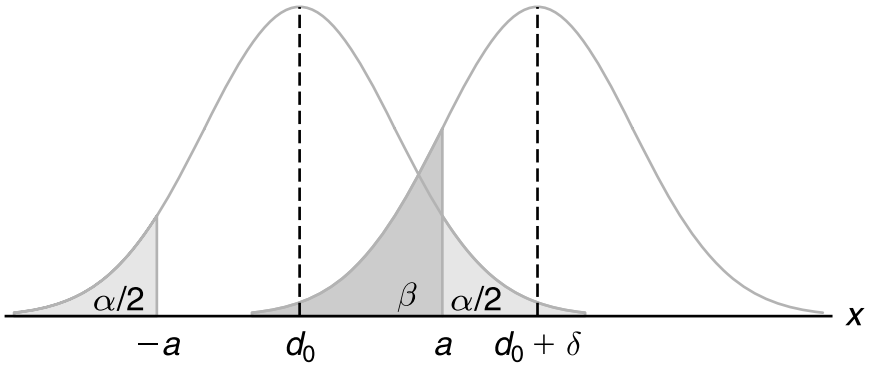


Figura 10.15: Prueba de $\mu_1 - \mu_2 = d_0$ contra $\mu_1 - \mu_2 = d_0 + \delta$.

$$\begin{aligned}\beta &= P(-a < \bar{X}_1 - \bar{X}_2 < a \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta) \\ &= P\left[\frac{-a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < \frac{(\bar{X}_1 - \bar{X}_2) - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right. \\ &\qquad \left. < \frac{a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta\right].\end{aligned}$$

Con la hipótesis alternativa $\mu_1 - \mu_2 = d_0 + \delta$, el estadístico

$$\frac{\bar{X}_1 - \bar{X}_2 - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}$$

es la variable normal estándar Z. Ahora bien, al escribir

$$-z_{\alpha/2} = \frac{-a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \quad \text{y} \quad z_{\alpha/2} = \frac{a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

tenemos

$$\beta = P\left[-z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2+\sigma_2^2)/n}} < Z < z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2+\sigma_2^2)/n}}\right],$$

de donde concluimos que

$$-z_\beta \approx z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2+\sigma_2^2)/n}},$$

y, por lo tanto,

$$n \approx \frac{(z_{\alpha/2}+z_\beta)^2(\sigma_1^2+\sigma_2^2)}{\delta^2}.$$

Para la prueba de una sola cola, la expresión para el tamaño requerido de la muestra cuando $n = n_1 = n_2$ es

$$\text{Elección del tamaño de la muestra: } n = \frac{(z_\alpha+z_\beta)^2(\sigma_1^2+\sigma_2^2)}{\delta^2}.$$

Cuando se desconoce la varianza poblacional (o varianzas en la situación de dos muestras), la elección del tamaño de la muestra no es directa. Al probar la hipótesis $\mu = \mu_0$ cuando el valor real es $\mu = \mu_0 + \delta$, el estadístico

$$\frac{\bar{X} - (\mu_0+\delta)}{S/\sqrt{n}}$$

no sigue la distribución t, como podría esperarse, sino que más bien sigue la **distribución *t* no central**. Sin embargo, existen tablas o gráficas que se basan en la distribución t no central para determinar el tamaño adecuado de la muestra, si se dispone de alguna estimación de σ o si δ es un múltiplo de σ. La tabla A.8 da los tamaños muestrales necesarios para controlar los valores de α y δ para diversos valores de

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu-\mu_0|}{\sigma}$$

para pruebas de una y de dos colas. En el caso de la prueba t de dos muestras en la que se desconocen las varianzas, pero se suponen iguales, obtenemos los tamaños muestrales $n = n_1 = n_2$ necesarios para controlar los valores de α y β para diversos valores de

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu_1-\mu_2-d_0|}{\sigma}$$

de la tabla A.9.

Ejemplo 10.9: Al comparar el comportamiento de dos catalizadores sobre el efecto del rendimiento de una reacción, se realiza una prueba t de dos muestras con $\alpha = 0.05$. Las varianzas de los rendimientos se consideran las mismas para los dos catalizadores. ¿De qué tamaño se necesita una muestra para cada catalizador, si se desea probar la hipótesis

$$H_0\text{: } \mu_1 = \mu_2,$$
$$H_1\text{: } \mu_1 \neq \mu_2,$$

si es esencial detectar una diferencia de 0.8σ entre los catalizadores con probabilidad 0.9?

Solución: De la tabla A.9, con $\alpha = 0.05$ para una prueba de dos colas, $\beta = 0.1$, y

$$\Delta = \frac{|0.8\sigma|}{\sigma} = 0.8,$$

encontramos que el tamaño de la muestra que se requiere es $n = 34$.

Se enfatiza que en situaciones prácticas sería difícil forzar a un científico o a un ingeniero a hacer un compromiso sobre la información de la que se puede encontrar un valor de Δ. Se recuerda al lector que el valor Δ cuantifica la clase de diferencia entre las medias que el científico considera importantes; es decir, una diferencia que se considere *significativa* desde un punto de vista científico, no estadístico. El ejemplo 10.9 ilustra cómo se hace a menudo esta elección; a saber, mediante la selección de una fracción de σ. Evidentemente, si el tamaño de la muestra se basa en una elección de $|\delta|$ que es una fracción pequeña de σ, el tamaño muestral que resulta puede ser bastante grande comparado con lo que permite el estudio.

10.10 Métodos gráficos para comparar medias

En el capítulo 3 se pone una considerable atención hacia la presentación de datos en forma gráfica. Los diagramas de tallo y hojas, en el capítulo 8, y las gráficas de caja y extensión, gráficas de cuantiles y gráficas normales cuantil-cuantil se utilizan para brindar una "imagen" y resumir así un conjunto de datos experimentales. Muchos paquetes de software computacional producen representaciones gráficas. A medida que procedamos con otras formas de análisis de datos (por ejemplo, el análisis de regresión y el análisis de varianza), los métodos gráficos se vuelven aún más informativos.

Las ayudas gráficas que se utilizan junto con la prueba de hipótesis no se usan como un reemplazo del procedimiento de prueba. En realidad, el valor del estadístico de prueba indica el tipo adecuado de evidencia en apoyo de H_0 o H_1. Sin embargo, una representación como imagen ofrece una buena ilustración y a menudo es un mejor comunicador de evidencia para el beneficiario del análisis. Además, una imagen con frecuencia dejará claro por qué se encontró una diferencia significativa. La falla de una suposición importante se puede descubrir mediante un resumen gráfico.

Para la comparación de las medias, las gráficas de caja y extensión de lado a lado tienen una presentación reveladora. El lector debería recordar que estas gráficas muestran el percentil 25, el percentil 75 y la mediana en un conjunto de datos. Además, las extensiones muestran los extremos en un conjunto de datos. Considere el ejercicio 10.40 que sigue a esta sección. Se miden los niveles en plasma de ácido ascórbico en dos grupos de mujeres embarazadas: fumadoras y no fumadoras. La figura 10.16 muestra las gráficas de caja y extensión para ambos grupos de mujeres. Dos cuestiones son muy evidentes. Al tomar en cuenta la variabilidad en los dos grupos parece haber una diferencia despreciable en las medias muestrales. Además, la variabilidad en los dos grupos parece ser algo diferente. Por supuesto, el analista debe tener presentes, en este caso, las diferencias más bien considerables entre los tamaños muestrales.

Considere el ejercicio 9.40 de la sección 9.9. La figura 10.17 presenta la gráfica múltiple de caja y extensión para los datos de 10 retoños, la mitad con nitrógeno y la otra mitad sin nitrógeno. Tal gráfica revela una valiabilidad menor para el grupo

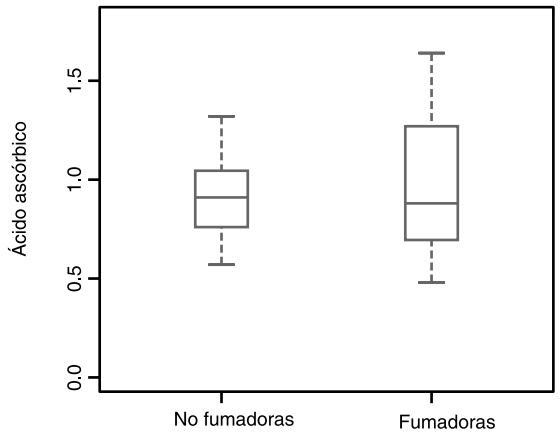


Figura 10.16: Dos gráficas de caja y extensión de los datos de ácido ascórbico para mujeres fumadoras y no fumadoras.

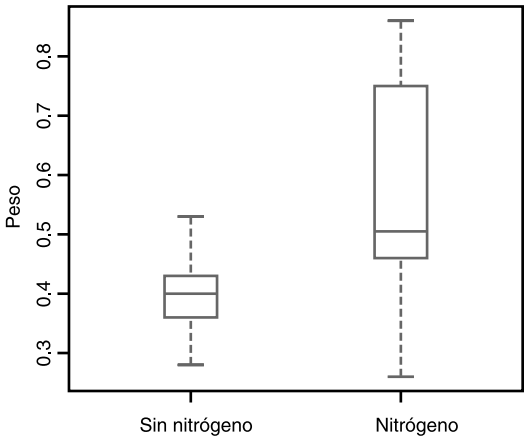


Figura 10.17: Dos gráficas de caja y extensión para los datos de los retoños.

sin nitrógeno. Además, la falta de traslape de las cajas sugiere una diferencia significativa entre los pesos medios de los tallos entre ambos grupos. Parecería que la presencia de nitrógeno aumenta el peso de los tallos y quizás aumente la variabilidad en los pesos.

No existen ciertas reglas generales con respecto a cuando las gráficas de caja y extensión brindan evidencia de diferencias significativas entre las medias. Sin embargo, una pauta aproximada es que si la línea del percentil 25 para una muestra excede la línea de la mediana de la otra muestra, hay evidencia sólida de una diferencia entre las medias.

Se hará más énfasis en los métodos gráficos en un estudio de caso de la vida real, que se presenta más adelante en este capítulo.

Salida por computadora comentada para una prueba T de dos muestras

Considere los datos del ejercicio 9.40, sección 9.9, donde se reunieron los datos de los retoños en condiciones con nitrógeno y sin nitrógeno. Pruebe

$$H_0: \mu_{\text{NIT}} = \mu_{\text{NON}}$$
$$H_1: \mu_{\text{NIT}} > \mu_{\text{NON}}$$

donde las medias poblacionales indican los pesos medios. La figura 10.18 es una salida por computadora comentada del paquete SAS. Observe que las desviaciones estándar y el error estándar se muestran para ambas muestras. Se da el estadístico t bajo la suposición de "varianza igual" y de "varianza diferente". De la gráfica de caja y extensión de la figura 10.17 en realidad parecería que se transgrede la suposición de varianza igual. Un valor P de 0.0229 sugiere una conclusión de medias diferentes. Esto coincide con la información de diagnóstico que se da en la figura 10.18. A propósito, observe que t y t' son iguales en este caso, ya que $n_1 = n_2$.

```
                TTEST Procedure
Variable Weight
    Mineral     N    Mean   Std Dev Std Err
No nitrogen    10  0.3990    0.0728  0.0230
   Nitrogen    10  0.5650    0.1867  0.0591

Variances      DF    t Value    Pr > |t|
    Equal      18       2.62      0.0174
  Unequal    11.7       2.62      0.0229

          Test the Equality of Variances
Variable   Num DF    Den DF    F Value    Pr > F
  Weight        9         9       6.58    0.0098
```

Figura 10.18: Salida del SAS para la prueba t de dos muestras.

Ejercicios

10.19 Una empresa de material eléctrico fabrica bombillas de luz que tienen una duración que se distribuye de forma aproximadamente normal con una media de 800 horas y una desviación estándar de 40 horas. Pruebe la hipótesis de que $\mu = 800$ horas contra la alternativa $\mu \neq 800$ horas, si una muestra aleatoria de 30 bombillas tiene una duración promedio de 788 horas. Utilice un valor P en su respuesta.

10.20 Una muestra aleatoria de 64 bolsas de palomitas (rosetas) de maíz con queso chedar pesan, en promedio, 5.23 onzas con una desviación estándar de 0.24 onzas. Pruebe la hipótesis de que $\mu = 5.5$ onzas contra la hipótesis alternativa, $\mu < 5.5$ onzas con un nivel de significancia de 0.05.

10.21 En un informe de investigación de Richard H. Weindruch de la Escuela de Medicina de la UCLA, se afirma que los ratones con una vida promedio de 32 meses vivirían hasta alrededor de 40 meses de edad, cuando 40% de las calorías en su dieta se reemplacen con vitaminas y proteínas. ¿Hay alguna razón para creer que $\mu < 40$, si 64 ratones que se sujetan a esa dieta tienen una vida promedio de 38 meses con una desviación estándar de 5.8 meses? Utilice un valor P en su conclusión.

10.22 La estatura promedio de mujeres en el grupo de primer año de cierta universidad es de 162.5 centímetros con una desviación estándar de 6.9 centímetros. ¿Hay alguna razón para creer que hay un cambio en la estatura promedio, si una muestra aleatoria de 50 mujeres en el grupo actual de primer año tiene una altura promedio de 165.2 centímetros? Utilice un valor P en su conclusión. Suponga que la desviación estándar permanece constante.

10.23 Se afirma que un automóvil se maneja en promedio más de 20,000 kilómetros por año. Para probar tal afirmación, se pide a una muestra de 100 propietarios de automóviles que lleven un registro de los kilómetros que recorran. ¿Estaría usted de acuerdo con esta afirmación, si la muestra aleatoria mostró un promedio de 23,500 kilómetros y una desviación estándar de 3900 kilómetros? Utilice un valor P en su conclusión.

10.24 En el boletín de la Asociación Estadounidense del Corazón, *Hypertension*, investigadores reportan que los individuos que practican la meditación trascendental (MT) bajan su presión sanguínea de forma significativa. Si una muestra aleatoria de 225 hombres practicantes de MT meditan 8.5 horas a la semana, con una desviación estándar de 2.25 horas, ¿esto sugiere que, en promedio, los hombres que utilizan la MT meditan más de 8 horas a la semana? Cite un valor P en su conclusión.

10.25 Pruebe la hipótesis de que el contenido promedio de los envases de un lubricante específico es de 10 litros, si los contenidos de una muestra aleatoria de 10 envases son 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 y 9.8 litros. Utilice un nivel de significancia de 0.01 y suponga que la distribución del contenido es normal.

10.26 De acuerdo con un estudio dietético una ingesta alta de sodio se puede relacionar con úlceras, cáncer estomacal y migraña. El requerimiento humano de sal es de tan sólo 220 miligramos diarios, el cual se rebasa en la mayoría de las porciones individuales de cereales listos para comerse. Si una muestra aleatoria de 20 porciones similares de cierto cereal tiene un contenido medio de 244 miligramos de sodio y una desviación estándar de 24.5 miligramos, ¿esto sugiere, en el nivel de significancia de 0.05, que el contenido promedio de

sodio para porciones individuales de tal cereal es mayor que 220 miligramos? Suponga que la distribución de contenidos de sodio es normal.

10.27 Un estudio de la Universidad de Colorado en Boulder muestra que correr aumenta el porcentaje de la tasa metabólica de descanso (TMD) en mujeres ancianas. La TMD promedio de 30 ancianas corredoras fue 34.0% más alta que la TMD promedio de 30 ancianas sedentarias, en tanto que las desviaciones estándar reportadas fueron de 10.5 y 10.2%, respectivamente. ¿Hay un aumento significativo en la TMD de las corredoras con respecto a las sedentarias? Suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales. Utilice un valor P en sus conclusiones.

10.28 De acuerdo con el *Chemical Engineering* una propiedad importante de la fibra es su absorción de agua. Se encuentra que el porcentaje promedio de absorción de 25 piezas de fibra de algodón que se seleccionan al azar es 20 con una desviación estándar de 1.5. Una muestra aleatoria de 25 piezas de acetato dan un porcentaje promedio de 12 con una desviación estándar de 1.25. Hay evidencia sólida de que el porcentaje promedio de absorción de la población para la fibra de algodón es significativamente mayor que la media para el acetato. Suponga que el porcentaje de absorción se distribuye de forma aproximadamente normal, y que las varianzas de la población en el porcentaje de absorción para las dos fibras son las mismas. Utilice un nivel de significancia de 0.05.

10.29 La experiencia indica que el tiempo para que los estudiantes de último año de preparatoria terminen un examen estandarizado es una variable aleatoria normal con una media de 35 minutos. Si a una muestra aleatoria de 20 estudiantes de último año de preparatoria le toma un promedio de 33.1 minutos completar dicho examen con una desviación estándar de 4.3 minutos, con un nivel de significancia de 0.025, pruebe la hipótesis de que $\mu = 35$ minutos contra la alternativa de que $\mu < 35$ minutos.

10.30 Una muestra aleatoria de tamaño $n_1 = 25$, que se toma de una población normal con una desviación estándar $\sigma_1 = 5.2$, tiene una media $\bar{x}_1 = 81$. Una segunda muestra aleatoria de tamaño $n_2 = 36$, que se toma de una población normal diferente con una desviación estándar $\sigma_2 = 3.4$, tiene una media $\bar{x}_2 = 76$. Pruebe la hipótesis de que $\mu_1 = \mu_2$ contra la alternativa $\mu_1 \neq \mu_2$. Cite un valor P en su conclusión.

10.31 Un fabricante afirma que la resistencia a la tensión promedio del hilo A excede la resistencia a la tensión promedio del hilo B, en al menos 12 kilogramos. Para probar esta afirmación, se prueban 50 piezas de cada tipo de hilo bajo condiciones similares. El hilo tipo A tiene una resistencia a la tensión promedio de 86.7 kilogramos con una desviación estándar de 6.28 kilogramos; mientras que el hilo tipo B tiene una resistencia a la tensión promedio de 77.8 kilogramos con una desviación estándar de 5.61 kilogramos. Pruebe la afirmación del fabricante usando un nivel de significancia de 0.05.

10.32 El *Amstat News* (diciembre de 2004) lista los sueldos medios de profesores asociados de estadística en instituciones de investigación, en escuelas de humanidades y en otras instituciones en Estados Unidos. Suponga que una muestra de 200 profesores asociados de instituciones de investigación que tienen un sueldo promedio de \$70,750 anuales con una desviación estándar de \$6000. Suponga también una muestra de 200 profesores asociados de otros tipos de institución que tienen un sueldo promedio de \$65,200 con una desviación estándar de \$5000. Pruebe la hipótesis de que el sueldo medio de profesores asociados en instituciones de investigación es \$2000 mayor que los de los de otras instituciones. Utilice un nivel de significancia de 0.01.

10.33 Se lleva a cabo un estudio para saber si el aumento de la concentración de sustrato tiene un efecto apreciable sobre la velocidad de una reacción química. Con una concentración de sustrato de 1.5 moles por litro, la reacción se realizó 15 veces, con una velocidad promedio de 7.5 micromoles por 30 minutos y una desviación estándar de 1.5. Con una concentración de sustrato de 2.0 moles por litro, se realizan 12 reacciones, que dan una velocidad promedio de 8.8 micromoles por 30 minutos y una desviación estándar muestral de 1.2. ¿Hay alguna razón para creer que este incremento en la concentración de sustrato ocasiona un aumento en la velocidad media de más de 0.5 micromoles por 30 minutos? Utilice un nivel de significancia de 0.01 y suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales.

10.34 Se realiza un estudio para determinar si los temas de la materia en un curso de física se comprenden mejor cuando se emplea un laboratorio en parte del curso. Se seleccionan estudiantes al azar para que participen, ya sea en un curso de tres semestres-hora sin laboratorio o en un curso de cuatro semestres-hora con laboratorio. En la sección con laboratorio 11 estudiantes tuvieron una calificación promedio de 85 con una desviación estándar de 4.7; mientras que en la sección sin laboratorio 17 estudiantes tuvieron una calificación promedio de 79 con una desviación estándar de 6.1. ¿Diría usted que el curso con laboratorio aumenta la calificación promedio hasta en 8 puntos? Utilice un valor P en su conclusión y suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales.

10.35 Para indagar si un nuevo suero frena el desarrollo de la leucemia, se seleccionan 9 ratones, todos con una etapa avanzada de la enfermedad. Cinco ratones reciben el tratamiento y cuatro no. Los tiempos de supervivencia, en años, a partir del momento en que comienza el experimento son los siguientes:

Con tratamiento	2.1	5.3	1.4	4.6	0.9
Sin tratamiento	1.9	0.5	2.8	3.1	

¿Se puede decir en el nivel de significancia de 0.05 que el suero es efectivo? Suponga que las dos distribuciones se distribuyen de forma normal con varianzas iguales.

10.36 Una compañía grande armadora de automóviles trata de decidir si compra llantas de la marca A o de la B para sus modelos nuevos. Para ayudar a tomar una decisión, se realiza un experimento en donde se usan 12 llantas de cada marca. Las llantas se utilizan hasta que se acaban. Los resultados son

$$\begin{aligned} \text{Marca A:}\ & \bar{x}_1 = 37{,}900 \text{ kilómetros}, \\ & s_1 = 5{,}100 \text{ kilómetros}, \\ \text{Marca B:}\ & \bar{x}_1 = 39{,}800 \text{ kilómetros}, \\ & s_2 = 5{,}900 \text{ kilómetros}. \end{aligned}$$

Pruebe la hipótesis de que no hay diferencia en el desgaste promedio de las 2 marcas de llantas. Suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales. Use un valor P.

10.37 En el ejercicio 9.42 de la página 298, pruebe la hipótesis de que los camiones compactos Volkswagen, en promedio, exceden a los camiones compactos Toyota, equipados de forma similar, en cuatro kilómetros por litro. Utilice un nivel de significancia de 0.10.

10.38 Un investigador de la UCLA afirma que la vida promedio de un ratón se puede prolongar hasta 8 meses más cuando las calorías en su dieta se reducen en aproximadamente 40% desde el momento en que se destetan. Las dietas restringidas se enriquecen a niveles normales con vitaminas y proteína. Suponga que se alimenta a una muestra aleatoria de 10 ratones con una dieta normal y tiene una vida promedio de 32.1 meses con una desviación estándar de 3.2 meses; mientras que una muestra aleatoria de 15 ratones se alimenta con la dieta restringida y viven un promedio de 37.6 meses con una desviación estándar de 2.8 meses. Con un nivel de significancia de 0.05 pruebe la hipótesis de que la vida promedio de los ratones con esta dieta restringida aumenta 8 meses contra la alternativa de que el aumento es menor que 8 meses. Suponga que las distribuciones de las vidas con las dietas regular y restringida son aproximadamente normales con varianzas iguales.

10.39 Los siguientes datos representan los tiempos de duración de películas producidas por 2 compañías cinematográficas:

Compañía	Tiempo (minutos)						
1	102	86	98	109	92		
2	81	165	97	134	92	87	114

Pruebe la hipótesis de que el tiempo de duración promedio de las películas producidas por la compañía 2 excede el tiempo promedio de duración de las que produce la compañía 1 en 10 minutos, contra la alternativa unilateral de que la diferencia es de menos de 10 minutos. Utilice un nivel de significancia de 0.1 y suponga que las distribuciones de los tiempos son aproximadamente normales con varianzas iguales.

10.40 En un estudio realizado en el Instituto Politécnico y Universidad Estatal de Virginia, se compararon los niveles de ácido ascórbico en plasma en mujeres embarazadas fumadoras contra las no fumadoras. Para el estudio se seleccionaron 32 mujeres en los últimos tres meses de embarazo, libres de padecimientos importantes y con edades de entre 15 y 32 años. Antes de tomar las muestras de 20 ml de sangre, a las participantes se les solicitó ir en ayunas, no consumir sus complementos vitamínicos y evitar comidas con alto contenido de ácido ascórbico. De las muestras de sangre se determinaron los siguientes valores, en miligramos por 100 mililitros, de ácido ascórbico en plasma de cada mujer:

Valores de ácido ascórbico en plasma

No fumadoras		Fumadoras
0.97	1.16	0.48
0.72	0.86	0.71
1.00	0.85	0.98
0.81	0.58	0.68
0.62	0.57	1.18
1.32	0.64	1.36
1.24	0.98	0.78
0.99	1.09	1.64
0.90	0.92	
0.74	0.78	
0.88	1.24	
0.94	1.18	

¿Existe suficiente evidencia para concluir que hay una diferencia entre los niveles de ácido ascórbico en plasma entre fumadoras y no fumadoras? Suponga que los dos conjuntos de datos provienen de poblaciones normales con varianzas diferentes. Utilice un valor P.

10.41 El Departamento de Zoología del Instituto Politécnico y Universidad Estatal de Virginia llevó a cabo un estudio, para determinar si hay una diferencia significativa en la densidad de organismos en dos estaciones diferentes ubicadas en Cedar Run, un río secundario que se localiza en la cuenca del río Roanoke. El drenaje de una planta de tratamiento de aguas negras y el sobreflujo del estanque de sedimentación de la Federal Mogul Corporation entran al flujo cerca del nacimiento del río. Los siguientes datos dan las medidas de densidad, en número de organismos por metro cuadrado, en las dos diferentes estaciones colectoras:

Número de organismos por metro cuadrado

Estación 1		Estación 2	
5030	4980	2800	2810
13,700	11,910	4670	1330
10,730	8130	6890	3320
11,400	26,850	7720	1230
860	17,660	7030	2130
2200	22,800	7330	2190
4250	1130		
15,040	1690		

¿Con un nivel de significancia de 0.05 podemos concluir que son iguales las densidades promedio en las dos estaciones? Suponga que las observaciones provienen de poblaciones normales con varianzas diferentes.

10.42 Cinco muestras de una sustancia ferrosa se usan para determinar si hay una diferencia entre un análisis

químico de laboratorio y un análisis de fluorescencia de rayos X del contenido de hierro. Cada muestra se divide en dos submuestras y se aplican los dos tipos de análisis. A continuación se presentan los datos codificados que muestran los análisis de contenido de hierro:

	Muestra				
Análisis	**1**	**2**	**3**	**4**	**5**
Rayos X	2.0	2.0	2.3	2.1	2.4
Química	2.2	1.9	2.5	2.3	2.4

Suponiendo que las poblaciones son normales, pruebe con un nivel de significancia de 0.05 si los dos métodos de análisis dan, en promedio, el mismo resultado.

10.43 El administrador de una compañía de taxis trata de decidir si el uso de llantas radiales en lugar de llantas regulares cinturadas mejora la economía de combustible. Se equipan 12 automóviles con llantas radiales y se manejan durante un recorrido de prueba preestablecido. Sin cambiar a los conductores, los mismos automóviles se equipan con llantas regulares cinturadas y se manejan otra vez en el recorrido de prueba. El consumo de gasolina, en kilómetros por litro, se registró de la siguiente manera:

	Kilómetros por litro	
Automóvil	**Llantas radiales**	**Llantas cinturadas**
1	4.2	4.1
2	4.7	4.9
3	6.6	6.2
4	7.0	6.9
5	6.7	6.8
6	4.5	4.4
7	5.7	5.7
8	6.0	5.8
9	7.4	6.9
10	4.9	4.7
11	6.1	6.0
12	5.2	4.9

¿Podemos concluir que los automóviles equipados con llantas radiales dan una economía de combustible mejor que aquellos equipados con llantas cinturadas? Suponga que las poblaciones se distribuyen normalmente. Utilice un valor P en su conclusión.

10.44 En el ejercicio 9.88 de la página 315, utilice la distribución t para probar la hipótesis de que la dieta reduce el peso de un individuo en 4.5 kilogramos, en promedio, contra la hipótesis alternativa de que la diferencia media en peso es menor que 4.5 kilogramos. Utilice un valor P.

10.45 De acuerdo con informes publicados, el ejercicio bajo condiciones de fatiga altera los mecanismos que determinan el desempeño. Se realizó un experimento donde se usaron 15 estudiantes universitarios hombres, entrenados para realizar un movimiento horizontal continuo del brazo, de derecha a izquierda, desde un microinterruptor hasta una barrera, golpeando sobre la barrera en coincidencia con la llegada de una manecilla del reloj a la posición de las 6 en punto. Se registra el valor absoluto de la diferencia entre el tiempo, en milisegundos, que toma golpear sobre la barrera y el tiempo para que la manecilla alcance la posición de las 6 en punto (500 mseg). Cada participante ejecuta la tarea cinco veces en condiciones sin fatiga y con fatiga, y se registraron las sumas de las diferencias absolutas para las cinco ejecuciones como sigue:

	Diferencias absolutas de tiempo	
Sujeto	**Sin fatiga**	**Con fatiga**
1	158	91
2	92	59
3	65	215
4	98	226
5	33	223
6	89	91
7	148	92
8	58	177
9	142	134
10	117	116
11	74	153
12	66	219
13	109	143
14	57	164
15	85	100

Un aumento en las diferencias medias absolutas de tiempo cuando la tarea se ejecuta bajo condiciones de fatiga apoyaría la afirmación de que el ejercicio, en condiciones de fatiga, altera el mecanismo que determina el desempeño. Suponiendo de que las poblaciones se distribuyen normalmente, pruebe tal afirmación.

10.46 En un estudio realizado por el Departamento de Nutrición Humana y Alimentos del Instituto Politécnico y Universidad Estatal de Virginia, se registraron los siguientes datos acerca de la comparación de residuos de ácido sórbico, en partes por millón, en jamón inmediatamente después de sumergirlo en una solución de ácido y después de 60 días de almacenamiento:

	Residuos de ácido sórbico en jamón	
Rebanada	**Antes del almacenamiento**	**Después del almacenamiento**
1	224	116
2	270	96
3	400	239
4	444	329
5	590	437
6	660	597
7	1400	689
8	680	576

Si se supone que las poblaciones se distribuyen normalmente, ¿hay suficiente evidencia, al nivel de significancia de 0.05, para decir que la duración del almacenamiento influye en las concentraciones residuales de ácido sórbico?

10.47 ¿Qué tan grande se requiere que sea la muestra del ejercicio 10.20, si la potencia de nuestra prueba debe ser 0.90 cuando la media real es 5.20? Suponga que $\sigma = 0.24$.

10.48 Si la distribución del tiempo de vida en el ejercicio 10.21 es aproximadamente normal, ¿qué tan grande se requiere que sea una muestra, para que la probabilidad de cometer un error tipo II sea 0.1 cuando la media real es 35.9 meses? Suponga que $\sigma = 5.8$ meses.

10.49 ¿Qué tan grande se requiere que sea la muestra del ejercicio 10.22, si la potencia de nuestra prueba debe ser 0.95 cuando la estatura promedio real difiere de 162.5 en 3.1 centímetros?

10.50 ¿Qué tan grandes deberían ser las muestras del ejercicio 10.31, si la potencia de nuestra prueba debe ser 0.95 cuando la diferencia real entre los tipos de hilo A y B es 8 kilogramos?

10.51 ¿Qué tan grande se requiere que sea la muestra del ejercicio 10.24 si la potencia de nuestra prueba será 0.8 cuando el tiempo medio real de meditación exceda el valor hipotético en 1.2 σ? Utilice $\alpha = 0.05$.

10.52 Se considera una prueba t de nivel $\alpha = 0.05$ para probar

$$H_0\text{: } \mu = 14,$$
$$H_1\text{: } \mu \neq 14.$$

¿Qué tamaño de la muestra se necesita para que la probabilidad sea 0.1 de no rechazar de manera errónea H_0, cuando la media poblacional real difiera de 14 en 0.5? A partir de una muestra preliminar estimamos que σ es 1.25.

10.53 Se llevó a cabo un estudio en el Departamento de Veterinaria del Instituto Politécnico y Universidad Estatal de Virginia, para determinar si la "resistencia" de una herida de incisión quirúrgica resulta afectada por la temperatura del bisturí. Se utilizaron 8 perros en el experimento. La incisión se realizó en el abdomen de los animales. Se aplicaron una incisión "caliente" y una "fría" a cada perro, y se midió la resistencia. Los datos que resultaron aparecen abajo.

a) Escriba una hipótesis apropiada para determinar si hay una diferencia significativa en la resistencia entre las incisiones caliente y fría.

b) Pruebe la hipótesis mediante el uso de una prueba t pareada. Utilice un valor P en su conclusión.

Perro	Bisturí	Resistencia
1	Caliente	5120
1	Frío	8200
2	Caliente	10000
2	Frío	8600
3	Caliente	10000
3	Frío	9200
4	Caliente	10000
4	Frío	6200
5	Caliente	10000
5	Frío	10000
6	Caliente	7900
6	Frío	5200
7	Caliente	510
7	Frío	885
8	Caliente	1020
8	Frío	460

10.54 Se utilizaron 9 sujetos en un experimento para determinar si una atmósfera que implica la exposición a monóxido de carbono tiene un impacto sobre la capacidad de respiración. Los datos fueron recolectados por el personal del Departamento de Salud y Educación Física del Instituto Politécnico y Universidad Estatal de Virginia. Los datos se analizaron en el Centro de Consulta Estadística en Hokie Land. Los sujetos se colocaron en cámaras de respiración, una de las cuales contenía una alta concentración de CO. Se realizaron varias mediciones de respiración para cada sujeto en cada cámara. Los sujetos se colocaron en las cámaras de respiración en una secuencia aleatoria. Los siguientes datos dan la frecuencia respiratoria en número de respiraciones por minuto. Realice una prueba unilateral de la hipótesis de que la frecuencia respiratoria media es la misma para los dos ambientes. Utilice $\alpha = 0.05$. Suponga que la frecuencia respiratoria es aproximadamente normal

Sujeto	Con CO	Sin CO
1	30	30
2	45	40
3	26	25
4	25	23
5	34	30
6	51	49
7	46	41
8	32	35
9	30	28

10.11 Una muestra: Prueba sobre una sola proporción

Las pruebas de hipótesis que se relacionan con proporciones se requieren en muchas áreas. Seguramente el político se interesará en conocer qué fracción de votantes lo favorecerá en la siguiente elección. Todas las empresas manufactureras se preocupan

por la proporción de artículos defectuosos cuando se realiza un embarque. El jugador depende de un conocimiento de la proporción de resultados que cree favorables.

Consideraremos el problema de probar la hipótesis de que la proporción de éxitos en un experimento binomial es igual a algún valor específico. Es decir, probaremos la hipótesis nula H_0 que $p = p_0$, donde p es el parámetro de la distribución binomial. La hipótesis alternativa puede ser una de las alternativas unilaterales o bilaterales usuales:

$$p < p_0, \quad p > p_0, \quad \text{o} \quad p \neq p_0.$$

La variable aleatoria adecuada sobre la que basamos nuestro criterio de decisión es la variable aleatoria binomial X; aunque también podríamos usar sólo el estadístico $\hat{p} = X/n$. Los valores de X que están lejos de la media $\mu = np_0$ conducirán al rechazo de la hipótesis nula. Como X es una variable binomial discreta, es poco probable que se pueda establecer una región crítica, cuyo tamaño sea *exactamente* igual a un valor prescrito de α. Por tal razón es preferible, al tratar con muestras pequeñas, basar nuestra decisión en valores P. *P*ara probar la hipótesis

$$H_0\text{: } p = p_0,$$
$$H_1\text{: } p < p_0,$$

utilizamos la distribución binomial para calcular el valor P

$$P = P(X \leq \text{x cuando } p = p_0).$$

El valor x es el número de éxitos en nuestra muestra de tamaño n. Si este valor P es menor que o igual a α, nuestra prueba es significativa en el nivel α y rechazamos H_0 a favor de H_1. De manera similar, para probar la hipótesis

$$H_0\text{: } p = p_0,$$
$$H_1\text{: } p > p_0,$$

en el nivel de significancia α, calculamos

$$P = P(X > x \text{ cuando } p = p_0)$$

y rechazamos H_0 a favor de H_1 si este valor P es menor que o igual a α. Finalmente, para probar la hipótesis

$$H_0\text{: } p = p_0,$$
$$H_1\text{: } p \neq p_0,$$

al nivel de significancia α, calculamos

$$P = 2P(X \leq x \text{ cuando } p = p_0) \qquad \text{si } x < np_0, \quad \text{o}$$
$$P = 2P(X \geq x \text{ cuando } p = p_0) \qquad \text{si } x > np_0$$

y se rechaza H_0 a favor de H_1, si el valor P calculado es menor que o igual a α.

Los pasos para probar una hipótesis nula acerca de una proporción contra varias alternativas usando las probabilidades binomiales de la tabla A.1 son los siguientes:

Prueba de una proporción: muestras pequeñas

1. H_0**:** $p = p_0$.
2. Una de las alternativas H_1: $p < p_0$, $p > p_0$, o $p \neq p_0$.
3. Elija un nivel de significancia igual a α.
4. Estadístico de prueba: variable binomial X con $p = \ p_0$.
5. Cálculos: Encuentre x, el número de éxitos, y calcule el valor P adecuado.
6. Decisión: Obtenga las conclusiones apropiadas basadas en el valor P.

Ejemplo 10.10: Un constructor afirma que se instalan bombas de calor en 70% de todas las casas que se construyen actualmente en la ciudad de Richmond, Virginia. ¿Estaría de acuerdo con esta afirmación, si una encuesta aleatoria de casas nuevas en esta ciudad demuestra que 8 de 15 tienen instaladas bombas de calor? Utilice un nivel de significancia de 0.10.

Solución:

1. H_0: $p = 0.7$.

2. H_1: $p \neq 0.7$.

3. $\alpha = 0.10$.

4. Estadístico de prueba: Variable binomial X con $p = 0.7$ y $n = 15$.

5. Cálculos: $x = 8$ y $np_0 = (15)(0.7) = 10.5$. Por lo tanto, de la tabla A.1, el valor P calculado es

$$P = 2P(X \leq 8 \text{ cuando } p = 0.7) = 2\sum_{x=0}^{8} b(x; 15, 0.7) = 0.2622 > 0.10.$$

6. Decisión: No rechace H_0. Concluya que no hay razón suficiente para dudar de la afirmación del constructor.

En la sección 5.3, vimos que las probabilidades binomiales se obtienen de la fórmula binomial real o de la tabla A.1 cuando n es pequeña. Para n grande, se requieren procedimientos de aproximación. Cuando el valor hipotético p_0 está muy cercano a 0 o a 1, se puede utilizar la distribución de Poisson con parámetro $\mu = np_0$. Sin embargo, la aproximación de la curva normal, con parámetros $\mu = np_0$ y $\sigma^2 = np_0q_0$, por lo general, se prefiere para n grande y es muy precisa, en tanto que p_0 no esté extremadamente cerca de 0 o de 1. Si utilizamos la aproximación normal, el **valor** z **para probar** $p = p_0$ está dado por

$$z = \frac{x - np_0}{\sqrt{np_0q_0}} = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}},$$

que es un valor de la variable normal estándar Z. De aquí que, para una prueba de dos colas al nivel de significancia α, la región crítica es $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$. Para la alternativa unilateral $p < p_0$, la región crítica es $z < -z_\alpha$, y para la alternativa $p > p_0$, la región crítica es $z > z_\alpha$.

Ejemplo 10.11: Un medicamento que se prescribe comúnmente para aliviar la tensión nerviosa se considera que es efectivo en tan sólo 60%. Resultados experimentales con un nuevo fármaco que se suministra a una muestra aleatoria de 100 adultos que padecen de tensión nerviosa demuestran que 70 tuvieron alivio. ¿Ésta es evidencia suficiente para concluir que el nuevo medicamento es superior a la que se prescribe actualmente? Utilice un nivel de significancia de 0.05.

Solución: **1.** H_0: $p = 0.6$.

2. H_1: $p > 0.6$.

3. $\alpha = 0.05$.

4. Región crítica: $z > 1.645$.

5. Cálculos: $x = 70$, $n = 100$, $\hat{p} = 70/100 = 0.7$, y

$$z = \frac{0.7 - 0.6}{\sqrt{(0.6)(0.4)/100}} = 2.04, \quad P = P(Z > 2.04) < 0.0207.$$

6. Decisión: Rechace H_0 y concluya que el nuevo fármaco es superior.

10.12 Dos muestras: Pruebas sobre dos proporciones

A menudo hay situaciones donde deseamos probar la hipótesis de que dos proporciones son iguales. Por ejemplo, podemos tratar de mostrar evidencia de que la proporción de doctores que son pediatras en una entidad es igual a la proporción de pediatras en otra entidad. Quizás un individuo decida dejar de fumar sólo si se convence de que la proporción de fumadores con cáncer pulmonar excede la proporción de no fumadores con ese tipo de cáncer.

En general, deseamos probar la hipótesis nula de que dos proporciones, o parámetros binomiales, son iguales. Es decir, probamos $p_1 = p_2$ contra una de las alternativas $p_1 < p_2$, $p_1 > p_2$, o $p_1 \neq p_2$. Desde luego, esto es equivalente a probar la hipótesis nula de que $p_1 - p_2 = 0$ contra una de las alternativas $p_1 - p_2 < 0$, $p_1 - p_2 > 0$ o $p_1 - \mathrm{p}_2 \neq 0$. El estadístico sobre el que basamos nuestra decisión es la variable aleatoria $\hat{P}_1 - \hat{P}_2$. Se seleccionan al azar muestras independientes de tamaño n_1 y n_2 de dos poblaciones binomiales y se calcula la proporción de éxitos $\hat{P}_1$ y $\hat{P}_2$ para las dos muestras.

En nuestra construcción de intervalos de confianza para p_1 y p_2 señalamos, para n_1 y n_2 suficientemente grandes, que el estimador puntual $\hat{P}_1$ menos $\hat{P}_2$ estaba distribuido de forma aproximadamente normal con media

$$\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$$

y varianza

$$\sigma^2_{\hat{P}_1 - \hat{P}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.$$

Por lo tanto, nuestra(s) región(es) crítica(s) se puede(n) establecer usando la variable normal estándar

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}}.$$

Cuando H_0, es verdadera, podemos sustituir $p_1 = p_2 = p$ y $q_1 = q_2 = q$ (donde p y q son los valores comunes) en la fórmula anterior para Z y obtener la forma

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{pq(1/n_1 + 1/n_2)}}.$$

Para calcular un valor de Z, no obstante, debemos estimar los parámetros p y q que aparecen en el radical. Al combinar los datos de ambas muestras, la **estimación combinada de la proporción p** es

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2},$$

donde x_1 y x_2 son el número de éxitos en cada una de las dos muestras. Al sustituir $\hat{p}$ por p y $\hat{q} = 1 - \hat{p}$ por q, el **valor z para probar $p_1 = p_2$** se determina a partir de la fórmula

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}.$$

Las regiones críticas para las hipótesis alternativas adecuadas se establecen como antes, utilizando puntos críticos de la curva normal estándar. De aquí, para la alternativa $p_1 \neq p_2$ con un nivel de significancia α, la región crítica es $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$. Para una prueba donde la alternativa sea $p_1 < p_2$, la región crítica será $z < -z_\alpha$; y cuando la alternativa sea $p_1 > p_2$, la región crítica será $z > z_\alpha$.

Ejemplo 10.12: Se tomará el voto entre los residentes de una ciudad y el condado circundante, para determinar si se debe construir la planta química que se propone. El lugar de construcción está dentro de los límites de la ciudad y, por esa razón, muchos votantes del condado consideran que la propuesta pasará debido a la gran proporción de votantes que favorecen la construcción. Para determinar si hay una diferencia significativa en la proporción de votantes de la ciudad y votantes del condado que favorecen la propuesta, se realiza una encuesta. Si 120 de 200 votantes de la ciudad favorecen la propuesta y 240 de 500 residentes del condado también lo hacen, ¿estaría usted de acuerdo en que la proporción de votantes de la ciudad que favorecen la propuesta es mayor que la proporción de votantes del condado? Utilice un nivel de significancia de 0.025.

Solución: Sean p_1 y p_2 las proporciones reales de votantes en la ciudad y el condado, respectivamente, que favorecen la propuesta.

1. H_0: $p_1 = p_2$.

2. H_1: $p_1 > p_2$.

3. $\alpha = 0.05$.

4. Región crítica: $z > 1.645$.

5. Cálculos:

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{120}{200} = 0.60, \quad \hat{p}_2 = \frac{x_2}{n_2} = \frac{240}{500} = 0.48, \quad \text{y}$$
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{120 + 240}{200 + 500} = 0.51.$$

Por lo tanto,

$$z = \frac{0.60 - 0.48}{\sqrt{(0.51)(0.49)(1/200 + 1/500)}} = 2.9,$$
$$P = P(Z > 2.9) = 0.0019.$$

6. Decisión: Rechace H_0 y esté de acuerdo en que la proporción de votantes de la ciudad a favor de la propuesta es mayor que la proporción de votantes del condado.

Ejercicios

10.55 Un experto en marketing de una compañía fabricante de pasta considera que 40% de los amantes de la pasta prefieren la lasagna. Si 9 de 20 amantes de la pasta eligen lasagna sobre otras pastas, ¿qué se puede concluir acerca de la afirmación del experto? Utilice un nivel de significancia de 0.05.

10.56 Suponga que, en el pasado, 40% de todos los adultos favorecían la pena capital. ¿Tenemos razón para creer que la proporción de adultos que actualmente favorecen la pena capital ha aumentado si, en una muestra aleatoria de 15 adultos, 8 están a favor de la pena capital? Utilice un nivel de significancia de 0.05.

10.57 Se lanza 20 veces una moneda, con un resultado de 5 caras. ¿Ésta es suficiente evidencia para rechazar la hipótesis de que la moneda está balanceada, a favor de la alternativa de que las caras ocurren menos de 50% de las veces? Cite un valor *P*.

10.58 Se cree que al menos 60% de los residentes de cierta área favorecen un demanda de anexión de una ciudad vecina. ¿Qué conclusión extraería, si sólo 110 en una muestra de 200 votantes están a favor de la demanda? Utilice un nivel de significancia de 0.05.

10.59 Una compañía petrolera afirma que un quinto de las casas en cierta ciudad se calientan con petróleo. ¿Tenemos razón para creer que menos de 1/5 se calientan con petróleo si, en una muestra aleatoria de 1000 casas en esta ciudad, se encuentra que 136 se calientan con petróleo? Utilice un valor P en su conclusión.

10.60 En cierta universidad se estima que a lo más 25% de los estudiantes van en bicicleta a la escuela. ¿Ésta parece ser una estimación válida si, en una muestra aleatoria de 90 estudiantes universitarios, se encuentra que 28 van en bicicleta a la escuela? Utilice un nivel de significancia de 0.05.

10.61 Se considera un nuevo dispositivo de radar para cierto sistema de misiles de defensa. El sistema se verifica experimentando con aeronaves reales, en las cuales se simula una situación de *muerte* o de *sin muerte*. Si en 300 pruebas ocurren 250 muertes, acepte o rechace, con un nivel de significancia de 0.04, la afirmación de que la probabilidad de una muerte con el sistema nuevo no excede la probabilidad de 0.8 del sistema existente.

10.62 En un experimento de laboratorio controlado, científicos de la Universidad de Minnesota descubrieron que 25% de cierta camada de ratas sujetas a una dieta de 20% de grano de café, y luego forzadas a consumir un poderoso químico causante de cáncer, desarrollaron tumores cancerosos. ¿Tendríamos razones para creer que la proporción de ratas que desarrollan tumores cuando se sujeta a esta dieta aumenta, si el experimento se repite y 16 de 48 ratas desarrollan tumores? Utilice un nivel de significancia de 0.05.

10.63 En un estudio para estimar la proporción de residentes de cierta ciudad y sus suburbios que están a favor de la construcción de una planta de energía nuclear, se encuentra que 63 de 100 residentes urbanos favorecen la construcción, mientras que sólo 59 de 125 residentes suburbanos la favorecen. ¿Hay una diferencia significativa entre la proporción de residentes urbanos y suburbanos que favorecen la construcción de la planta nuclear? Utilice un valor *P*.

10.64 En un estudio sobre la fertilidad de mujeres casadas conducido por Martin O'Connell y Carolyn C. Rogers para la Oficina de Censos en 1979, se seleccionaron al azar dos grupos de esposas con edades de 25 a 29 años y sin hijos, y a cada una se le preguntó si a final de cuentas planeaba tener un hijo. Se seleccionó un grupo entre las mujeres con menos de dos años de casadas y otro entre las que tenían cinco años de casadas. Suponga que 240 de 300 con menos de dos años de casadas planean tener un hijo algún día, comparadas con 288 de las 400 con cinco años de casadas. ¿Podemos concluir que la proporción de mujeres con menos de dos años de casadas que planean tener hijos es significativamente mayor que la proporción con cinco años de casadas? Utilice un valor *P*.

10.65 Una comunidad urbana quiere demostrar que la incidencia de cáncer de seno es mayor en ella que en un área rural vecina. (Se encontró que los niveles de PCB son más altos en el suelo de la comunidad urbana.) Si se encuentra que 20 de 200 mujeres adultas en la comunidad urbana tienen cáncer de seno y 10 de 150 mujeres adultas en la comunidad rural tienen cáncer de seno, ¿podríamos concluir con un nivel de significancia de 0.05 que este tipo de cáncer prevalece más en la comunidad urbana?

10.66 En un invierno con epidemia de gripe, una compañía farmacéutica bien conocida estudió a 2000 bebés, para determinar si el nuevo medicamento de la compañía era eficaz después de dos días. Entre 120 bebés que tenían gripe y se les suministró el medicamento, 29 se curaron dentro de dos días. Entre 280 bebés que tenían gripe pero que no recibieron el fármaco, 56 se curaron dentro de dos días. ¿Hay alguna indicación significativa que apoye la afirmación de la compañía de la efectividad del medicamento?

10.13 Pruebas de una y dos muestras referentes a varianzas

En esta sección nos referimos a la prueba de hipótesis relacionada con varianzas o desviaciones estándar poblacionales. Las pruebas de una y dos muestras sobre varianzas en realidad no son difíciles de motivar. Los ingenieros y los científicos constantemente se enfrentan a estudios donde se les pide demostrar que las mediciones que tienen que ver con productos o procesos caen dentro de las especificaciones que fijan los consumidores. Las especificaciones a menudo se cumplen si la varianza del proceso es suficientemente pequeña. La atención también se concentra en experimentos comparativos entre métodos o procesos, donde la reproductibilidad o variabilidad inherentes se deben comparar de manera formal. Además, con frecuencia se aplica una prueba que compara dos varianzas antes de llevar a cabo una prueba t sobre dos medias. El objetivo es determinar si se viola la suposición de varianzas iguales.

Primero consideremos el problema de probar la hipótesis nula H_0 de que la varianza poblacional σ^2 es igual a un valor específico σ_0^2, contra una de las alternativas comunes $\sigma^2 < \sigma_0^2$, $\sigma^2 > \sigma_0^2$ o $\sigma^2 \neq \sigma_0^2$. El estadístico apropiado sobre el que basamos nuestra decisión es el mismo estadístico chi cuadrado del teorema 8.4 que se utiliza en el capítulo 9 para construir un intervalo de confianza para σ^2. Por lo tanto, si suponemos que la distribución de la población que se muestrea es normal, el valor de chi cuadrada para probar $\sigma^2 = \sigma_0^2$. está dado por

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

donde n es el tamaño de la muestra, s^2 es la varianza muestral y σ_0^2 es el valor de σ^2 dado por la hipótesis nula. Si H_0 es verdadera, χ^2 es un valor de la distribución chi cuadrada con $v = n - 1$ grados de libertad. De aquí que, para una prueba de dos colas en el nivel de significancia α, la región crítica es $\chi^2 < \chi^2_{1-\alpha/2}$ o $\chi^2 > \chi^2_{\alpha/2}$. Para la alternativa unilateral $\sigma^2 < \sigma_0^2$, la región crítica es $\chi^2 < \chi^2_{1-\alpha}$; y para la alternativa unilateral $\sigma^2 > \sigma_0^2$, la región crítica es $\chi^2 > \chi^2_{\alpha}$.

Robustez de la prueba χ^2 para la suposición de normalidad

El lector puede percibir que varias pruebas dependen, al menos en teoría, de la suposición de normalidad. En general, muchos procedimientos en estadística aplicada tienen fundamentos teóricos que dependen de la distribución normal. Estos procedimientos varían en el grado de su dependencia de la suposición de la normalidad. Un procedimiento que es razonablemente insensible a la suposición se denomina **procedimiento robusto** (es decir, robusto para la normalidad). La prueba χ^2 sobre una sola varianza no es robusta en absoluto hacia la normalidad (es decir, el éxito práctico del procedimiento depende de la normalidad). Como resultado, el valor P calculado puede ser apreciablemente diferente del valor P real si la población muestreada no es normal. En realidad, resulta bastante factible que un valor P estadísticamente significativo quizá no sea una verdadera señal de H_1: $\sigma \neq \sigma_0$ sino, más bien, que un valor significativo puede ser un resultado de la transgresión de las suposiciones de normalidad. Por lo tanto, el analista se debería aproximar con precaución al uso de esta prueba χ^2 específica.

Ejemplo 10.13: Un fabricante de baterías para automóvil afirma que la duración de sus baterías se distribuye de forma aproximadamente normal con una desviación estándar igual a

0.9 años. Si una muestra aleatoria de 10 de tales baterías tiene una desviación estándar de 1.2 años, ¿considera que $\sigma > 0.9$ años? Utilice un nivel de significancia de 0.05.

Solución: **1.** H_0: $\sigma^2 = 0.81$.

2. H_1: $\sigma^2 > 0.81$.

3. $\alpha = 0.05$.

4. Región crítica: De la figura 10.19 vemos que se rechaza la hipótesis nula cuando $\chi^2 > 16.919$, donde $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ con $v = 9$ grados de libertad.

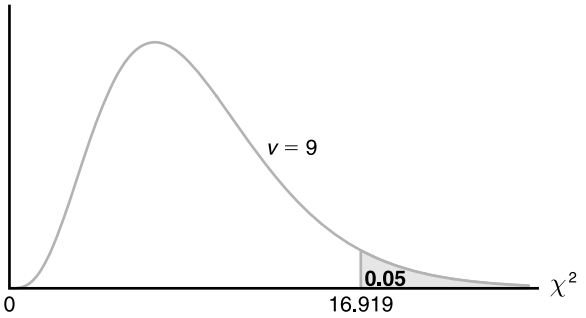


Figura 10.19: Región crítica para la hipótesis alternativa $\sigma > 0.9$.

5. Cálculos: $s^2 = 1.44$, $n = 10$ y

$$\chi^2 = \frac{(9)(1.44)}{0.81} = 16.0, \qquad P \approx 0.07.$$

6. Decisión: El estadístico χ^2 no es significativo en el nivel 0.05. Sin embargo, con base en el valor P de 0.07, hay alguna evidencia de que $\sigma > 0.9$.

Consideremos ahora el problema de probar la igualdad de las varianzas σ_1^2 y σ_2^2 de dos poblaciones. Esto es, probaremos la hipótesis nula H_0 de que $\sigma_1^2 = \sigma_2^2$ contra una de las alternativas usuales

$$\sigma_1^2 < \sigma_2^2, \quad \sigma_1^2 > \sigma_2^2, \quad \text{o} \quad \sigma_1^2 \neq \sigma_2^2.$$

Para muestras aleatorias independientes de tamaño n_1 y n_2, respectivamente, de las dos poblaciones, el valor f **para probar** $\sigma_1^2 = \sigma_2^2$ es la razón

$$f = \frac{s_1^2}{s_2^2},$$

donde s_1^2 y s_2^2 son las varianzas calculadas de las dos muestras. Si las dos poblaciones se distribuyen de forma aproximadamente normal y la hipótesis nula es verdadera, de acuerdo con el teorema 8.8 la razón $f = s_1^2/s_2^2$ es un valor de la distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Por lo tanto, las regiones críticas de tamaño α que corresponden a las alternativas unilaterales $\sigma_1^2 < \sigma_2^2$ y $\sigma_1^2 > \sigma_2^2$ son, respectivamente, $f < f_{1-\alpha}(v_1, v_2)$ y $f > f_\alpha(v_1, v_2)$. Para la alternativa bilateral $\sigma_1^2 \neq \sigma_2^2$, la región crítica es $f < f_{1-\alpha/2}(v_1, v_2)$ o $f > f_{\alpha/2}(v_1, v_2)$.

Ejemplo 10.14: Al probar la diferencia en el desgaste abrasivo de los dos materiales del ejemplo 10.6, supusimos que eran iguales las dos varianzas poblacionales desconocidas. ¿Se justifica tal suposición? Utilice un nivel de significancia de 0.10.

Solución: Sean σ_1^2 y σ_2^2 las varianzas poblacionales para el desgaste abrasivo del material 1 y del material 2, respectivamente.

1. H_0: $\sigma_1^2 = \sigma_2^2$.

2. H_1: $\sigma_1^2 \neq \sigma_2^2$.

3. $\alpha = 0.10$.

4. Región crítica: De la figura 10.20, sabemos que $f_{0.05}(11, 9) = 3.11$ y, usando el teorema 8.7,

$$f_{0.95}(11, 9) = \frac{1}{f_{0.05}(9, 11)} = 0.34.$$

Por lo tanto, se rechaza la hipótesis nula cuando $f < 0.34$ o $f > 3.11$, donde $f = s_1^2/s_2^2$ con $v_1 = 11$ y $v_2 = 9$ grados de libertad.

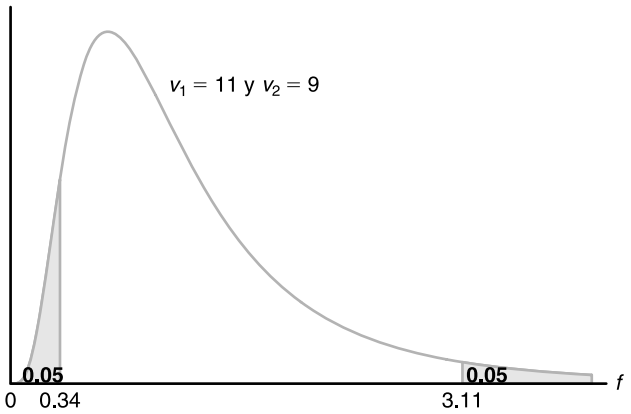


Figura 10.20: Región crítica para la hipótesis alternativa $\sigma_1^2 \neq \sigma_2^2$.

5. Cálculos: $s_1^2 = 16$, $s_2^2 = 25$ y, por ende, $f = \frac{16}{25} = 0.64$.

6. Decisión: no rechace H_0. Concluya que no hay suficiente evidencia de que las varianzas difieran.

Prueba *F* para probar varianzas con el SAS

La figura 10.18 de la página 357 muestra una prueba t de dos muestras donde se comparan dos medias, como ejercicio, con los datos de los retoños. La gráfica de caja y extensión en la figura 10.17 de la página 356 sugiere que las varianzas no son homogéneas y, por ello, el estadístico t' y su valor P correspondiente son relevantes. Note también que la salida muestra el estadístico F para H_0: $\sigma_1 = \sigma_2$. con un valor P de 0.0098, es decir, evidencia adicional de que se debe esperar más variabilidad cuando se usa nitrógeno, en comparación con la condición "sin nitrógeno".

Ejercicios

10.67 Se sabe que el volumen de los envases de un lubricante específico se distribuye normalmente con una varianza de 0.03 litros. Pruebe la hipótesis de que $\sigma^2 = 0.03$ contra la alternativa de que $\sigma^2 \neq 0.03$ para la muestra aleatoria de 10 envases del ejercicio 10.25 de la página 357. Use un valor P en sus conclusiones.

10.68 Por experiencia se sabe que el tiempo que se requiere para que los estudiantes de preparatoria de último año completen una prueba estandarizada es una variable aleatoria normal, con una desviación estándar de 6 minutos. Pruebe la hipótesis de que $\sigma = 6$ contra la alternativa de que $\sigma < 6$, si una muestra aleatoria de 20 estudiantes de preparatoria de último año tiene una desviación estándar $s = 4.51$. Utilice un nivel de significancia de 0.05.

10.69 Se deben supervisar las aflotoxinas ocasionadas por moho en cosechas de cacahuate en Virginia. Una muestra de 64 lotes de cacahuate revela niveles de 24.17 ppm, en promedio, con una varianza de 4.25 ppm. Pruebe la hipótesis de que $\sigma^2 = 4.2$ ppm con la alternativa de que $\sigma^2 \neq 4.2$ ppm. Utilice un valor P en sus conclusiones..

10.70 Datos históricos indican que la cantidad de dinero que aportaron los residentes trabajadores de una ciudad grande para un escuadrón de rescate voluntario es una variable aleatoria normal con una desviación estándar de $1.40. Se sugiere que las contribuciones al escuadrón de rescate sólo de los empleados del departamento de sanidad son mucho más variables. Si las contribuciones de una muestra aleatoria de 12 empleados del departamento de sanidad tienen una desviación estándar de $1.75, ¿podemos concluir con un nivel de significancia de 0.01 que la desviación estándar de las contribuciones de todos los trabajadores de sanidad es mayor que la de todos los trabajadores que viven en dicha ciudad?

10.71 Se dice que una máquina despachadora de bebida gaseosa está fuera de control si la varianza de los contenidos excede 1.15 decilitros. Si una muestra aleatoria de 25 bebidas de esta máquina tiene una varianza de 2.03 decilitros, ¿esto indica con un nivel de significancia de 0.05 que la máquina está fuera de control? Suponga que los contenidos se distribuyen de forma aproximadamente normal.

10.72 Prueba de $\sigma^2 = \sigma_0^2$ para una muestra grande: Cuando $n \geq 30$ podemos probar la hipótesis nula de que $\sigma^2 = \sigma_0^2$ o $\sigma^2 = \sigma_0$, al calcular

$$z = \frac{s - \sigma_0}{\sigma_0/\sqrt{2n}},$$

que es un valor de una variable aleatoria cuya distribución de muestreo es aproximadamente la distribución normal estándar.

a) Con referencia al ejemplo 10.5 pruebe, con un nivel de significancia de 0.05, si $\sigma = 10.0$ años contra la alternativa de que $\sigma \neq 10.0$ años.

b) Se sospecha que la varianza de la distribución de distancias en kilómetros logrados en 5 litros de combustible, por un modelo nuevo de automóvil equipado con un motor diesel, es menor que la varianza de la distribución de distancias lograda por el mismo modelo equipado con un motor de gasolina de seis cilindros, que se sabe es $\sigma^2 = 6.25$. Si 72 recorridos de prueba en el modelo diesel tienen una varianza de 4.41, ¿podemos concluir con un nivel de significancia de 0.05 que la varianza de las distancias alcanzadas por el modelo diesel es menor que la del modelo de gasolina?

10.73 Se realiza un estudio para comparar la longitud de tiempo entre hombres y mujeres para ensamblar cierto producto. La experiencia indica que la distribución de los tiempos tanto para hombres como para mujeres es aproximadamente normal, pero que la varianza de los tiempos para las mujeres es menor que para los hombres. Una muestra aleatoria de tiempos para 11 hombres y 14 mujeres da los siguientes datos:

Hombres	**Mujeres**
$n_1 = 11$	$n_2 = 14$
$s_1 = 6.1$	$s_2 = 5.3$

Pruebe la hipótesis de que $\sigma_1^2 = \sigma_2^2$ contra la alternativa de que $\sigma_1^2 > \sigma_2^2$. Utilice un valor P en su conclusión.

10.74 En el ejercicio 10.41 de la página 359, pruebe la hipótesis al nivel de significancia de 0.05 de que $\sigma_1^2 = \sigma_2^2$ contra la alternativa de que $\sigma_1^2 \neq \sigma_2^2$, donde σ_1^2 y σ_2^2 son las varianzas para el número de organismos por metro cuadrado en los dos diferentes lugares de Cedar Run.

10.75 Con referencia al ejercicio 10.39 de la página 359, pruebe la hipótesis de que $\sigma_1^2 = \sigma_2^2$ contra la alternativa de que $\sigma_1^2 \neq \sigma_2^2$, donde σ_1^2 y σ_2^2 son las varianzas para los tiempos de duración de películas producidas por la compañía 1 y la compañía 2, respectivamente. Utilice un valor P.

10.76 Se comparan dos tipos de instrumentos para medir la cantidad de monóxido de azufre en la atmósfera, en un experimento sobre la contaminación del aire. Se desea determinar si los dos tipos de instrumentos dan mediciones que tengan la misma variabilidad. Se registran las siguientes lecturas para los dos instrumentos:

Monóxido de azufre	
Instrumento A	**Instrumento B**
0.86	0.87
0.82	0.74
0.75	0.63
0.61	0.55

Monóxido de azufre	
Instrumento A	**Instrumento B**
0.89	0.76
0.64	0.70
0.81	0.69
0.68	0.57
0.65	0.53

Suponiendo que las poblaciones de mediciones se distribuyen de forma aproximadamente normal, pruebe la hipótesis de que $\sigma_A = \sigma_B$, contra la alternativa de que $\sigma_A \neq \sigma_B$. Use un valor P.

10.77 Se lleva a cabo un experimento para comparar el contenido de alcohol en una salsa de soya en dos líneas de producción diferentes. La producción se supervisa ocho veces al día. Los datos son los que aquí se muestran.

```
Línea de producción 1:
    0.48 0.39 0.42 0.52 0.40 0.48 0.52 0.52
Línea de producción 2:
    0.38 0.37 0.39 0.41 0.38 0.39 0.40 0.39
```

Suponga que ambas poblaciones son normales. Se sospecha que la línea de producción 1 no produce con la consistencia de la línea 2 en términos de contenido de alcohol. Pruebe la hipótesis de que $\sigma_1 = \sigma_2$ contra la alternativa de que $\sigma_1 \neq \sigma_2$. Utilice un valor P.

10.78 Se sabe que las emisiones de hidrocarburos disminuyeron de forma dramática durante la década de 1980. Se realizó un estudio para comparar las emisiones de hidrocarburos a velocidad estacionaria, en partes por millón (ppm), para automóviles de 1980 y 1990. Se seleccionaron al azar 20 automóviles de cada modelo y se registraron sus niveles de emisión de hidrocarburos. Los datos son los siguientes:

```
Modelos 1980:
 141 359 247 940 882 494 306 210 105 880
 200 223 188 940 241 190 300 435 241 380
Modelos 1990:
 140 160  20 20 223  60  20  95 360 70
 220 400 217 58 235 380 200 175  85 65
```

Pruebe la hipótesis de que $\sigma_1 = \sigma_2$ contra la alternativa de que $\sigma_1 \neq \sigma_2$. Suponga que ambas poblaciones son normales. Utilice un valor P.

10.14 Prueba de la bondad de ajuste

A lo largo de este capítulo nos ocupamos de la prueba de hipótesis estadísticas acerca de parámetros de una sola población como μ, σ^2 y p. Ahora consideraremos una prueba para determinar si una población tiene una distribución teórica específica. La prueba se basa en qué tan buen ajuste tenemos, entre la frecuencia de ocurrencia de las observaciones en una muestra observada y las frecuencias esperadas que se obtienen a partir de la distribución hipotética.

Tabla 10.4: Frecuencias esperadas y observadas de 120 lanzamientos de un dado

Frecuencia:	**1**	**2**	**3**	**4**	**5**	**6**
Observada	20	22	17	18	19	24
Esperada	20	20	20	20	20	20

Para ilustrar, considere el lanzamiento de un dado. Elaboramos la hipótesis de que el dado es legal, lo cual equivale a probar la hipótesis de que la distribución de resultados es la distribución uniforme discreta

$$f(x) = \frac{1}{6}, \quad x = 1, 2, ..., 6.$$

Suponga que el dado se lanza 120 veces y que se registra cada resultado. Teóricamente, si el dado está balanceado, esperaríamos que cada cara ocurriera 20 veces. Los resultados se dan en la tabla 10.4. Al comparar las frecuencias observadas con las frecuencias esperadas correspondientes, debemos decidir si es posible que tales discrepancias ocurran como resultado de fluctuaciones del muestreo y de que el dado está balanceado o que éste no es legal, y de que la distribución de resultados no es

uniforme. Es práctica común referirse a cada resultado posible de un experimento como una celda. De aquí, en nuestro caso tenemos 6 celdas. La estadística adecuada en la que basamos nuestro criterio de decisión para un experimento que incluye k celdas se define mediante el siguiente teorema.

Prueba de la bondad de ajuste

Una **prueba de la bondad de ajuste** entre las frecuencias observadas y esperadas se basa en la cantidad

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

donde χ^2 es un valor de una variable aleatoria cuya distribución muestral se aproxima muy de cerca con la distribución chi cuadrada con $v = k - 1$ grados de libertad. Los símbolos o_i y e_i representan las frecuencias observada y esperada, respectivamente, para la i-ésima celda.

El número de grados de libertad que se asocia con la distribución chi cuadrada que se utiliza aquí es igual a $k - 1$, pues hay sólo $k - 1$ frecuencias de celdas libremente determinadas. Es decir, una vez que se determinan las frecuencias de $k - 1$ celdas queda determinada la frecuencia para la k-ésima celda.

Si las frecuencias observadas están cerca de las frecuencias esperadas correspondientes, el valor χ^2 será pequeño, lo cual indica un buen ajuste. Si las frecuencias observadas difieren de manera considerable de las frecuencias esperadas, el valor χ^2 será grande, y el ajuste, deficiente. Un buen ajuste conduce a la aceptación de H_0; mientras que un ajuste deficiente conduce a su rechazo. La región crítica caerá, por lo tanto, en la cola derecha de la distribución chi cuadrada. Para un nivel de significancia igual a α, encontramos el valor crítico χ^2_α de la tabla A.5 y, entonces, $\chi^2 > \chi^2_\alpha$ constituye la región crítica. **El criterio de decisión que aquí se describe no se debería utilizar, a menos que cada una de las frecuencias esperadas sea al menos igual a 5**. Esta restricción podría requerir la combinación de celdas adyacentes, lo que tiene como resultado una reducción en el número de grados de libertad.

De la tabla 10.4, encontramos que el valor χ^2 es

$$\begin{aligned}\chi^2 =& \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} \\ &+ \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1.7.\end{aligned}$$

Usando la tabla A.5, encontramos $\chi^2_{0.05} = 11.070$ para $v = 5$ grados de libertad. Como 1.7 es menor que el valor crítico, no se rechaza H_0. Concluimos que no hay suficiente evidencia de que el dado no está balanceado.

Como segunda ilustración, probemos la hipótesis de que la distribución de frecuencias de las duraciones de baterías dadas en la tabla 1.7 de la página 23 puede aproximarse mediante una distribución normal con media $\mu = 3.5$ y una desviación estándar $\sigma = 0.7$. Las frecuencias esperadas para las 7 clases (celdas), que se listan en la tabla 10.5, se obtienen al calcular las áreas bajo la curva normal hipotética que caen entre los diversos límites de clase.

Por ejemplo, los valores z que corresponden a los límites de la cuarta clase son

$$z_1 = \frac{2.95 - 3.5}{0.7} = -0.79 \quad \text{y} \quad z_2 = \frac{3.45 - 3.5}{0.7} = -0.07.$$

Tabla 10.5: Frecuencias observadas y esperadas para las duraciones de las baterías con la suposición de normalidad

Límites de clase	o_i		e_i	
$1.45 - 1.95$	2		0.5	
$1.95 - 2.45$	1	7	2.1	8.5
$2.45 - 2.95$	4		5.9	
$2.95 - 3.45$	15		10.3	
$3.45 - 3.95$	10		10.7	
$3.95 - 4.45$	5	8	7.0	10.5
$4.45 - 4.95$	3		3.5	

De la tabla A.3 encontramos que el área entre $z_1 = -0.79$ y $z_2 = -0.07$ es

$$\begin{aligned}\text{área} &= P(-0.79 < Z < -0.07) = P(Z < -0.07) - P(Z < -0.79)\\ &= 0.4721 - 0.2148 = 0.2573.\end{aligned}$$

De aquí, la frecuencia esperada para la cuarta clase es

$$e_4 = (0.2573)\ (40) = 10.3.$$

Se acostumbra redondear estas frecuencias a un decimal.

La frecuencia esperada para el primer intervalo de clase se obtiene al utilizar el área total bajo la curva normal a la izquierda del límite 1.95. Para el último intervalo de clase, usamos el área total a la derecha del límite 4.45. Todas las demás frecuencias esperadas se determinan utilizando el método que se describe para la cuarta clase. Observe que combinamos clases adyacentes en la tabla 10.5, donde las frecuencias esperadas son menores que 5. En consecuencia, el número total de intervalos se reduce de 7 a 4, lo cual tiene como resultado $v = 3$ grados de libertad. El valor χ^2 está dado entonces por

$$\chi^2 = \frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} = 3.05.$$

Como el valor χ^2 calculado es menor que $\chi^2_{0.05} = 7.815$ para 3 grados de libertad, no tenemos razón para rechazar la hipótesis nula y concluimos que la distribución normal con $\mu = 3.5$ y $\sigma = 0.7$ brinda un buen ajuste para la distribución de duraciones de las baterías.

La prueba de bondad de ajuste chi cuadrada es un recurso importante, en particular dado que muchos procedimientos estadísticos en la práctica dependen, en un sentido teórico, de la suposición de que los datos reunidos provienen de un tipo de distribución específico. Como ya vimos, la suposición de normalidad se hace con bastante frecuencia. En los siguientes capítulos continuaremos haciendo *suposiciones de normalidad*, con la finalidad de proporcionar una base teórica para ciertas pruebas e intervalos de confianza.

En la literatura hay pruebas que son más poderosas que la prueba chi cuadrada para demostrar la normalidad. Una de tales pruebas es la **prueba de Geary**, la cual se basa en un estadístico muy sencillo que es una razón de dos estimadores de la desviación estándar poblacional σ. Suponga que se toma una muestra aleatoria

$X_1, X_2, \ldots, X_n$ de una distribución normal, $N(\mu, \sigma)$. Considere la razón

$$U = \frac{\sqrt{\pi/2}\sum_{i=1}^{n}|X_i - \bar{X}|/n}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n}}.$$

El lector debería reconocer que el denominador es un estimador razonable de σ si la distribución es normal o no. El numerador es un buen estimador de σ si la distribución es normal; sin embargo, podría sobrestimar o subestimar σ cuando haya desviaciones de la normalidad. Así, los valores de U que difieren considerablemente de 1.0 representan la señal de que se debería rechazar la hipótesis de normalidad.

Para muestras grandes, una prueba razonable se basa en la normalidad aproximada de U. El estadístico de prueba es, entonces, una estandarización de U, dada por

$$Z = \frac{U - 1}{0.2661/\sqrt{n}}.$$

Desde luego, el procedimiento de prueba incluye la región crítica bilateral. Calculamos un valor de z a partir de los datos y no rechazamos la hipótesis de normalidad cuando

$$-z_{\alpha/2} < Z < z_{\alpha/2}.$$

En la bibliografía se cita un artículo que trata sobre la prueba de Geary.

10.15 Prueba de independencia (datos categóricos)

El procedimiento de prueba de chi cuadrada, que se presenta en la sección 10.14, también se puede usar para probar la hipótesis de independencia de dos variables de clasificación. Suponga que deseamos determinar si las opiniones de los votantes residentes del estado de Illinois con respecto a una nueva reforma de impuestos son independientes de sus niveles de ingreso. Una muestra aleatoria de 1000 votantes registrados del estado de Illinois se clasifican de acuerdo con su posición en las categorías de ingreso bajo, medio o alto, y si están a favor o no de la nueva reforma de impuestos. Las frecuencias observadas se presentan en la tabla 10.6, que se conoce como **tabla de contingencia**.

Tabla 10.6: Tabla de contingencia 2 × 3

Reforma de impuestos	**Nivel de ingreso**			
	Bajo	**Medio**	**Alto**	**Total**
A favor	182	213	203	598
En contra	154	138	110	402
Total	336	351	313	1000

Una tabla de contingencia con r renglones y c columnas se denomina tabla $r \times c$ ("$r \times c$" se lee "r por c"). Los totales de renglones y columnas en la tabla 10.6 se denominan **frecuencias marginales**. Nuestra decisión de aceptar o rechazar la hipótesis nula, H_0, de independencia entre la opinión de un votante, con respecto a

la nueva reforma de impuestos y su nivel de ingreso, se basa en qué tan buen ajuste tengamos entre las frecuencias observadas en cada una de las 6 celdas de la tabla 10.6, y las frecuencias que esperaríamos para cada celda bajo la suposición de que H_0 es verdadera. Para encontrar estas frecuencias esperadas, definamos los siguientes eventos:

L: Una persona seleccionada está en el nivel de ingresos bajo.

M: Una persona seleccionada está en el nivel de ingresos medio.

H: Una persona seleccionada está en el nivel de ingresos alto.

F: Una persona seleccionada está a favor de la nueva reforma de impuestos.

A: Una persona seleccionada está en contra de la nueva reforma de impuestos.

Usando las frecuencias marginales, podemos listar las siguientes estimaciones de probabilidad:

$$P(L) = \frac{336}{1000}, \qquad P(M) = \frac{351}{1000}, \qquad P(H) = \frac{313}{1000},$$
$$P(F) = \frac{598}{1000}, \qquad P(A) = \frac{402}{1000}.$$

Ahora bien, si H_0 es verdadera y las dos variables son independientes, deberíamos tener

$$\begin{aligned}
P(L \cap F) &= P(L)P(F) = \left(\frac{336}{1000}\right)\left(\frac{598}{1000}\right), \\
P(L \cap A) &= P(L)P(A) = \left(\frac{336}{1000}\right)\left(\frac{402}{1000}\right), \\
P(M \cap F) &= P(M)P(F) = \left(\frac{351}{1000}\right)\left(\frac{598}{1000}\right), \\
P(M \cap A) &= P(M)P(A) = \left(\frac{351}{1000}\right)\left(\frac{402}{1000}\right), \\
P(H \cap F) &= P(H)P(F) = \left(\frac{313}{1000}\right)\left(\frac{598}{1000}\right), \\
P(H \cap A) &= P(H)P(A) = \left(\frac{313}{1000}\right)\left(\frac{402}{1000}\right).
\end{aligned}$$

Las frecuencias esperadas se obtienen al multiplicar cada probabilidad de una celda por el número total de observaciones. Como antes, redondeamos estas frecuencias a un decimal. Así, se estima que el número esperado de votantes de bajo ingreso en nuestra muestra que favorecen la nueva reforma fiscal es

$$\left(\frac{336}{1000}\right)\left(\frac{598}{1000}\right)(1000) = \frac{(336)(598)}{1000} = 200.9,$$

cuando H_0 es verdadera. La regla general para obtener la frecuencia esperada de cualquier celda está dada por la siguiente fórmula:

$$\text{frecuencia esperada} = \frac{(\text{total de la columna}) \times (\text{total del renglón})}{\text{gran total}}$$

En la tabla 10.7 la frecuencia esperada para cada celda se registra entre paréntesis a un lado del valor observado real. Observe que las frecuencias esperadas en cual-

quier renglón o columna se suman al total marginal apropiado. En nuestro ejemplo necesitamos calcular sólo las dos frecuencias esperadas en el renglón superior de la tabla 10.7 y, después, encontrar las otras por sustracción. El número de grados de libertad asociados con la prueba chi cuadrada que aquí se usa es igual al número de frecuencias de celdas que se pueden llenar libremente, cuando se nos dan los totales marginales y el gran total, y en este caso este número es 2. Una fórmula simple que proporciona el número correcto de grados de libertad es

$$v = (r - 1)(c - 1).$$

Tabla 10.7: Frecuencias observadas y esperadas

	Nivel de ingreso			
Reforma fiscal	**Bajo**	**Medio**	**Alto**	**Total**
A favor	182 (200.9)	213 (209.9)	203 (187.2)	598
En contra	154 (135.1)	138 (141.1)	110 (125.8)	402
Total	336	351	313	1000

De aquí, para nuestro ejemplo, $v = (2 - 1)(3 - 1) = 2$ grados de libertad. Para probar la hipótesis nula de independencia, usamos el siguiente criterio de decisión:

Prueba de independencia

Calcule

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

donde la suma se extiende a todas las celdas rc en la tabla de contingencia $r \times c$. Si $\chi^2 > \chi^2_\alpha$ con $v = (r - 1)(c - 1)$ grados de libertad, rechace la hipótesis nula de independencia al nivel de significancia α; en cualquier otro caso, no rechace la hipótesis nula.

Al aplicar este criterio a nuestro ejemplo, encontramos que

$$\chi^2 = \frac{(182 - 200.9)^2}{200.9} + \frac{(213 - 209.9)^2}{209.9} + \frac{(203 - 187.2)^2}{187.2} + \frac{(154 - 135.1)^2}{135.1} + \frac{(138 - 141.1)^2}{141.1} + \frac{(110 - 125.8)^2}{125.8} = 7.85,$$

$P \approx 0.02$.

De la tabla A.5 encontramos que $\chi^2_{0.05} = 5.991$ para $v = (2 - 1)(3 - 1) = 2$ grados de libertad. Se rechaza la hipótesis nula y concluimos que la opinión de un votante con respecto a la nueva reforma fiscal y su nivel de ingresos no son independientes.

Es importante recordar que el estadístico sobre el cual basamos nuestra decisión tiene una distribución que sólo se aproxima por la distribución chi cuadrada. Los valores χ^2 calculados dependen de las frecuencias de las celdas y, en consecuencia, son discretos. La distribución chi cuadrada continua parece aproximar muy bien la distribución de muestreo discreta de χ^2, dado que el número de grados de libertad es mayor que 1. En una tabla de contingencia de 2×2, donde sólo tenemos 1 grado de libertad, se aplica una corrección llamada **corrección de Yates para continuidad**.

La fórmula corregida se vuelve entonces

$$\chi^2(\text{corregida}) = \sum_i \frac{(|o_i - e_i| - 0.5)^2}{e_i}.$$

Si las frecuencias de celdas esperadas son grandes, los resultados corregidos y sin corrección son casi los mismos. Cuando las frecuencias esperadas están entre 5 y 10, se debería aplicar la corrección de Yates. Para frecuencias esperadas menores que 5, se debería utilizar la prueba exacta de Fisher-Irwin. Una discusión de esta prueba se puede encontrar en *Basic Concepts of Probability and Statistics* de Hodges and Lehmann (véase la bibliografía). La prueba de Fisher-Irwin se puede evitar, sin embargo, mediante la elección de una muestra grande.

10.16 Prueba de homogeneidad

Cuando probamos la independencia en la sección 10.15, se seleccionó una muestra aleatoria de 1000 votantes, y los totales de renglón y columna para nuestra tabla de contingencia se determinaron al azar. Otro tipo de problema para el que se aplica el método de la sección 10.15 es aquel donde se predeterminan los totales de renglón y columna. Suponga, por ejemplo, que decidimos de antemano seleccionar 200 demócratas, 150 republicanos y 150 independientes de los votantes del estado de Carolina del Norte y registrar si favorecen una iniciativa de ley para el aborto, están en contra o están indecisos. Las respuestas observadas se dan en la tabla 10.8.

Tabla 10.8: Frecuencias observadas y esperadas

	Afiliación política			
Ley del aborto	**Demócrata**	**Republicano**	**Independiente**	**Total**
A favor	82	70	62	214
En contra	93	62	67	222
Indecisos	25	18	21	64
Total	200	150	150	500

Ahora bien, en vez de probar la independencia, probamos la hipótesis de que las proporciones de población dentro de cada renglón son las mismas. Es decir, probamos la hipótesis de que las proporciones de demócratas, republicanos e independientes que favorecen la ley sobre el aborto son las mismas; las proporciones de cada afiliación política contra la ley son las mismas; y las proporciones de cada afiliación política de quienes están indecisos son las mismas. Básicamente nos interesamos en determinar si las tres categorías de votantes son **homogéneas** con respecto a sus opiniones acerca de la iniciativa de ley sobre el aborto. Tal prueba se llama prueba de homogeneidad.

Al suponer homogeneidad, de nuevo encontramos las frecuencias esperadas de las celdas al multiplicar los totales de renglón y columna correspondientes, y después dividir entre el gran total. El análisis entonces continúa al utilizar el mismo estadístico chi cuadrada como antes. Ilustramos este proceso en el siguiente ejemplo para los datos de la tabla 10.8.

Ejemplo 10.15: Con referencia a los datos de la tabla 10.8, pruebe la hipótesis de que las opiniones con respecto a la ley del aborto propuesta son las mismas dentro de cada afiliación política. Utilice un nivel de significancia de 0.05.

Solución: **1.** H_0: Para cada opinión las proporciones de demócratas, republicanos e independientes son las mismas.

2. H_1: Para al menos una opinión, las proporciones de demócratas, republicanos e independientes no son las mismas.

3. $\alpha = 0.05$.

4. Región crítica: $\chi^2 > 9.488$ con $v = 4$ grados de libertad.

5. Cálculos: Al usar la fórmula de las frecuencias de celdas esperadas de la página 375 necesitamos calcular las 4 frecuencias de celdas. Todas las demás frecuencias se encuentran por sustracción. Las frecuencias de celdas observadas y esperadas se muestran en la tabla 10.9.

Tabla 10.9: Frecuencias observadas y esperadas

	Afiliación política			
Ley del aborto	**Demócrata**	**Republicano**	**Independiente**	**Total**
A favor	82 (85.6)	70 (64.2)	62 (64.2)	214
En contra	93 (88.8)	62 (66.6)	67 (66.6)	222
Indecisos	25 (25.6)	18 (19.2)	21 (19.2)	64
Total	200	150	150	500

Así,

$$\begin{aligned}\chi^2 &= \frac{(82-85.6)^2}{85.6} + \frac{(70-64.2)^2}{64.2} + \frac{(62-64.2)^2}{64.2} \\ &\quad + \frac{(93-88.8)^2}{88.8} + \frac{(62-66.6)^2}{66.6} + \frac{(67-66.6)^2}{66.6} \\ &\quad + \frac{(25-25.6)^2}{25.6} + \frac{(18-19.2)^2}{19.2} + \frac{(21-19.2)^2}{19.2} \\ &= 1.53.\end{aligned}$$

6. Decisión: No rechace H_0. No hay suficiente evidencia para concluir que la proporción de demócratas, republicanos e independientes difieren para cada opinión establecida.

10.17 Prueba para varias proporciones

El estadístico chi cuadrada para probar la homogeneidad también se aplica cuando se prueba la hipótesis de que k parámetros binomiales tienen el mismo valor. Ésta es, por lo tanto, una extensión de la prueba que se presentó en la sección 10.12, para determinar diferencias entre dos proporciones a una prueba para determinar diferencias entre k proporciones. Por ello, nos interesamos en probar la hipótesis nula

$$H_0\text{: } p_1 = p_2 = \cdots = P_k$$

contra la hipótesis alternativa, H_1, de que las proporciones de la población *no son todas iguales*. Para ejecutar esta prueba, primero observamos muestras aleatorias

Tabla 10.10: k muestras binomiales independientes

Muestras:	**1**	**2**	$\cdots$	$\boldsymbol{k}$
Éxitos	x_1	x_2	$\cdots$	x_k
Fracasos	$n_1 - x_1$	$n_2 - x_2$	$\cdots$	$n_k - x_k$

independientes de tamaño n_1, $n_2, \ldots,$ n_k de las k poblaciones y acomodamos los datos como en la tabla de contingencia $2 \times k$, tabla 10.10.

Según si los tamaños de las muestras aleatorias se predeterminaron u ocurrieron al azar, el procedimiento de prueba es idéntico a la prueba de homogeneidad o a la prueba de independencia. Por lo tanto, las frecuencias de celdas esperadas se calculan como antes y se sustituyen junto con las frecuencias observadas en el estadístico chi cuadrada

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

con

$$v = (2-1)(k-1) = k-1$$

grados de libertad.
Al seleccionar la región crítica apropiada de la cola superior de la forma $\chi^2 > \chi^2_\alpha$, podemos llegar ahora a una decisión con respecto a H_0.

Ejemplo 10.16: En un estudio sobre un taller, se reúne un conjunto de datos para determinar si la proporción de artículos defectuosos producida por los trabajadores fue la misma para el turno matutino, el vespertino o el nocturno. Los datos se reunieron y se presentan en la tabla 10.11:

Tabla 10.11: Datos para el ejemplo 10.16

Turno:	**Matutino**	**Vespertino**	**Nocturno**
Defectuosos	45	55	70
No defectuosos	905	890	870

Utilice un nivel de significancia de 0.025 para determinar si la proporción de defectuosos es la misma para los tres turnos.

Solución: Representemos con p_1, p_2 y p_3 la proporción real de defectuosos para los turnos matutino, vespertino y nocturno, respectivamente.

1. H_0: $p_1 = p_2 = p_3$.
2. H_1: p_1, p_2 y p_3 no todas son iguales.
3. $\alpha = 0.025$.
4. Región crítica: $\chi^2 > 7.378$ para $v = 2$ grados de libertad.
5. Cálculos: En correspondencia con las frecuencias observadas $o_1 = 45$ y $o_2 = 55$, encontramos

$$e_1 = \frac{(950)(170)}{2835} = 57.0 \quad \text{y} \quad e_2 = \frac{(945)(170)}{2835} = 56.7.$$

Todas las demás frecuencias esperadas se encuentran por sustracción y se incluyen en la tabla 10.12.

Tabla 10.12: Frecuencias esperadas y observadas

Turno:	**Matutino**	**Vespertino**	**Nocturno**	**Total**
Defectuosos	45 (57.0)	55 (56.7)	70 (56.3)	170
No defectuosos	905 (893.0)	890 (888.3)	870 (883.7)	2665
Total	950	945	940	2835

Ahora bien,

$$\chi^2 = \frac{(45-57.0)^2}{57.0} + \frac{(55-56.7)^2}{56.7} + \frac{(70-56.3)^2}{56.3} + \frac{(905-893.0)^2}{893.0} + \frac{(890-888.3)^2}{888.3} + \frac{(870-883.7)^2}{883.7} = 6.29,$$

$P \approx 0.04$.

6. Decisión: no rechazamos H_0 con $\alpha = 0.025$. No obstante, con el anterior valor P calculado, ciertamente sería riesgoso concluir que la proporción de defectuosos producidos es la misma para todos los turnos. ∎

10.18 Estudio de caso de dos muestras

En esta sección consideramos un estudio donde demostramos un análisis completo usando tanto el análisis gráfico como el formal, junto con salidas por computadora comentados y conclusiones. En un estudio del análisis de datos que realizó el personal del Centro de Consulta Estadística del Tecnológico de Virginia, se compararon dos materiales diferentes, digamos la aleación A y la aleación B, en términos de la resistencia de rotura. La aleación B es más cara, aunque en realidad se debería adoptar si se demuestra que es más fuerte que la aleación A. Se debe tomar en cuenta la consistencia del rendimiento de las dos aleaciones.

Se seleccionaron muestras aleatorias de vigas de cada aleación y la resistencia se midió en una deflexión de 0.001 pulgadas cuando se aplicó una fuerza fija en ambos extremos de la viga. Se utilizaron 20 especímenes para cada una de las dos aleaciones. Los datos se presentan en la tabla 10.13.

Es importante que el ingeniero compare las dos aleaciones. La preocupación es la resistencia promedio y la reproducibilidad. Interesa determinar si hay una transgresión seria de la suposición de normalidad que requieren las pruebas t y F. Las figuras 10.21 y 10.22 son gráficas de cuantil-cuantil normales de las muestras para las dos aleaciones.

No parece haber ninguna transgresión seria de la suposición de normalidad. Además, la figura 10.23 muestra dos gráficos de caja y extensión en la misma gráfica. Los gráficos de caja y extensión sugieren que no hay una diferencia apreciable de la variabilidad en la deflexión para las dos aleaciones. Sin embargo, parece que la media de la aleación B es significativamente menor, lo cual sugiere (gráficamente al menos) que la aleación B es más fuerte. Las medias muestrales y las desviaciones estándar son

$$\bar{y}_A = 83.55, \quad s_A = 3.663; \qquad \bar{y}_B = 79.70, \quad s_B = 3.097.$$

Tabla 10.13: Datos para el estudio de caso de dos muestras

Aleación *A*			**Aleación *B***		
88	82	87	75	81	80
79	85	90	77	78	81
84	88	83	86	78	77
89	80	81	84	82	78
81	85		80	80	
83	87		78	76	
82	80		83	85	
79	78		76	79	

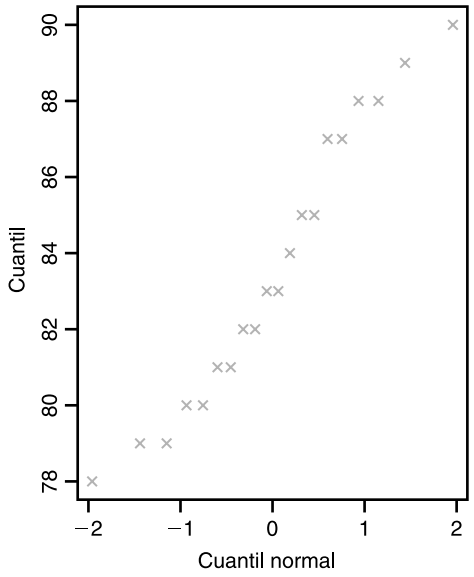


Figura 10.21: Gráfica de cuantil-cuantil normal de los datos para la aleación A.

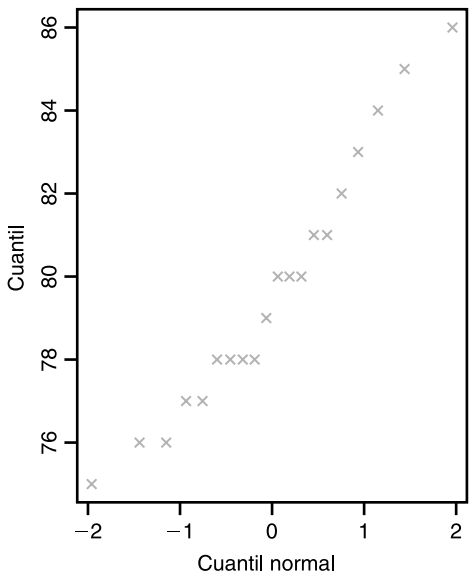


Figura 10.22: Gráfica de cuantil-cuantil normal de los datos para la aleación B.

La salida del SAS para el PROC TTEST se presenta en la figura 10.24. La prueba F sugiere que no hay diferencia significativa en las varianzas ($P = 0.4709$) y el estadístico t de dos muestras para probar

$$H_0\text{: } \mu_A = \mu_B,$$
$$H_1\text{: } \mu_A > \mu_B,$$

($t = 3.59$, $P = 0.0009$) rechaza H_0 en favor de H_1 y, de esta manera, confirma lo que sugiere la información gráfica. Aquí utilizamos la prueba t que reúne las varianzas de dos muestras a la luz de los resultados de la prueba F. Con base en este análisis sería adecuada la adopción de la aleación B.

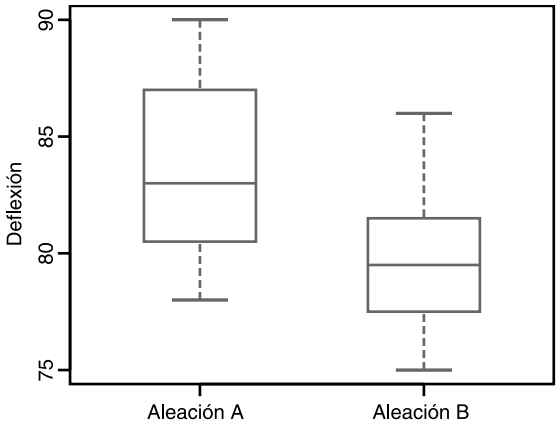


Figura 10.23: Gráficos de caja y extensión para ambas aleaciones.

```
                The TTEST Procedure
  Alloy          N    Mean  Std Dev  Std Err
Alloy A         20   83.55   3.6631   0.8191
Alloy B         20    79.7   3.0967   0.6924

  Variances      DF    t Value    Pr > |t|
  Equal          38       3.59      0.0009
  Unequal        37       3.59      0.0010
           Equality of Variances
   Num DF    Den DF    F Value    Pr > F
       19        19       1.40    0.4709
```

Figura 10.24: Salida del SAS comentada para los datos de las aleaciones.

Significancia estadística, y significancia científica o en ingeniería

Mientras que el estadístico se puede sentir muy a gusto con los resultados de la comparación entre las dos aleaciones en el estudio anterior, queda un dilema para el ingeniero. El análisis demuestra una mejoría estadísticamente significativa con el uso de la aleación B. Sin embargo, ¿se encuentra que la diferencia realmente valga la pena si la aleación B es más cara? Esta ilustración resalta un asunto muy importante que con frecuencia se pasa por alto por los estadísticos y el analista de datos: *la distinción entre significancia estadística y significancia científica o en ingeniería.* Aquí la diferencia promedio en la deflexión es $\bar{y}_A - \bar{y}_B = 0.00385$ pulgadas. En un análisis completo el ingeniero debe determinar si la diferencia es suficiente para justificar el costo extra en el largo plazo. Éste es un asunto económico y de ingeniería. El lector debería comprender que una diferencia estadísticamente significativa tan sólo implica que la diferencia en las medias muestrales que se encuentra en los datos con dificultad podría ocurrir al azar. Esto no implica que la diferencia en las medias poblacionales sea profunda o particularmente significativa en el contexto del problema. Por ejemplo, en la sección 10.7, se utiliza una salida por computadora

para demostrar la evidencia de que un medidor de pH está, de hecho, sesgado. Es decir, esto no demuestra un pH medio de 7.00 para el material en que se probó. Sin embargo, la variabilidad entre las observaciones en la muestra es muy pequeña. El ingeniero puede decidir que las desviaciones pequeñas de 7.0 hacen que el medidor de pH sea bastante adecuado.

Ejercicios

10.79 Se lanza 180 veces un dado con los siguientes resultados:

x	1	2	3	4	5	6
f	28	36	36	30	27	23

¿Es un dado balanceado? Utilice un nivel de significancia de 0.01.

10.80 En 100 lanzamientos de una moneda se observan 63 caras y 37 cruces. ¿Es una moneda balanceada? Utilice un nivel de significancia de 0.05.

10.81 Se supone que una máquina mezcla cacahuates, avellanas, anacardos y pacanas a razón de 5:2:2:1. Se encuentra que una lata que contiene 500 de tales nueces mezcladas tiene 269 cacahuates, 112 avellanas, 74 anacardos y 45 pacanas. Al nivel de significancia de 0.05, pruebe la hipótesis de que la máquina mezcla las nueces a una razón de 5:2:2:1.

10.82 Las calificaciones de un curso de estadística para un semestre específico fueron las siguientes:

Calificación	A	B	C	D
f	14	18	32	20

Pruebe la hipótesis, al nivel de significancia de 0.05, de que la distribución de calificaciones es uniforme.

10.83 Se extraen 3 cartas de una baraja ordinaria, con reemplazo, y se registra el número Y de espadas. Después de repetir el experimento 64 veces, se registran los siguientes resultados:

y	0	1	2	3
f	21	31	12	0

Con un nivel de significancia de 0.01, pruebe la hipótesis de que los datos registrados se pueden ajustar mediante la distribución binomial $b(y;\, 3,\, 1/4)$, $y = 0,\, 1,\, 2,\, 3$.

10.84 Se seleccionan tres canicas de una urna que contiene 5 canicas rojas y 3 verdes. Después de registrar el número X de canicas rojas, las canicas se reemplazan en la urna y el experimento se repite 112 veces. Los resultados que se obtienen son los siguientes:

x	0	1	2	3
f	1	31	55	25

Con un nivel de significancia de 0.05, pruebe la hipótesis de que los datos registrados se pueden ajustar con la distribución hipergeométrica $h(x,\, 8,\, 3,\, 5)$, $x = 0,\, 1,\, 2,\, 3$.

10.85 Se lanza una moneda hasta que sale una cara y se registra el número de lanzamientos X. Después de repetir el experimento 256 veces, obtenemos los siguientes resultados:

x	1	2	3	4	5	6	7	8
f	136	60	34	12	9	1	3	1

Con un nivel de significancia de 0.05 pruebe la hipótesis de que la distribución observada de X se puede ajustar por la distribución geométrica $g(x;\, 1/2)$, $x = 1,\, 2,\, 3, \ldots$

10.86 Repita el ejercicio 10.83 con un conjunto nuevo de datos obtenidos al llevar a cabo realmente 64 veces el experimento que se describe.

10.87 Repita el ejercicio 10.85 con el nuevo conjunto de datos obtenidos al realizar 256 veces el experimento que se describe.

10.88 En el ejercicio 1.18 de la página 28, pruebe la bondad de ajuste entre las frecuencias de clase que se observan, y las frecuencias esperadas correspondientes de una distribución normal con $\mu = 65$ y $\sigma = 21$. Utilice un nivel de significancia de 0.05.

10.89 En el ejercicio 1.19 de la página 28, pruebe la bondad del ajuste entre las frecuencias de clase que se observan y las frecuencias esperadas correspondientes de una distribución normal con $\mu = 1.8$ y $\sigma = 0.4$. Utilice un nivel de significancia de 0.01.

10.90 En un experimento para estudiar la dependencia de la hipertensión con respecto a los hábitos de fumar, se tomaron los siguientes datos de 180 individuos:

	No fumadores	**Fumadores moderados**	**Fumadores empedernidos**
Con hipertensión	21	36	30
Sin hipertensión	48	26	19

Pruebe la hipótesis de que la presencia o ausencia de la hipertensión es independiente de los hábitos de fumar. Utilice un nivel de significancia de 0.05.

10.91 Una muestra aleatoria de 90 adultos se clasifica de acuerdo con su género y el número de horas que pasan viendo la televisión durante una semana:

	Sexo	
	Masculino	**Femenino**
Más de 25 horas	15	29
Menos de 25 horas	27	19

Utilice un nivel de significancia de 0.01 y pruebe la hipótesis de que el tiempo que pasan viendo televisión es independiente de si el espectador es hombre o mujer.

10.92 Una muestra aleatoria de 200 hombres casados, todos jubilados, se clasifica de acuerdo con la educación y el número de hijos:

	Número de hijos		
Educación	**0–1**	**2–3**	**Más de 3**
Primaria	14	37	32
Secundaria	19	42	17
Universidad	12	17	10

Con un nivel de significancia de 0.05, pruebe la hipótesis de que el tamaño de la familia es independiente del nivel académico del padre.

10.93 Un criminólogo realizó una investigación para determinar si, en una ciudad grande, la incidencia de ciertos tipos de delitos varía de una parte a otra. Los crímenes específicos de interés son asalto (con violencia), robo en casa, hurto y homicidio. La siguiente tabla muestra el número de delitos cometidos en cuatro áreas de la ciudad durante el año pasado.

	Tipo de crimen			
Distrito	**Asalto**	**Robo en casa**	**Hurto**	**Homicidio**
1	162	118	451	18
2	310	196	996	25
3	258	193	458	10
4	280	175	390	19

¿A partir de tales datos podemos concluir, con un nivel de significancia de 0.01, que la ocurrencia de estos tipos de delitos es dependiente del distrito de la ciudad?

10.94 El hospital de una universidad realizó un experimento para determinar el grado de alivio que brindan tres remedios para la tos. Cada medicamento para la tos se trata en 50 estudiantes y se registran los siguientes datos:

	Remedio para la tos		
	NyQuil	**Robitussin**	**Triaminic**
Sin alivio	11	13	9
Cierto alivio	32	28	27
Alivio completo	7	9	14

Con un nivel de significancia de 0.05, pruebe la hipótesis de que los tres remedios para la tos son igualmente efectivos.

10.95 Para determinar las posiciones actuales acerca de las oraciones en escuelas públicas, se llevó a cabo una investigación en cuatro condados de Virginia. La siguiente tabla da las opiniones de 200 padres del condado de Craig, 150 padres del de Giles, 100 padres del de Franklin y 100 del de Montgomery:

	Condado			
Posición	**Craig**	**Giles**	**Franklin**	**Mont.**
A favor	65	66	40	34
En contra	42	30	33	42
Sin opinión	93	54	27	24

Pruebe la homogeneidad de las opiniones entre los 4 condados con respecto a las oraciones en escuelas públicas. Utilice un valor P en sus conclusiones.

10.96 De acuerdo con un estudio de la Universidad Johns Hopkins publicado en *American Journal of Public Health*, las viudas viven más que los viudos. Considere los siguientes datos de supervivencia de 100 viudas y 100 viudos después de la muerte del cónyuge:

Años vividos	**Viuda**	**Viudo**
Menos de 5	25	39
de 5 a 10	42	40
Más de 10	33	21

¿Con un nivel de significancia de 0.05 podemos concluir que las proporciones de viudas y viudos son iguales con respecto a los diferentes periodos que un cónyuge sobrevive luego de la muerte de su compañero?

10.97 Las siguientes respuestas con respecto al estándar de vida al momento de una encuesta de opinión independiente de 1000 familias contra un año antes parece estar de acuerdo con los resultados de un estudio publicado en *Across the Board* (junio de 1981):

	Estándar de vida			
Periodo	**Algo mejor**	**Igual**	**No tan bueno**	**Total**
1980: Enero	72	144	84	300
Mayo	63	135	102	300
Septiembre	47	100	53	200
1981: Enero	40	105	55	200

Pruebe la hipótesis de que las proporciones de familias dentro de cada estándar de vida son las mismas para cada uno de los cuatro periodos. Utilice un valor P.

10.98 Se lleva a cabo un estudio en Indiana, Kentucky y Ohio, para determinar la postura de los votantes con respecto al transporte escolar. Una encuesta de 200 votantes de cada uno de estos estados da los siguientes resultados:

	Postura del votante		
Estado	**Apoya**	**No apoya**	**Indeciso**
Indiana	82	97	21
Kentucky	107	66	27
Ohio	93	74	33

Con un nivel de significancia de 0.025, pruebe la hipótesis nula de que las proporciones de votantes dentro de cada categoría de postura son las mismas para cada uno de los tres estados.

10.99 Se lleva a cabo una investigación en dos ciudades de Virginia, para determinar la opinión de los votantes hacia los candidatos a la gubernatura en una elección próxima. En cada ciudad se seleccionan 500 votantes al azar y se registran los siguientes datos:

	Ciudad	
Opinión del votante	**Richmond**	**Norfolk**
Favorece a A	204	225
Favorece a B	211	198
Indeciso	85	77

Con un nivel de significancia de 0.05, pruebe la hipótesis nula de que las proporciones de votantes que favorecen al candidato A, al candidato B o están indecisos son las mismas para cada ciudad.

10.100 En un estudio para estimar la proporción de esposas que de manera regular ven telenovelas, se encuentra que 52 de 200 esposas en Denver, 31 de 150 en Phoenix, y 37 de 150 en Rochester ven al menos una telenovela. Utilice un nivel de significancia de 0.05 para probar la hipótesis de que no hay diferencia entre las proporciones reales de esposas que ven telenovelas en esas tres ciudades.

Ejercicios de repaso

10.101 Un genetista se interesa en la proporción de hombres y mujeres de una población que tiene cierto trastorno sanguíneo menor. En una muestra aleatoria de 100 hombres, se encuentra que 31 lo padecen, mientras que sólo 24 de 100 mujeres parecen tener el trastorno. ¿Con un nivel de significancia de 0.01 podemos concluir que la proporción de hombres en la población con este trastorno sanguíneo es significativamente mayor que la proporción de mujeres afectadas?

10.102 Considere la situación del ejercicio 10.54 de la página 361. El consumo de oxígeno en ml/kg/min también se midió en los nueve sujetos.

Sujeto	**Con CO**	**Sin CO**
1	26.46	25.41
2	17.46	22.53
3	16.32	16.32
4	20.19	27.48
5	19.84	24.97
6	20.65	21.77
7	28.21	28.17
8	33.94	32.02
9	29.32	28.96

Se conjetura que el consumo de oxígeno debería ser mayor en un ambiente relativamente libre de CO. Realice una prueba de significancia y discuta la conjetura.

10.103 Establezca las hipótesis nula y alternativa para utilizarse en la prueba de las siguientes afirmaciones y determine de manera general dónde se localiza la región crítica:

a) La caída de nieve promedio en el lago George durante el mes de febrero es 21.8 centímetros.

b) No más de 20% del cuerpo de profesores en la universidad local contribuyó a un fondo anual.

c) En promedio, los niños asisten a la escuela dentro de 6.2 kilómetros de sus casas en un suburbio de St. Louis.

d) Al menos 70% de los automóviles nuevos del siguiente año caerán en la categoría de compactos y semicompactos.

e) La proporción de votantes que favorecen al funcionario actual en la próxima elección es 0.58.

f) El filete rib-eye promedio en el restaurante Longhorn Steak es de al menos 340 gramos.

10.104 Se realiza un estudio para determinar si, en las bodas, más italianos que estadounidenses prefieren la champaña blanca en vez de la rosada. De los 300 italianos que se seleccionaron al azar, 72 prefieren champaña blanca, y de los 400 estadounidenses seleccionados 70 prefieren champaña blanca en vez de la rosada. ¿Podemos concluir que una proporción mayor de italianos que de estadounidenses prefiere champaña blanca en las bodas? Utilice un nivel de significancia de 0.05.

10.105 En un conjunto de datos analizados por el Centro de Consulta Estadística del Instituto Politécnico y Universidad Estatal de Virginia, se solicitó a un grupo de sujetos completar cierta tarea en la computadora. La respuesta medida fue el tiempo de terminación. El propósito del experimento fue probar un grupo de herramientas de ayuda desarrolladas por el Departamento de Ciencias Computacionales del mismo instituto. Participaron 10 sujetos. Con una asignación al azar, a 5 se les dio un procedimiento estándar con lenguaje Fortran para completar la tarea. A los otros 5 se les pidió realizar la tarea usando las herramientas de ayuda. A continuación se presentan los datos de los tiempos de terminación de la tarea. Suponiendo que las distribuciones poblacionales son normales y las varianzas son las mismas para los dos grupos, apoye o rechace la conjetura de que las herramientas de ayuda aumentan la velocidad con la que se realiza la tarea.

Grupo 1 (Procedimiento estándar)	Grupo 2 (Herramienta de ayuda)
161	132
169	162
174	134
158	138
163	133

10.106 Establezca las hipótesis nula y alternativa a utilizar en la prueba de las siguientes afirmaciones, y determine de manera general dónde se ubica la región crítica:

a) A lo más 20% de la cosecha de trigo del próximo año se exportará a la Unión Soviética.

b) En promedio, las amas de casa estadounidenses beben 3 tazas de café al día.

c) La proporción de graduados en Virginia este año que se especializan en ciencias sociales es al menos 0.15.

d) La donación promedio a la Asociación Estadounidense del Pulmón es no más de $10.

e) Los residentes del suburbio Richmond recorren, en promedio, 15 kilómetros hasta su lugar de trabajo.

10.107 Si se selecciona al azar una lata que contiene 500 nueces de cada uno de tres diferentes distribuidores de nueces surtidas y contienen, respectivamente, 345, 313 y 359 cacahuates en cada una de las latas, ¿con un nivel de significancia de 0.01 podemos concluir que las nueces surtidas de los tres distribuidores contienen proporciones iguales de cacahuates?

10.108 Valor *z* para probar $p_1 - p_2 = d_0$. Para probar la hipótesis nula H_0 de que $p_1 - p_2 = d_0$, donde $d_0 \neq 0$, basamos nuestra decisión en

$$z = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2}},$$

que es un valor de una variable aleatoria, cuya distribución aproxima la distribución normal estándar, en tanto que n_1 y n_2 sean grandes. Con referencia al ejemplo 10.12 de la página 365, pruebe la hipótesis de que el porcentaje de votantes de la ciudad que favorecen la construcción de la planta química no excederá el porcentaje de votantes del condado en más de 3%. Utilice un valor P en su conclusión.

10.109 Se realiza un estudio para determinar si hay una diferencia entre las proporciones de padres en los estados de Maryland (MD), Virginia (VA), Georgia (GA) y Alabama (AL) que están a favor de colocar Biblias en las escuelas primarias. En la siguiente tabla se registran las respuestas de 100 padres seleccionados al azar en cada uno de esos estados:

	Estado			
Preferencia	**MD**	**VA**	**GA**	**AL**
Sí	65	71	78	82
No	35	29	22	18

¿Podemos concluir que las proporciones de padres que están a favor de colocar Biblias en las escuelas son las mismas para estos cuatro estados? Utilice un nivel de significancia de 0.01.

10.110 Se lleva a cabo un estudio en el Centro de Medicina Veterinaria Equina de la Universidad Regional de Virginia-Maryland, para determinar si la realización de cierto tipo de cirugía en caballos jóvenes tiene algún efecto en ciertas clases de células sanguíneas en el animal. Se toman muestras del fluido de cada uno de seis potros antes y después de la cirugía. Se analizan las muestras para el número de leucogramos de glóbulos blancos (WBC) posoperatorios. También se realiza una medición de leucogramos WBC preoperatorios. Utilice una prueba t de una muestra pareada para determinar si hay un cambio significativo en los leucogramos WBC con la cirugía.

Potro	Precirugía*	Postcirugía*
1	10.80	10.60
2	12.90	16.60
3	9.59	17.20
4	8.81	14.00
5	12.00	10.60
6	6.07	8.60

*Todos los valores $\times 10^{-3}$.

10.111 Se lleva a cabo un estudio en el Departamento de Salud y Educación Física del Instituto Politécnico y Universidad Estatal de Virginia, para determinar si 8 semanas de entrenamiento realmente reducen los niveles de colesterol en los participantes. A un grupo de tratamiento que consiste en 15 personas se les dan conferencias dos veces a la semana de cómo reducir su nivel de colesterol. Otro grupo de 18 personas de edad similar se selecciona al azar como grupo de control. Se registran los niveles de colesterol de todos los participantes al final del programa de 8 semanas y se listan a continuación:

```
Grupo de tratamiento:
    129 131 154 172 115 126 175 191
    122 238 159 156 176 175 126
Grupo de control:
    151 132 196 195 188 198 187 168 115
    165 137 208 133 217 191 193 140 146
```

¿Podemos concluir, con un nivel de significancia de 5%, que el nivel de colesterol promedio se reduce como consecuencia del programa? Haga la prueba adecuada en las medias.

10.112 En un estudio que realiza el Departamento de Ingeniería Mecánica y que analiza el Centro de Consulta Estadística del Instituto Politécnico y Universidad Estatal de Virginia, se comparan las varillas de acero que proveen dos compañías diferentes. Se fabrican diez resortes de muestra con las varillas proporcionadas por

cada compañía y se estudia la "capacidad de rebote". Los datos son los siguientes:

```
Compañía A
 9.3 8.8 6.8  8.7  8.5 6.7  8.0  6.5  9.2 7.0
Compañía B
11.0 9.8 9.9 10.2 10.1 9.7 11.0 11.1 10.2 9.6
```

¿Puede concluir que casi no hay diferencia entre las varillas de acero proporcionadas por las dos compañías? Utilice un valor P para llegar a su conclusión. ¿Las varianzas deberían combinarse aquí?

10.113 En un estudio que conduce el Centro de Recursos Acuáticos y que analiza el Centro de Consulta Estadística del Instituto Politécnico y Universidad Estatal de Virginia, se comparan dos plantas de tratamiento para aguas residuales. La planta A se ubica donde el ingreso medio de los hogares está por abajo de $22,000 al año, y la planta B se ubica donde el ingreso medio de los hogares está por arriba de $60,000 anuales. La cantidad de agua residual que trata cada planta (miles de galones/día) se muestrea de forma aleatoria durante 10 días. Los datos son los siguientes:

```
Planta A:
21 19 20 23 22 28 32 19 13 18
Planta B:
20 39 24 33 30 28 30 22 33 24
```

¿Con un nivel de significancia de 5% podemos concluir que la cantidad promedio de agua residual tratada en el vecindario de altos ingresos es mayor que la del área de bajos ingresos?

10.114 Los siguientes datos muestran el número de defectos en 100,000 líneas de código en un tipo particular de software hecho en Estados Unidos y en Japón. ¿Hay suficiente evidencia para afirmar que existe una diferencia significativa entre los programas de los dos países? Pruebe las medias. ¿Deberían combinarse las varianzas?

```
E.U.    48 39 42 52 40 48 52 52
        54 48 52 55 43 46 48 52
Japón   50 48 42 40 43 48 50 46
        38 38 36 40 40 48 48 45
```

10.115 Estudios indican que la concentración de PCB es mucho más alta en tejido maligno de pecho que en tejido normal de pecho. Si un estudio de 50 mujeres con cáncer de seno revela una concentración promedio de PCB de 22.8×10^{-4} gramos, con una desviación estándar de 4.8×10^{-4} gramos, ¿la concentración media de PCB es menor que 24×10^{-4} gramos?

10.19 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

Una de las formas más sencillas de uso incorrecto de la estadística tiene que ver con la conclusión científica final que se obtiene cuando el analista no rechaza la hipótesis nula H_0. En esta obra intentamos aclarar lo que significan la hipótesis nula y la alternativa, así como en sentido amplio, la hipotesis alternativa es más importante. Como cuando, por ejemplo, el ingeniero intenta comparar dos calibradores y utiliza una prueba t de dos muestras, y H_0 es "los calibradores son equivalentes" mientras que H_1 es "los calibradores no son equivalentes", no rechazar H_0 no lleva a la conclusión de calibradores equivalentes". De hecho, puede darse el caso de que nunca se escriba o se diga "Accept H_0". El hecho de no rechazar H_0 tan sólo implica evidencia insuficiente. Dependiendo de la naturaleza de la hipótesis, no se descartan aun muchas posibilidades.

Como en el caso de los intervalos de confianza para muestras grandes que estudiamos en el capítulo 9, una prueba t con muestra grande que utiliza

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

con s que reemplaza σ es riesgoso utilizar para $n < 30$. Si $n \geq 30$ y la distribución no es normal sino que está algo cercana a la normal, se requiere el teorema del límite central y se confía en el hecho de que $n \geq 30$, $s \approx \sigma$.

Desde luego, cualquier prueba t va acompañada con la suposición concomitante de normalidad. Como en el caso de los intervalos de confianza, la prueba t es relativamente robusta para la normalidad. No obstante, incluso uno debería utilizar gráficas de probabilidad, pruebas del ajuste de bondad u otros procedimientos gráficos cuando la muestra no es demasiado pequeña.

Capítulo 11

Regresión lineal simple y correlación

11.1 Introducción a la regresión lineal

En la práctica, es frecuente que se requiera resolver problemas que implican conjuntos de variables de las cuales se sabe que tienen alguna relación inherente entre sí. Por ejemplo, en una situación industrial quizá se sepa que el contenido de alquitrán en la corriente de salida de un proceso químico está relacionado con la temperatura en la entrada. Podría ser de interés desarrollar un método de pronóstico, es decir, un procedimiento para estimar el contenido de alquitrán de varios combustibles de la temperatura de entrada, a partir de información experimental. Pero, por supuesto, es muy probable que para muchos ejemplos concretos en los que la temperatura de entrada sea la misma, por ejemplo 130 °C, el contenido de alquitrán a la salida no sea el mismo. Esto se parece mucho a lo que ocurre cuando se estudian varios automóviles con el mismo volumen en su motor. No todos recorrerán la misma distancia por unidad de gasolina. Si se consideraran viviendas en la misma parte del país que tuvieran la misma superficie habitable, no significaría que todas se venderían al mismo precio. El contenido de alquitrán, las millas por unidad de gasolina (mpg), y el precio de las casas (en miles de dólares) son **variables dependientes** naturales o respuestas en los tres escenarios. La temperatura en la entrada, el volumen del motor (pies cúbicos) y los pies cuadrados de área habitable son, respectivamente, **variables independientes** naturales o **regresores**. Una forma razonable de relación entre la **respuesta** $\boldsymbol{Y}$ y el regresor x es la relación lineal

$$Y = \alpha + \beta x,$$

donde, por supuesto, α es la **intersección** y β es la **pendiente**. La relación se ilustra en la figura 11.1.

Si la relación es exacta, entonces se trata de una **determinista** entre dos variables científicas, y no contiene ningún componente aleatorio o probabilístico. Sin embargo, en los ejemplos que se mencionaron, así como en muchos otros fenómenos científicos y de ingeniería, la relación no es determinista (es decir, una x dada no siempre produce el mismo valor de Y). Como resultado, existen problemas importantes que son de naturaleza probabilística, toda vez que la relación anterior no puede considerarse exacta. El concepto de **análisis de regresión** tiene que ver con

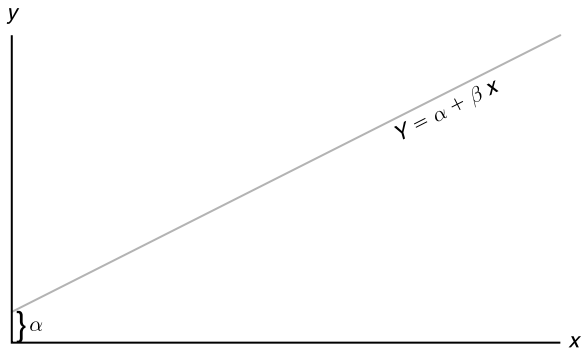


Figura 11.1: Una relación lineal.

encontrar la mejor relación entre Y y x, al cuantificar la intensidad de dicha relación y emplear métodos que permitan predecir los valores de la respuesta ante valores dados del regresor x.

En muchas aplicaciones, habrá más de un regresor (es decir, más de una variable independiente **que ayude a explicar a Y**). Por ejemplo, en el caso en que la respuesta es el precio de una casa, se esperaría que la edad de ésta contribuyera a la explicación del precio, por lo que en este caso la estructura múltiple de la regresión podría escribirse como

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2,$$

donde Y es el precio, x_1 son los pies cuadrados y x_2 es la edad en años. En el capítulo siguiente se estudiarán problemas con regresores múltiples. El análisis resultante se denomina **regresión múltiple**; en tanto que el análisis del caso con un solo regresor recibe el nombre de **regresión simple**. Un segundo ejemplo ilustrativo de la regresión múltiple sería el de un ingeniero químico que estudia la cantidad de hidrógeno perdido de las muestras de un metal específico que se tiene almacenado. En este caso habría dos entradas, el tiempo de almacenamiento x_1 en horas, y la temperatura de almacenamiento x_2 en grados centígrados. Entonces, la respuesta sería la pérdida de hidrógeno Y en partes por millón.

En este capítulo se estudia el tema de la **regresión lineal simple**, que trata el caso de una sola variable regresora. Para el caso de más de una variable regresora, el lector debe consultar el Capítulo 12. Sea una muestra aleatoria de tamaño n, denotada por el conjunto $\{(x_i, y_i); i = 1, 2, \ldots, n\}$. Si se tomaran muestras adicionales que tuvieran exactamente los mismos valores de x, se esperaría que los valores de y variaran. Así, el valor y_i de la pareja ordenada (x_i, y_i) es el valor de cierta variable aleatoria Y_i.

11.2 El modelo de regresión lineal simple

Hemos limitado el uso de los términos *análisis de regresión* a situaciones donde las relaciones entre las variables no son deterministas (esto es, no son exactas). En otras palabras, debe existir un **componente aleatorio** en la ecuación que relaciona las variables. Este componente aleatorio toma en cuenta consideraciones que no se miden, o que en realidad no son comprendidas por los científicos o los ingenieros. Es

seguro que en la mayoría de aplicaciones de la regresión, la ecuación lineal, digamos, $Y = \alpha + \beta x$ es una aproximación simplificada de algo desconocido y mucho más complejo. Por ejemplo, en nuestra ilustración que implica la respuesta $Y =$ contenido de alquitrán y $x =$ temperatura de entrada, es probable que $Y = \alpha + \beta x$ sea una aproximación razonable operativa dentro de un rango limitado de x. Se cumple, más que se infringe, el hecho de que los modelos que son simplificaciones de estructuras más complicadas y desconocidas son de naturaleza lineal (es decir, lineales en los **parámetros** α y β, o como en el caso del modelo que implica el precio, el tamaño y la edad de la casa, lineal en los **parámetros** α, β_1 y β_2). Estas estructuras lineales son sencillas y de naturaleza empírica, por lo que se denominan **modelos empíricos**.

Un análisis de la relación entre Y y x requiere el planteamiento de un **modelo estadístico**. Con frecuencia, un modelo es usado por un estadístico como representación de un **ideal** que, en esencia, define cómo percibimos que el sistema en cuestión generó los datos. El modelo debe incluir al conjunto $[(x_i, y_i); i = 1, 2, \ldots, n]$ de datos que implica n parejas de valores (x, y). Debe tenerse en cuenta que el valor de y_i depende de x_i por medio de una estructura lineal que también incluye el componente aleatorio. La base para el uso de un modelo estadístico relaciona la forma en que la variable aleatoria Y cambia con x y el componente aleatorio. El modelo también incluye las suposiciones acerca de las propiedades estadísticas del componente aleatorio. A continuación se da el modelo estadístico para la regresión lineal simple.

Modelo de regresión lineal simple

La respuesta Y se relaciona con la variable independiente x a través de la ecuación

$$Y = \alpha + \beta x + \epsilon.$$

En la cual α y β son los parámetros desconocidos de la intersección con el eje vertical y la pendiente, respectivamente, y ϵ es una variable aleatoria que se supone está distribuida con $E(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Es frecuente que a la cantidad σ^2 se le denomine varianza del error o varianza residual.

Del modelo anterior se hacen evidentes varias cuestiones. La cantidad Y es una variable aleatoria, ya que ϵ es aleatoria. El valor x de la variable regresora no es aleatorio y, de hecho, se mide con un error despreciable. La cantidad ϵ, que con frecuencia recibe el nombre de **error aleatorio** o **alteración aleatoria**, tiene varianza constante. Es frecuente que a esta parte de las suposiciones se le llame la suposición de varianza **homogénea**. La presencia de este error aleatorio, ϵ, impide que el modelo sea tan sólo una ecuación determinista. Ahora, el hecho de que $E(\epsilon) = 0$ implica que para una x específica los valores de y se distribuyen alrededor de la **recta verdadera** o **recta de regresión** de la población $y = \alpha + \beta x$. Si se elige bien el modelo, (esto es, no hay regresores adicionales de importancia y la aproximación lineal es buena dentro de los rangos de los datos), entonces son razonables los errores positivos y negativos alrededor de la regresión verdadera. Debe recordarse que en la práctica se desconocen α y β, y que deben estimarse a partir de los datos. Además, el modelo que se acaba de describir es de naturaleza conceptual. Como resultado, en la práctica nunca se observan los valores reales ϵ, por lo que nunca se puede trazar la verdadera recta de regresión (aunque se acepta que ahí está). Únicamente es posible dibujar una recta estimada. La figura 11.2 ilustra la naturaleza de los datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para un caso en que sólo se dispone de $n = 5$ observaciones. Debe destacarse que lo que observamos

en la figura 11.2 no es la recta que utilizan el científico o ingeniero. En vez de ello, ¡la ilustración únicamente describe el significado de las suposiciones! A continuación se describirá la regresión que el usuario tiene a su disposición.

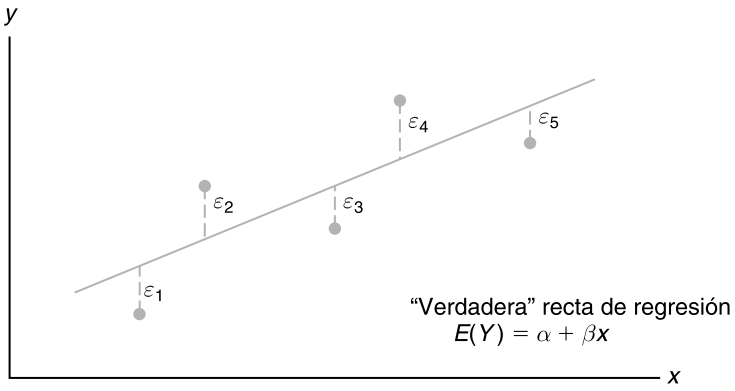


Figura 11.2: Datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para $n = 5$.

La recta de regresión ajustada

Un aspecto importante del análisis de regresión es, simplemente, estimar los parámetros α y β (es decir, estimar los llamados **coeficientes de regresión**). En la sección siguiente se estudiará el método para estimarlos. Suponga que los estimados de α y β se denotan con a y b, respectivamente. Entonces, la recta de **regresión ajustada**, o estimada, está dada por

$$\hat{y} = a + bx,$$

donde $\hat{y}$ es el valor pronosticado o ajustado. Es evidente que la recta ajustada es una estimación de la verdadera recta de regresión. Se espera que la recta ajustada esté más cerca de la verdadera línea de regresión cuando se disponga de una gran cantidad de datos. En el ejemplo siguiente se ilustra la recta ajustada para un estudio sobre contaminación en la vida real.

Uno de los problemas más desafiantes que se enfrentan en el área del control de la contaminación del agua lo representa la industria de la peletería. Los desechos de ésta tienen una complejidad química. Se caracterizan por valores elevados de demanda de oxígeno bioquímico, sólidos volátiles y otras medidas de la contaminación. Considere los datos experimentales de la tabla 11.1, que se obtuvo de 33 muestras de desechos tratados químicamente, en el estudio que se realizó en el Instituto Politécnico y Universidad Estatal de Virginia. Se registraron los valores de x, la reducción porcentual de los sólidos totales, y de y, el porcentaje de disminución de la demanda de oxígeno químico, para 33 muestras.

Los datos de la tabla 11.1 aparecen graficados en la figura 11.3, que es un **diagrama de dispersión**. Al inspeccionar dicho diagrama se observa que los puntos siguen de cerca una línea recta, lo cual indica que la suposición de linealidad entre las dos variables parece ser razonable.

Tabla 11.1: Medidas de los sólidos y la demanda de oxígeno químico

Reducción de sólidos x **(%)**	**Demanda de oxígeno químico,** y **(%)**	**Reducción de sólidos** x **(%)**	**Demanda de oxígeno químico,** y **(%)**
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

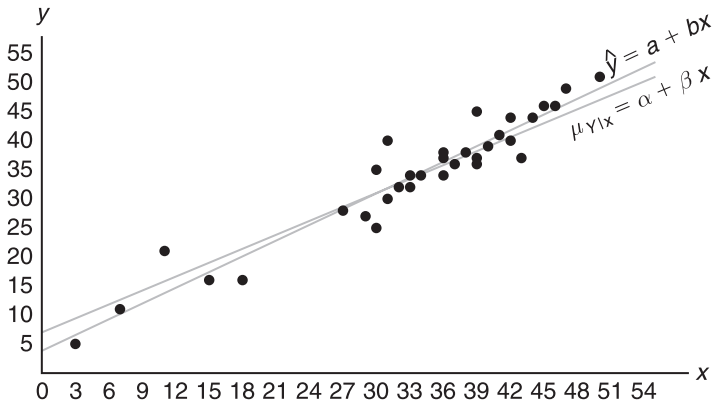


Figura 11.3: Diagrama de dispersión con rectas de regresión.

Con el diagrama de dispersión de la figura 11.3 se ilustran una verdadera recta hipotética de regresión y la recta de regresión ajustada. Este ejemplo se volverá a estudiar más adelante, en la sección 11.3, cuando se examine el método de estimación.

Otra mirada a las suposiciones del modelo

Resulta instructivo repasar el modelo de regresión lineal simple que se presentó con anterioridad, y analizar el sentido gráfico en que se relaciona con la denominada regresión verdadera. Se expandirá la figura 11.2 con la ilustración no sólo de dónde se localizan los ϵ_i en la gráfica, sino también lo que implica la suposición de normalidad de dichos ϵ_i.

Suponga que se tiene una regresión lineal simple con $n = 6$ valores de x equidistantes, y un valor único de y para cada x. Considere la gráfica de la figura 11.4. Esa ilustración debería dar al lector una representación clara del modelo y de las suposiciones implicadas. La recta que aparece en la gráfica es la de regresión verdadera. Los puntos son (y, x) reales dispersos alrededor de la recta. Cada punto tiene en sí mismo una distribución normal con el centro de la distribución (es decir, la media de y) sobre la recta. Ciertamente, esto es lo que se esperaba, ya que $E(Y) = \alpha + \beta x$. Como resultado, la verdadera recta de regresión **pasa a través de las medias de la respuesta**, y las observaciones reales se encuentran sobre la distribución alrededor de las medias. También observe que todas las distribuciones tienen la misma varianza, que se denota con σ^2. Por supuesto, la desviación entre una y individual y el punto sobre la recta será su valor individual de ϵ. Esto queda claro porque

$$y_i - E(Y_i) = y_i - (\alpha + \beta x_i) = \epsilon_i.$$

Así, en una x dada, tanto Y como el ϵ correspondiente tienen varianza σ^2.

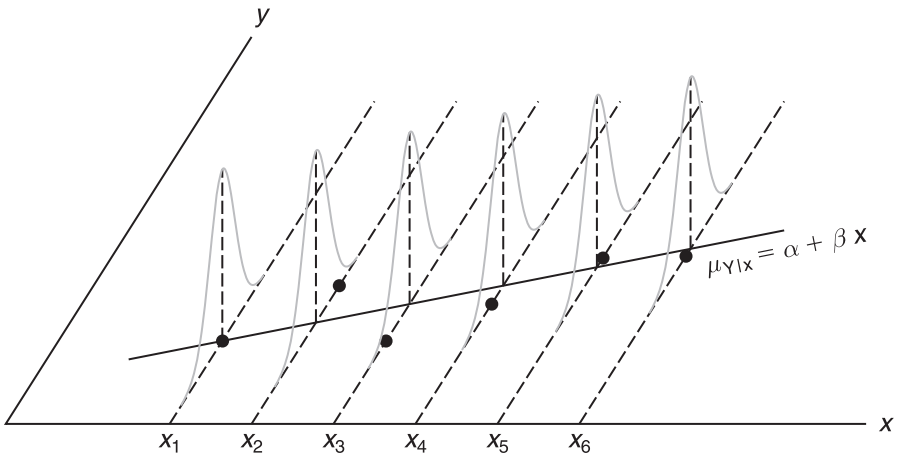


Figura 11.4: Observaciones individuales alrededor de la verdadera recta de regresión.

Note asimismo que aquí se ha escrito la verdadera recta de regresión como $\mu_{Y|x} = \alpha + \beta x$ con la finalidad de reafirmar que la recta pasa a través de la media de la variable aleatoria Y.

11.3 Los mínimos cuadrados y el modelo ajustado

En esta sección se estudia el método de ajustar una recta de regresión estimada a los datos, lo cual equivale a determinar las estimaciones a y b de α y de β, respectivamente. Por supuesto, esto permite el cálculo de los valores pronosticados a partir de la recta ajustada $\hat{y} = a + bx$, y hacer otros tipos de análisis y obtener otra información de diagnóstico que midan la intensidad de la relación y lo bien que se ajusta el modelo. Antes de estudiar el método de estimación de los mínimos cuadrados, resulta importante presentar el concepto de **residuo**. En esencia, un residuo es un error en el ajuste del modelo $\hat{y} = a + bx$.

Residuo: Error en el ajuste

Dado un conjunto de datos de regresión $[(x_i, y_i); i = 1, 2, \ldots, n]$ y un modelo ajustado $\hat{y}_i = a + bx_i$, el i-ésimo residuo e_i está dado por

$$e_i = y_i - \hat{y}_i, \qquad i = 1, 2, \ldots, n.$$

Es evidente que si un conjunto de n residuos es grande, entonces el ajuste del modelo no es bueno. Los residuos pequeños son una señal del buen ajuste. Otra relación interesante y que a veces es útil es la siguiente:

$$y_i = a + bx_i, + e_i.$$

El uso de la ecuación anterior debería dar como resultado la aclaración de la diferencia entre los residuos, e_i, y los errores del modelo conceptual, ϵ_i. El lector debe tener en cuenta que ϵ_i no son observados, y que e_i no sólo se observan sino que juegan un papel importante en el análisis general.

La figura 11.5 ilustra el ajuste de la recta a este conjunto de datos: $\hat{y} = a + bx$, y la recta que refleja el modelo $\mu_{Y|x} = \alpha + \beta x$. Por supuesto, ahora α y β son parámetros desconocidos. La recta ajustada es una estimación de la que genera el modelo estadístico. Hay que tener presente que la recta $\mu_{Y|x} = \alpha + \beta x$ es desconocida.

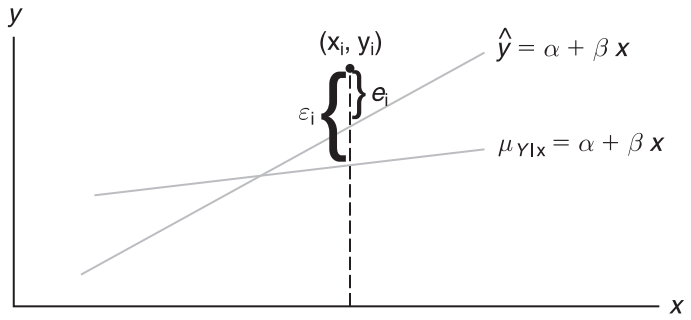


Figura 11.5: Comparación de ϵ_i con el residuo e_i.

Método de los mínimos cuadrados

Se deben encontrar los valores de a y b, estimadores de α y β, de manera que la suma de los cuadrados de los residuos sea mínima. La suma residual de los cuadrados con frecuencia se denomina suma de cuadrados de los errores respecto de la recta de regresión, y se denota como *sse*. Este procedimiento de minimización para estimar los parámetros se llama **método de los mínimos cuadrados**. Así, deben encontrarse a y b de modo que se minimice

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Al diferenciar sse con respecto a a y b, se obtiene

$$\frac{\partial(\text{SSE})}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i), \qquad \frac{\partial(\text{SSE})}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i.$$

Al igualar a cero las derivadas parciales y reacomodar los términos, obtenemos las ecuaciones siguientes (llamadas **ecuaciones normales**)

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

que se resuelven simultáneamente para obtener fórmulas de cálculo para a y b.

Estimación de los coeficientes de regresión

Dada la muestra $\{(x_i, y_i); i = 1, 2, \ldots, n\}$, los estimadores de mínimos cuadrados de a y b de los coeficientes de regresión α y β, se calculan mediante las fórmulas

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

y

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}.$$

En el ejemplo siguiente se ilustra el cálculo de a y b usando los datos de la tabla 11.1.

Ejemplo 11.1: Estime la recta de regresión para los datos de contaminación de la tabla 11.1.

Solución:

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41,355, \quad \sum_{i=1}^{33} x_i^2 = 41,086$$

Por lo tanto,

$$b = \frac{(33)(41,355) - (1104)(1124)}{(33)(41,086) - (1104)^2} = 0.903643,$$

y

$$a = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Así, la recta de regresión estimada está dada por

$$\hat{y} = 3.8296 + 0.9036x.$$

Con la recta de regresión del ejemplo 11.1 pronosticaríamos una reducción de 31% en la demanda de oxígeno químico cuando la reducción de los sólidos totales fuera de 30%. El 31% de reducción en la demanda de oxígeno químico puede interpretarse como una estimación de la media de la población $\mu_{Y|30}$, o como una estimación de una observación nueva en la que la reducción de sólidos totales es de 30%. Sin embargo, dichas estimaciones están sujetas a error. Aun cuando el experimento estuviera controlado de manera que la reducción de los sólidos totales fuera de 30%, es improbable que la reducción en la demanda de oxígeno químico que se midiera fuera exactamente igual a 31%. En realidad, los datos originales registrados en la tabla 11.1 indican que se registraron medidas de 25% y 35% en la reducción de la demanda de oxígeno, cuando la disminución de los sólidos totales era de 30%.

¿Qué es lo bueno de los mínimos cuadrados?

Debería observarse que el criterio de los mínimos cuadrados está diseñado para brindar una línea ajustada que resulte en la "cercanía" entre la recta y los puntos graficados. Existen muchas formas de medir dicha cercanía. Por ejemplo, quizá se deseara determinar los valores de a y b para los que $\sum_{i=1}^{n} |y_i - \hat{y}_i|$ es mínima, o para los que $\sum_{i=1}^{n} |y_i - \hat{y}_i|^{1.5}$ es mínima. Ambos métodos son viables y razonables. Observe que los dos, así como el procedimiento de los mínimos cuadrados, hacen que se fuerce a que los residuos sean "pequeños" en cierto sentido. Debe recordarse que los residuos son la contraparte empírica de los valores ϵ. La figura 11.6 ilustra un conjunto de residuos. Note que la línea ajustada tiene valores predichos como puntos sobre la recta y, por ello, los residuos son desviaciones verticales entre los puntos y la recta. Como resultado, el procedimiento de los mínimos cuadrados genera una recta que **minimiza la suma de los cuadrados de las desviaciones verticales** entre los puntos y la recta.

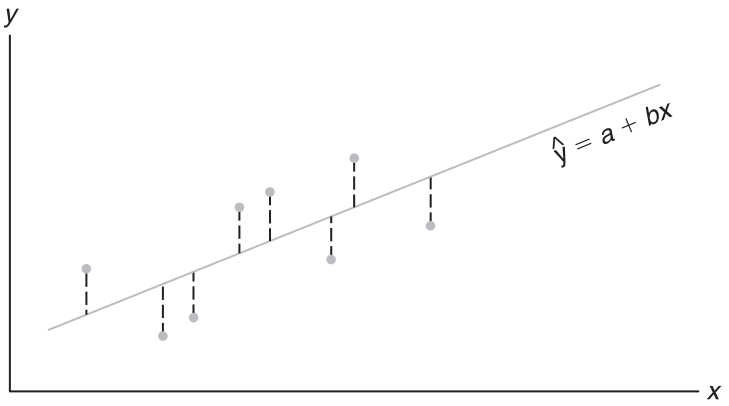


Figura 11.6: Los residuos como desviaciones verticales.

Ejercicios

11.1 Un estudio efectuado por VPI&SU para determinar si las mediciones estáticas de la fuerza de un brazo tienen influencia sobre las características de "levantamiento dinámico" de cierto individuo. Veinticinco individuos se sometieron a pruebas de fortaleza y luego se les pidió que hicieran una prueba de levantamiento de un peso, en el que éste se elevaba en forma dinámica por encima de la cabeza. A continuación se presentan los datos.

Individuo	Fuerza del brazo, x	Levantamiento dinámico, y
1	17.3	71.7
2	19.3	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.2	68.3
12	28.7	96.7

a) Estime los valores de α y β para la curva de regresión lineal $\mu_{Y|x} = \alpha + \beta x$.

b) Encuentre una estimación puntal de $\mu_{Y|30}$.

c) Grafique los residuos contra las X (fuerza del brazo). Comente los resultados.

Individuo	Fuerza del brazo, x	Levantamiento dinámico, y
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.7

11.2 Las siguientes son las calificaciones de un grupo de 9 estudiantes en un examen parcial (x) y en el examen final (y):

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

a) Estime la recta de regresión lineal.

b) Calcule la calificación final de un estudiante que obtuvo 85 en el examen parcial.

11.3 Se realizó un estudio sobre la cantidad de azúcar convertida, en cierto proceso, a distintas temperaturas. Los datos se codificaron y registraron como sigue:

Temperatura, x	Azúcar convertida, y
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

a) Estime la recta de regresión lineal.

b) Calcule la cantidad media de azúcar convertida que se produce cuando la temperatura registrada es 1.75.

c) Grafique los residuos contra la temperatura. Comente el resultado.

11.4 En cierto tipo de espécimen de prueba metálico, se sabe que la tensión normal sobre éste se relaciona de manera funcional con la resistencia al corte. Los siguientes son un conjunto de datos experimentales obtenidos para las dos variables:

Tensión normal, x	Resistencia al corte, y
26.8	26.5
25.4	27.3
28.9	24.2
23.6	27.1
27.7	23.6
23.9	25.9
24.7	26.3
28.1	22.5
26.9	21.7
27.4	21.4
22.6	25.8
25.6	24.9

a) Estime la recta de regresión $\mu_{Y|x} = \alpha + \beta x$.

b) Estime la resistencia al corte para una tensión normal de 24.5 kilogramos por centímetro cuadrado.

11.5 Se registraron las cantidades de un compuesto químico, y, que se disolvía en 100 gramos de agua a distintas temperaturas:

x (°C)	y (gramos)		
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

a) Encuentre la ecuación de la recta de regresión.

b) Grafique la recta en un diagrama de dispersión.

c) Estime la cantidad de producto químico que se disolverá en 100 gramos de agua a 50 °C.

11.6 Se aplicará un examen de colocación de matemáticas a todos los estudiantes de nuevo ingreso en una universidad pequeña. Se niega la inscripción al curso regular de matemáticas a los estudiantes que obtengan menos de 35, y se les envía a una clase remedial. Se registraron los resultados del examen de colocación, y las calificaciones finales de 20 estudiantes que tomaron el curso regular:

Examen de colocación	Calificación en curso
50	53
35	41
35	61
40	56
55	68
65	36
35	11
60	70
90	79
35	59
90	54
80	91
60	48
60	71
60	71
40	47
55	53
50	68
65	57
50	79

a) Elabore un diagrama de dispersión.

b) Encuentre la ecuación de la recta de regresión con la finalidad de predecir las calificaciones en el curso a partir de las del examen de colocación.

c) Grafique la recta en el diagrama de dispersión.

d) Si la calificación aprobatoria mínima es 60, ¿por debajo de cuál calificación en el examen de colocación debería negarse a los estudiantes futuros el derecho de admisión a ese curso?

11.7 Un comerciante al detalle realizó un estudio para determinar la relación que hay entre los gastos de la publicidad semanal y las ventas. Registró los datos siguientes:

Costos de publicidad ($)	**Ventas ($)**
40	385
20	400
25	395
20	365
30	475
50	440
40	490
20	420
50	560
40	525
25	480
50	510

a) Elabore un diagrama de dispersión.

b) Encuentre la ecuación de regresión para pronosticar las ventas semanales, a partir de los gastos en publicidad.

c) Estime las ventas semanales cuando los costos de la publicidad sean de $35.

d) Grafique los residuos contra los costos de publicidad. Haga comentarios.

11.8 Se recabaron los siguientes datos para determinar la relación entre la presión y la lectura correspondiente en la escala, para fines de calibración.

Presión, x (lb/pulg^2)	**Lectura de la escala, y**
10	13
10	18
10	16
10	15
10	20
50	86
50	90
50	88
50	88
50	92

a) Obtenga la ecuación de la recta de regresión.

b) En esta aplicación el propósito de la calibración es estimar la presión a partir de una lectura observada en la escala. Estime la presión para una lectura en la escala de 54, usando $\bar{x} = (54 - a)/b$.

11.9 Un estudio sobre la cantidad de lluvia y la de contaminación removida del aire produjo los siguientes datos:

Cantidad de lluvia diaria, x (0.01 cm)	**Partículas removidas, y (μg/m^3)**
4.3	126
4.5	121
5.9	116
5.6	118
6.1	114
5.2	118
3.8	132
2.1	141
7.5	108

a) Obtenga la ecuación de la recta de regresión para pronosticar las partículas removidas, a partir de la cantidad de lluvia diaria.

b) Estime la cantidad de partículas removidas cuando la lluvia diaria es $x = 4.8$ unidades.

11.10 Los siguientes datos son los precios de venta, z, de cierta marca y modelo de automóvil usado de w años de edad:

w (años)	**z (dólares)**
1	6350
2	5695
2	5750
3	5395
5	4985
5	4895

Ajuste una curva de la forma $\mu_{z|w} = \gamma\delta^w$ mediante la ecuación de regresión muestral no lineal $\hat{z} = cd^w$. [*Sugerencia:* Escriba $\ln \hat{z} = \ln c + (\ln d)w = a + bw$.]

11.11 El empuje de un motor (y) es función de la temperatura de escape (x) en °F, cuando otras variables de importancia se mantienen constantes. Considere los siguientes datos.

y	x	y	x
4300	1760	4010	1665
4650	1652	3810	1550
3200	1485	4500	1700
3150	1390	3008	1270
4950	1820		

a) Grafique los datos.

b) Ajuste una recta de regresión simple a los datos y grafíquela a través de ellos.

11.12 Se realizó un estudio para analizar el efecto de la temperatura ambiente, x, sobre la energía eléctrica consumida por una planta química, y. Se mantuvieron constantes otros factores y se recabaron los datos a partir de una planta piloto experimental.

a) Grafique los datos.

b) Estime la pendiente y la intersección en un modelo de regresión lineal simple.

c) Pronostique el consumo de energía para una temperatura ambiente de 65 °F.

y (BTU)	x (°F)	y (BTU)	x (°F)
250	27	265	31
285	45	298	60
320	72	267	34
295	58	321	74

11.13 Los siguientes datos son una parte de un conjunto clásico denominado "datos piloto de graficación", que aparecen en *Fitting Equations to Data*, de Daniel y Wood, publicado en 1971. La respuesta y es el contenido de ácido del material producido por valoración; mientras que el regresor x es el contenido de ácido orgánico producido por extracción y ponderación.

y	x	y	x
76	123	70	109
62	55	37	48
66	100	82	138
58	75	88	164
88	159	43	28

a) Grafique los datos; ¿la regresión lineal simple parece un modelo adecuado?

b) Haga un ajuste de regresión lineal simple; calcule la pendiente y la intersección.

c) Grafique la recta de regresión en la gráfica del inciso *a*).

11.14 Un profesor de la Escuela de Negocios de una universidad encuestó a una docena de sus colegas acerca del número de reuniones profesionales a que acudieron en los últimos cinco años (X), y el número de artículos que publicaron en revistas arbitradas (Y) durante el mismo periodo. A continuación se presenta el resumen de los datos:

$$n = 12, \quad \bar{x} = 4, \quad \bar{y} = 12,$$
$$\sum_{i=1}^{n} x_i^2 = 232, \quad \sum_{i=1}^{n} x_i y_i = 318.$$

Ajuste un modelo de regresión lineal simple entre x y y averiguando las estimaciones de la intersección y la pendiente. Comente acerca de si la asistencia a reuniones profesionales originaría una mayor cantidad de artículos.

11.4 Propiedades de los estimadores de los mínimos cuadrados

Además de los supuestos de que el término del error en el modelo

$$Y_i = \alpha + bx_i + \epsilon_i$$

es una variable aleatoria con media igual a cero y varianza σ^2 constante, suponga que además se acepta que $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ son independientes entre una ejecución y otra del experimento, lo cual brinda un fundamento para calcular las medias y varianzas de los estimadores de α y β.

Es importante recordar que nuestros valores de a y b, con base en una muestra dada de n observaciones, tan sólo son estimaciones de parámetros verdaderos α y β. Si se repite el experimento una y otra vez, usando en cada ocasión los mismos valores muestrales fijos de x, las estimaciones resultantes de α y β muy probablemente difieran de un experimento a otro. Estas estimaciones distintas pueden verse como valores adoptados por las variables aleatorias A y B; en tanto que a y b son realizaciones específicas.

Como los valores de x permanecen fijos, los valores de A y B dependen de las variaciones de los valores de y o, con más precisión, de los valores de las variables aleatorias $Y_1, Y_2, \ldots, Y_n$. Las suposiciones sobre la distribución implican que las Y_i, $i = 1, 2, \ldots, n$, también estén distribuidas con independencia, con media $\mu_{Y|x_i} = \alpha + \beta x_i$ y varianzas iguales σ^2; es decir,

$$\sigma^2_{Y|x_i} = \sigma^2 \quad \text{para} \quad i = 1, 2, \ldots, n.$$

Media y varianza de los estimadores

En la exposición que sigue se demuestra que el estimador B está insesgado para β, y se obtienen las varianzas tanto de A como de B. Esto inicia una serie de desarrollos

que llevan a la prueba de hipótesis y a la estimación de intervalos de confianza para la intersección y la pendiente.

Como el estimador

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

es de la forma $\sum_{i=1}^{n} c_i Y_i$, donde

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad i = 1, 2, \ldots, n,$$

y del corolario 4.4 se concluye que

$$\mu_B = E(B) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(Y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta,$$

y después, con el corolario 4.10,

$$\sigma_B^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_{Y_i}^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Puede demostrarse (ejercicio 11.15 en la página 412) que la variable aleatoria A tiene la media

$$\mu_A = \alpha \quad \text{y la varianza} \quad \sigma_A^2 = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2.$$

De estos resultados, es evidente que los **estimadores de mínimos cuadrados tanto para α como para β son insesgados**.

Partición de la variabilidad total y estimación de σ^2

Para hacer inferencias sobre α y β, es necesario llegar a una estimación del parámetro σ^2 que aparece en las dos fórmulas anteriores de la varianza de A y de B. El parámetro σ^2, el modelo de la varianza del error, refleja una variación aleatoria o variación del error experimental alrededor de la recta de regresión. En gran parte de lo que sigue se recomienda emplear la notación

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

De manera que la suma de los cuadrados de los errores puede escribirse así:

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2b\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + b^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= S_{yy} - 2bS_{xy} + b^2S_{xx} = S_{yy} - bS_{xy}, \end{aligned}$$

que es el paso final que surge del hecho de que $b = S_{xy}/S_{xx}$.

Teorema 11.1: Un estimador insesgado de σ^2 es

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}.$$

La prueba del teorema 11.1 se deja como ejercicio para el lector (consulte el ejercicio de repaso 11.61).

El estimador de σ^2 como error cuadrático medio

Con la finalidad de obtener cierta intuición sobre el estimador σ^2, hay que observar el resultado del teorema 11.1. El parámetro σ^2 mide la varianza o las desviaciones cuadradas entre los valores de Y y su media, dada por $\mu_{Y|x}$ (es decir, desviaciones cuadradas entre Y y $\alpha + \beta x$). Por supuesto, $\alpha + \beta x$ se estima con $\hat{y} = a + bx$. Así, tendría sentido que la varianza σ^2 quedara mejor descrita como la desviación cuadrada de una observación cualquiera, y_i, con respecto a la media estimada, $\hat{y}_i$, que es el punto correspondiente sobre la recta ajustada. Entonces, los valores $(y_i - \hat{y}_i)^2$ revelan la varianza apropiada, en forma muy parecida a como los valores $(y_i - \bar{y})^2$ miden la varianza cuando se muestrea en un escenario que no es de regresión. En otras palabras $\bar{y}$ estima la media en la última situación sencilla, en la cual $\hat{y}_i$ estima la media de y_i en una estructura de regresión. Ahora, ¿qué significa el divisor $n - 2$? En las secciones que siguen, se observará que éstos son los grados de libertad asociados con el estimador s^2 para σ^2. En el escenario i.i.d. normal estándar se resta de n un grado de libertad en el denominador. Y una explicación razonable es que se estima un parámetro, que es la media μ por medio de $\bar{y}$, pero en el problema de la regresión **se estiman dos parámetros**, que son α y β, con a y b. Así, el parámetro importante σ^2, que se estima mediante

$$s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2),$$

se denomina **error cuadrático medio**, e ilustra un tipo de media (división entre $n - 2$) de los residuos cuadrados.

11.5 Inferencias que conciernen a los coeficientes de regresión

Además de tan solo estimar la relación lineal entre x y Y para fines de predicción, el experimentador podría estar interesado en hacer ciertas inferencias acerca de la

pendiente y la intersección. Debe estarse dispuesto a hacer la suposición adicional de que cada ϵ_i, $i = 1, 2, \ldots, n$, tiene distribución normal, con la finalidad de permitir la prueba de hipótesis y la construcción de intervalos de confianza sobre α y β. Esta suposición implica que $Y_1, Y_2, \ldots, Y_n$ también están distribuidas en forma normal, cada una con una distribución de probabilidad $n(y_i; \alpha + \beta x_i, \sigma)$. Como A y B son funciones lineales de variables normales independientes, del teorema 7.11 se deduce que A y B tienen distribución normal con distribuciones de probabilidad $n(a; \alpha, \sigma_A)$ y $n(b; \beta, \sigma_A)$, respectivamente.

Se ve que la suposición de normalidad, un resultado mucho más análogo al dado en el teorema 8.4, permite concluir que $(n-2)S^2/\sigma^2$ es una variable chi-cuadrada con $n-2$ grados de libertad, independiente de la variable aleatoria B. Entonces, el teorema 8.5 garantiza que el estadístico

$$T = \frac{(B-\beta)/(\sigma/\sqrt{S_{xx}})}{S/\sigma} = \frac{B-\beta}{S/\sqrt{S_{xx}}}$$

tenga una distribución t con $n-2$ grados de libertad . El estadístico T se usa para construir un intervalo de confianza de $(1-\alpha)100\%$ para el coeficiente β.

Intervalo de confianza para β

Un intervalo de confianza de $100(1-\alpha)100\%$ para el parámetro β en la recta de regresión $\mu_{Y|x} = \alpha + \beta x$ es

$$b - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta < b + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n-2$ grados de libertad.

Ejemplo 11.2: Encuentre un intervalo de confianza de 95% para β en la recta de regresión $\mu_{Y|x} = \alpha + \beta x$, con base en los datos de contaminación de la tabla 11.1.

Solución: A partir de los resultados dados en el ejemplo 11.1, se determina que

$$S_{xx} = 4152.18, \qquad S_{xy} = 3752.09.$$

Además, se encuentra que $S_{yy} = 3713.88$. Recuerde que $b = 0.903643$. Entonces,

$$s^2 = \frac{S_{yy} - bS_{xy}}{n-2} = \frac{3713.88 - (0.903643)(3752.09)}{31} = 10.4299.$$

Por lo tanto, al sacar raíz cuadrada obtenemos $s = 3.2295$. Usando la tabla A.4, se encuentra que $t_{0.025} \approx 2.045$ para 31 grados de libertad. Así, un intervalo de confianza de 95% para β es

$$0.903643 - \frac{(2.045)(3.2295)}{\sqrt{4152.18}} < \beta < 0.903643 + \frac{(2.045)(3.2295)}{\sqrt{4152.18}},$$

que se simplifica a

$$0.8012 < \beta < 1.0061.$$

Prueba de hipótesis sobre la pendiente

Para probar la hipótesis nula H_0 de que $\beta = \beta_0$, contra una alternativa posible, utilizamos de nuevo la distribución t con $n - 2$ grados de libertad, con la finalidad de establecer una región crítica y después basar nuestra decisión sobre el valor de

$$t = \frac{b - \beta_0}{s/\sqrt{S_{xx}}}.$$

El método se ilustra con el ejemplo siguiente.

Ejemplo 11.3: Usando el valor estimado de $b = 0.903643$ del ejemplo 11.1, pruebe la hipótesis de que $\beta = 1.0$ contra la alternativa de que $\beta < 1.0$.

Solución: Las hipótesis son H_0: $\beta = 1.0$ y H_1: $\beta < 1.0$. Por lo tanto,

$$t = \frac{0.903643 - 1.0}{3.2295/\sqrt{4152.18}} = -1.92,$$

con $n - 2 = 31$ grados de libertad ($P \approx 0.03$).

Decisión: El valor t es significativo al nivel de 0.03, lo cual sugiere evidencia sólida de que $\beta < 1.0$.

Una prueba t importante sobre la pendiente es la prueba de hipótesis

$$H_0\text{: } \beta = 0,$$
$$H_1\text{: } \beta \neq 0.$$

Cuando no se rechaza la hipótesis nula, la conclusión es que no hay relación lineal significativa entre $E(y)$ y la variable independiente x. La gráfica de los datos del ejemplo 11.1 sugeriría que existe una relación lineal. Sin embargo, en ciertas aplicaciones en las que σ^2 es grande y por ende hay "ruido" considerable en los datos, una gráfica, aunque útil, quizá no produzca información clara para el investigador. El rechazo anterior de H_0 implica que hay una relación lineal significativa.

La figura 11.7 muestra una salida de *MINITAB* de la prueba t para

$$H_0\text{: } \beta = 0,$$
$$H_1\text{: } \beta \neq 0,$$

con los datos del ejemplo 11.1. Observe el coeficiente de regresión (Coef), el error estándar (Coef. SE), el valor t (T), y el valor P (P). Se rechaza la hipótesis nula. Es claro que existe una relación lineal significativa entre la demanda media del oxígeno químico y la reducción de los sólidos. Note que el estadístico t se calculó como

$$t = \frac{\text{coeficiente}}{\text{error estándar}} = \frac{b}{s/\sqrt{S_{xx}}}.$$

El no rechazar H_0: $\beta = 0$, sugiere que no hay una relación lineal entre Y y x. La figura 11.8 es una ilustración de la implicación de este resultado. Puede significar que los cambios de x tienen poco efecto sobre los cambios de Y, como se ve en el inciso a). Sin embargo, también puede indicar que la relación verdadera es no lineal, como se aprecia en b).

```
Regression Analysis: COD versus Per_Red
The regression equation is COD = 3.83 + 0.904 Per_Red

Predictor     Coef  SE Coef      T      P
Constant     3.830    1.768   2.17  0.038
Per_Red    0.90364  0.05012  18.03  0.000

S = 3.22954   R-Sq = 91.3%   R-Sq(adj) = 91.0%
Analysis of Variance
Source          DF      SS      MS       F      P
Regression       1  3390.6  3390.6  325.08  0.000
Residual Error  31   323.3    10.4
Total           32  3713.9
```

Figura 11.7: Salida de *MINITAB* de la prueba t para los datos del ejemplo 11.1.

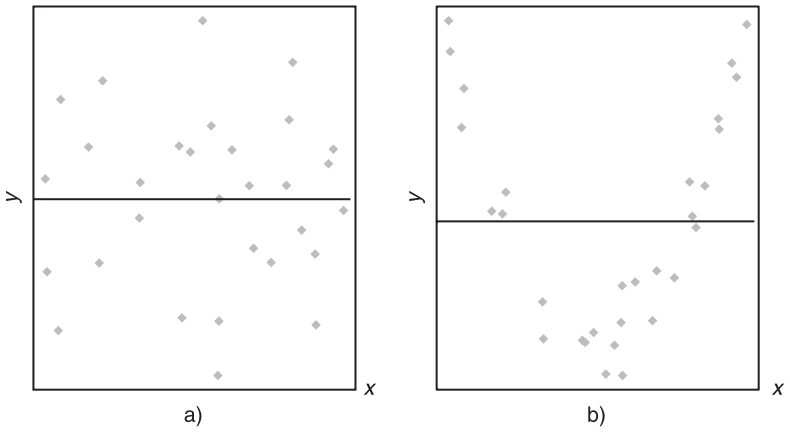


Figura 11.8: No se rechaza la hipótesis H_0: $\beta = 0$.

Cuando se rechaza H_0: $\beta = 0$, existe la implicación de que el término lineal en x que reside en el modelo explica una porción significativa de la variabilidad de Y. Las dos gráficas que aparecen en la figura 11.9 ilustran los escenarios posibles. Como se ilustra en el inciso a) de la figura, el rechazo sugiere que la relación es, en efecto, lineal. Como se ve en el inciso b), se sugiere que aunque el modelo no contenga un efecto lineal, se tendría una mejor representación si se incluye un término polinomial (tal vez cuadrático) (es decir, términos que complementen el término lineal).

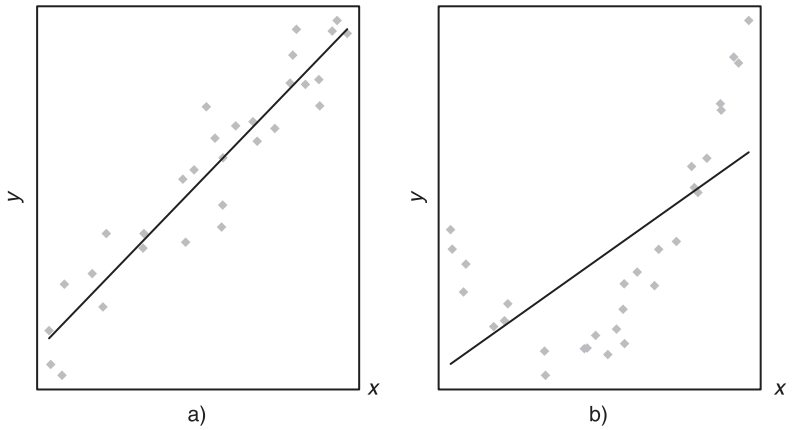


Figura 11.9: Se rechaza la hipótesis de que H_0: $\beta = 0$.

Inferencia estadística sobre la intersección

Los intervalos de confianza y la prueba de hipótesis sobre el coeficiente α pueden establecerse por el hecho de que A está distribuida en forma normal. No es difícil demostrar que

$$T = \frac{A - \alpha}{S\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}$$

tiene una distribución t con $n - 2$ grados de libertad, de manera que podemos construir un intervalo de confianza de $(1 - \alpha)100\%$ para α.

Intervalo de confianza para α

Un intervalo de confianza de $100(1 - \alpha)\%$ para el parámetro α en la recta de regresión $\mu_{Y|x} = \alpha + \beta x$ es

$$a - t_{\alpha/2}\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}} < \alpha < a + t_{\alpha/2}\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Observe que el símbolo α se utiliza aquí en dos formas sin relación alguna entre sí: primero como el nivel de significancia y, luego, como la intersección de la recta de regresión.

Ejemplo 11.4: Encuentre un intervalo de confianza de 95% para α en la recta de regresión $\mu_{Y|x} = \alpha + \beta x$, con base en los datos de la tabla 11.1.

Solución: En los ejemplos 11.1 y 11.2 se encontró que

$$S_{xx} = 4152.8 \qquad \text{y} \qquad s = 3.2295.$$

Del ejemplo 11.1 se tiene que

$$\sum_{i=1}^{n} x_i^2 = 41,086 \qquad \text{y} \qquad a = 3.829633.$$

Con el empleo de la tabla A.4, se encuentra que $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de confianza de 95% para α es

$$3.829633 - \frac{(2.045)(3.2295)\sqrt{41,086}}{\sqrt{(33)(4152.18)}} < \alpha < 3.829633 + \frac{(2.045)(3.2295)\sqrt{41,086}}{\sqrt{(33)(4152.18)}},$$

que se simplifica a $0.2132 < \alpha < 7.4461$.

Para probar la hipótesis nula H_0 de que $\alpha = \alpha_0$ contra una alternativa posible, utilizamos la distribución t con $n - 2$ grados de libertad para establecer una región crítica y, luego, basar la decisión sobre el valor de

$$t = \frac{a - \alpha_0}{s\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}.$$

Ejemplo 11.5: Usando el valor estimado de $\alpha = 3.829640$ del ejemplo 11.1, pruebe la hipótesis de que $\alpha = 0$ con un nivel de significancia de 0.05, contra la alternativa de que $\alpha \neq 0$.

Solución: Las hipótesis son H_0: $\alpha = 0$ y H_1: $\alpha \neq 0$. Por lo tanto,

$$t = \frac{3.829633 - 0}{3.2295\sqrt{41,086/((33)(4152.18))}} = 2.17,$$

con 31 grados de libertad. Así, $P =$ valor $P \approx 0.038$ y concluimos que $\alpha \neq 0$. Observe que esto tan sólo es Coef/StDev, como se aprecia en la salida de *MINITAB* de la figura 11.7. El SE Coef es el error estándar de la intersección estimada.

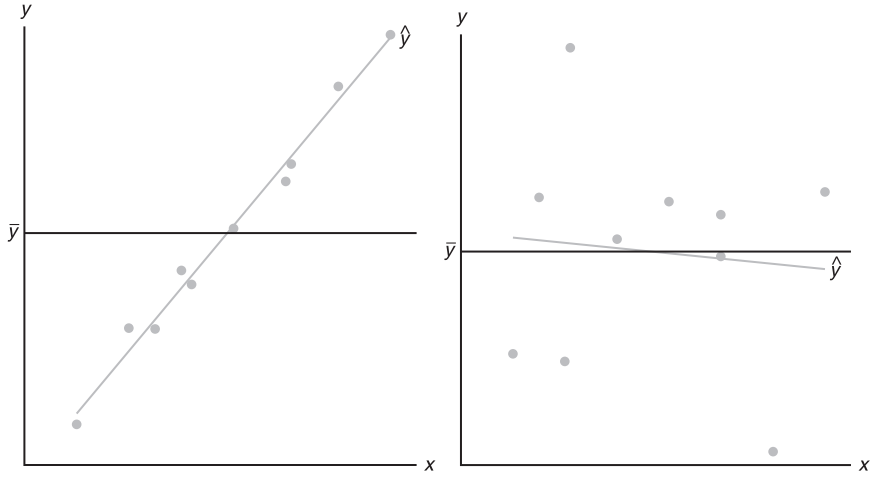
Una medida de la calidad del ajuste: el coeficiente de determinación

Observe el lector que en la figura 11.7 está dado un parámetro denotado con R-Sq, cuyo valor es 91.3%. Esta cantidad, R^2, se denomina **coeficiente de determinación** y es una medida de la **proporción de la variabilidad explicada por el modelo ajustado**. En la sección 11.8 se introducirá el concepto del enfoque del análisis de varianza, para la prueba de hipótesis en la regresión. El enfoque del análisis de varianza utiliza la suma cuadrática de los errores $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ y de la **suma total de los cuadrados corregida** $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$. Esta última representa la variación en los valores de respuesta que *idealmente* serían explicados con el modelo. El valor *SSE* es la variación debida al error, o **variación no explicada**. Resulta claro que si $SSE = 0$, toda variación queda explicada. La cantidad que representa la variación explicada es $SST - SSE$. R^2 es el

$$\text{Coeficiente de determinación: } R^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Note que si el ajuste es perfecto, *todos los residuos son cero*, y así $R^2 = 1.0$. Pero si *SSE* es tan sólo un poco menor que *SST*, $R^2 \approx 0$. Observe en la salida de la figura 11.7 que el coeficiente de determinación sugiere que el modelo ajustado a los datos explica el 91.3% de la variabilidad de la respuesta, la demanda de oxígeno químico.

Las gráficas que de la figura 11.10 brindan una ilustración de un buen ajuste ($R^2 \approx 1.0$) en *a*), y un ajuste deficiente ($R^2 \approx 0$) en *b*).



(a) $R^2 \approx 1.0$ (b) $R^2 \approx 0$

Figura 11.10: Gráficas que ilustran un ajuste muy bueno y otro deficiente.

Errores en el uso de R^2

Los analistas citan con mucha frecuencia los valores de R^2, quizá debido a su simplicidad. Sin embargo, hay errores en su interpretación. La confiabilidad de R^2 es función del tamaño del conjunto de los datos de la regresión y del tipo de aplicación. Es claro que $0 \leq R^2 \leq 1$, y el límite superior se alcanza cuando el ajuste a los datos es perfecto (es decir, todos los residuos son cero). ¿Cuál es un valor aceptable de R^2? Se trata de una pregunta difícil de contestar. Es seguro que un químico que tratara de establecer una calibración lineal de una pieza de equipo de alta precisión, por experiencia, esperaría un valor muy alto de R^2 (quizá superior a 0.99); mientras que un científico del comportamiento, que trabaja con datos afectados por la variabilidad del comportamiento humano, quizá se sentiría afortunado si experimentara un valor de R^2 tan grande como 0.70. Un individuo con pericia en el ajuste de modelos tiene la sensibilidad para saber cuándo un valor es suficientemente grande, dada la situación que enfrente. Es claro que algunos fenómenos científicos llevan por sí mismos a modelar con mayor precisión que otros.

El criterio de R^2 es peligroso de utilizar al comparar *modelos en competencia* para el mismo conjunto de datos. Cuando se agregan términos adicionales al modelo (como un regresor más), disminuye *SSE* y con ello se incrementa R^2 (al menos no disminuye), lo cual implica que R^2 puede hacerse artificialmente alto con la práctica inapropiada de **sobreajustar** (es decir, incluir demasiados términos en el modelo). Así, el incremento inevitable de R^2 al agregar un términos adicionales no implica

que éstos fueran necesarios. En realidad, para predecir los valores de la respuesta el modelo simple puede ser superior. En el capítulo 12 se estudiará con detalle el papel del sobreajuste y su influencia sobre la capacidad de predicción, cuando se vea el concepto de los modelos que implican **más de un solo regresor**. En este momento baste decir que *para seleccionar un modelo no se debe suscribir un proceso de selección que únicamente incluya la consideración de* R^2.

11.6 Predicción

Hay varias razones para construir un modelo de regresión lineal. Una de ellas es, desde luego, predecir valores de respuesta para uno o más valores de la variable independiente. Esta sección se centra en los errores asociados con la predicción.

La ecuación $\hat{y} = a + bx$ puede utilizarse para predecir o estimar la **respuesta media** $\mu_{Y|x_0}$ en $x = x_0$, donde x_0 no necesariamente es uno de los valores preestablecidos, o puede emplearse para pronosticar un solo valor y_0 de la variable Y_0, cuando $x = x_0$. Se esperaría que el error de predicción fuera mayor para el caso de un solo valor pronosticado, que para aquel en que se predice una media. Entonces, esto afectaría el ancho de los intervalos para los valores que se predicen.

Suponga el lector que el experimentador desea construir un intervalo de confianza para $\mu_{Y|x_0}$. Se debe usar el estimador puntual $\hat{Y}_0 = A + Bx_0$ para estimar $\mu_{Y|x_0} = \alpha + \beta x$. Puede demostrarse que la distribución muestral de $\hat{Y}_0$ es normal con media

$$\mu_{Y|x_0} = E(\hat{Y}_0) = E(A + Bx_0) = \alpha + \beta x_0 = \mu_{Y|x_0}$$

y varianza

$$\sigma^2_{\hat{Y}_0} = \sigma^2_{A+Bx_0} = \sigma^2_{\bar{Y}+B(x_0-\bar{x})} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right],$$

esta última surge del hecho de que $Cov(\bar{Y}, B) = 0$ (véase el ejercicio 11.6 en la página 412). Así, ahora es posible construir un intervalo de confianza de $(1 - \alpha)100\%$ sobre la respuesta media $\mu_{Y|x_0}$ a partir del estadístico

$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S\sqrt{1/n + (x_0 - \bar{x})^2/S_{xx}}},$$

que tiene distribución t con $n - 2$ grados de libertad.

Intervalo de confianza para $\mu_{Y|x_0}$

Un intervalo de confianza de $(1 - \alpha)100\%$ para la respuesta media $\mu_{Y|x_0}$ es

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Ejemplo 11.6: Con los datos de la tabla 11.1, construya límites de confianza de 95% para la respuesta media $\mu_{Y|x_0}$.

Solución: De la ecuación de regresión encontramos que para $x_0 = 20\%$ de reducción de sólidos, digamos,

$$\hat{y}_0 = 3.829633 + (0.903643)(20) = 21.9025.$$

Además, $\bar{x} = 33.4545$, $S_{xx} = 4152.18$, $s = 3.2295$ y $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de confianza de 95% para $\mu_{Y|20}$ es

$$21.9025 - (2.045)(3.2295)\sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < \mu_{Y|20}$$
$$< 21.9025 + (2.045)(3.2295)\sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}},$$

o, simplemente, $20.1071 < \mu_{Y|20} < 23.6979$.

Al repetir los cálculos anteriores para cada uno de los diferentes valores de x_0, se obtienen los límites de confianza correspondientes para cada $\mu_{Y|x_0}$. En la figura 11.11 se muestran los datos de los puntos, la recta de regresión estimada y los límites de confianza superior e inferior sobre la media de $Y|x$.

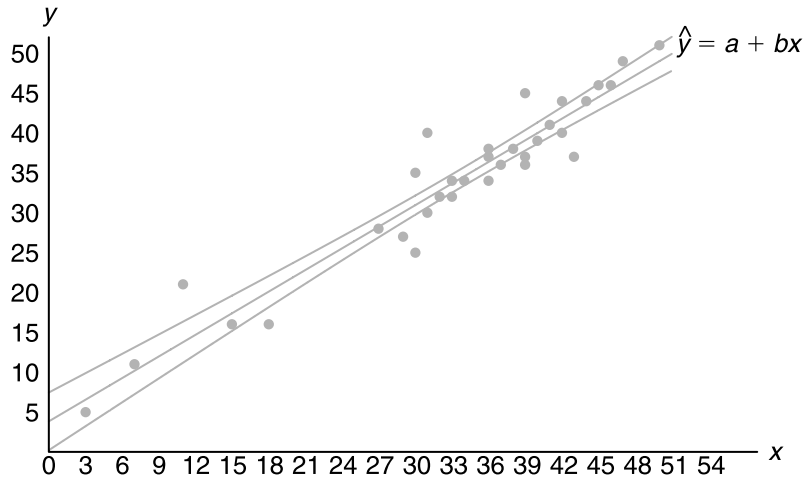


Figura 11.11: Límites de confianza para el valor medio de $Y|x$.

En el ejemplo 11.6 se tiene el 95% de confianza en que la demanda de oxígeno químico de la población estará entre 20.1071% y el 23.6979%, cuando la reducción de sólidos sea de 20%.

Predicción del intervalo

Otro tipo de intervalo que con frecuencia se malinterpreta y se confunde con aquel dado para $\mu_{Y|x}$ es el intervalo de la predicción para una respuesta futura observada. En realidad, en muchos casos el intervalo de la predicción es más relevante para el científico o ingeniero, que el intervalo de confianza sobre la media. En el ejemplo del contenido de alquitrán y la temperatura de entrada, mencionado en la sección 11.1, sería de interés no sólo estimar la media del contenido de alquitrán a una temperatura específica, sino también en la construcción de un intervalo que refleje el error

en la predicción de una cantidad futura observada del contenido de alquitrán a la temperatura dada.

Para obtener un **intervalo de predicción** para cualquier valor único y_0 de la variable Y_0, es necesario estimar la varianza de las diferencias entre las ordenadas $\hat{y}_0$, obtenidas de las rectas de regresión calculadas en el muestreo repetido cuando $x = x_0$, y la ordenada verdadera correspondiente y_0. Se puede pensar en la diferencia $\hat{y}_0 - y_0$ como un valor de la variable aleatoria $\hat{Y}_0 - Y_0$, cuya distribución muestral se demuestre que sea normal con media

$$\mu_{\hat{Y}_0 - Y_0} = E(\hat{Y}_0 - Y_0) = E[A + Bx_0 - (\alpha + \beta x_0 + \epsilon_0)] = 0$$

y varianza

$$\sigma^2_{\hat{Y}_0 - Y_0} = \sigma^2_{A + Bx_0 - \epsilon_0} = \sigma^2_{\bar{Y} + B(x_0 - \bar{x}) - \epsilon_0} = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right].$$

Así, un intervalo de predicción de $(1 - \alpha)100\%$ para un solo valor pronosticado y_0 puede construirse a partir del estadístico

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{xx}}},$$

que tiene una distribución t con $n - 2$ grados de libertad.

Intervalo de predicción para y_0

Un intervalo de predicción de $(1 - \alpha)100\%$ para una sola respuesta y_0 está dado por

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Es claro que hay una diferencia entre el concepto de un intervalo de confianza y el del intervalo de predicción antes descrito. La interpretación del intervalo de confianza es idéntica a la que se describió para todos los intervalos de confianza sobre los parámetros de la población estudiados en el libro. En verdad, $\mu_{Y|x_0}$ es un parámetro de la población. Sin embargo, el intervalo de la predicción calculado representa un intervalo que tiene una probabilidad igual a $1 - \alpha$ de contener no un parámetro sino un valor futuro de y_0 de la variable aleatoria Y_0.

Ejemplo 11.7: Con los datos de la tabla 11.1, construya un intervalo de predicción de 95% para y_0 cuando $x_0 = 20\%$.

Solución: Tenemos que $n = 33$, $x_0 = 20$, $\bar{x} = 33.4545$, $\hat{y}_0 = 21.9025$, $S_{xx} = 4252.18$, $s = 3.2295$ y $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de predicción de 95% para y_0 es

$$21.9025 - (2.045)(3.2295)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < y_0$$
$$< 21.9025 + (2.045)(3.2295)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}},$$

que se simplifica para $15.0585 < y_0 < 28.7464$.

La figura 11.12 muestra otra gráfica de los datos de la demanda de oxígeno químico, con los intervalos de confianza de la respuesta media y con el intervalo de predicción sobre una respuesta individual graficada. En el caso de la respuesta media, la gráfica refleja un intervalo mucho más angosto alrededor de la recta de regresión.

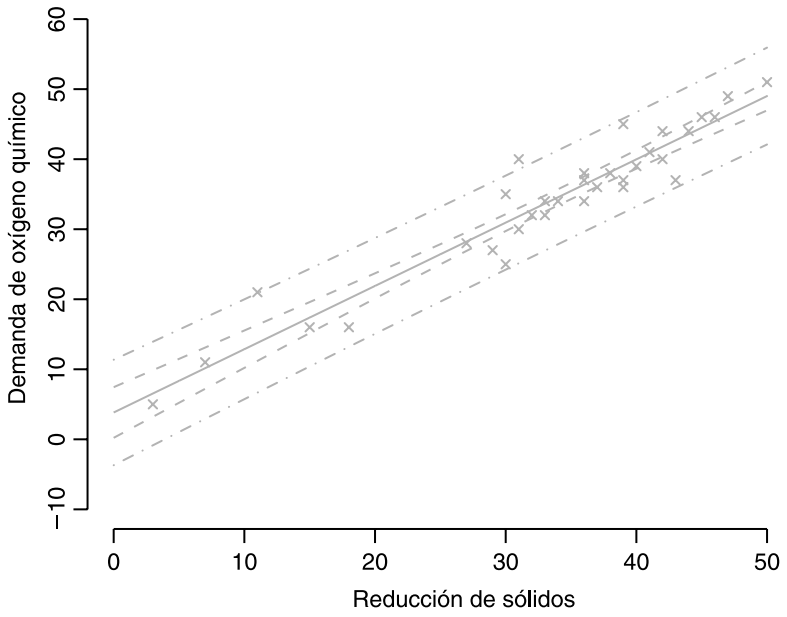


Figura 11.12: Intervalos de confianza y predicción para los datos de la demanda de oxígeno químico; las bandas interiores indican los límites de confianza para las respuesta medias, y las exteriores señalan los límites de predicción para las respuestas futuras.

Ejercicios

11.15 Suponga que los ϵ_i son normales, independientes, con media igual a cero y varianza común σ^2, demuestre que A, el estimador de mínimos cuadrados de α en $\mu_{Y|x} = \alpha + \beta x$ tiene distribución normal con media α y varianza

$$\sigma_A^2 = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2.$$

11.16 Para un modelo de regresión lineal simple

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

donde los ϵ_i son independientes y tienen distribución normal, con medias iguales a cero y varianzas σ^2 iguales, demuestre que $\hat{Y}$ y

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

tienen covarianza de cero.

11.17 En relación con el ejercicio 11.1, de la página 397,

a) evalúe s^2;

b) pruebe la hipótesis de que $\beta = 0$ contra la alternativa de que $\beta \neq 0$ con un nivel de significancia de 0.05, e interprete la decisión resultante.

11.18 En relación con el ejercicio 11.2 de la página 398,

a) evalúe s^2;

b) construya un intervalo de confianza de 95% para α;

c) construya un intervalo de confianza de 95% para β.

11.19 Con referencia al ejercicio 11.3 de la página 398,

a) evalúe s^2;

b) construya un intervalo de confianza de 95% para α;

c) construya un intervalo de confianza de 95% para β.

11.20 En relación con el ejercicio 11.4 de la página 398,

a) evalúe s^2;

b) construya un intervalo de confianza de 99% para α;

c) construya un intervalo de confianza de 99% para β.

11.21 Para el ejercicio 11.5 de la página 398,

a) evalúe s^2;

b) construya un intervalo de confianza de 99% para α;

c) construya un intervalo de confianza de 99% para β.

11.22 Pruebe la hipótesis de que $\alpha = 10$ en el ejercicio 11.6 de la página 398, contra la alternativa de que $\alpha < 10$. Utilice un nivel de significancia de 0.05.

11.23 Pruebe la hipótesis de que $\beta = 6$ en el ejercicio 11.7 de la página 399, contra la alternativa de que $\beta < 6$. Utilice un nivel de significancia de 0.025.

11.24 Con el valor de s^2 que se halló en el ejercicio 11.18*a*), construya un intervalo de confianza de 95% para $\mu_{Y|85}$ en el ejercicio 11.2 de la página 398.

11.25 En relación con el ejercicio 11.4 de la página 398, utilice el valor de s^2 que encontró en el ejercicio 11.20*a*) para calcular

a) un intervalo de confianza de 95% para la resistencia media al corte cuando $x = 24.5$;

b) un intervalo de predicción de 95% para un solo valor pronosticado de la resistencia al corte cuando $x = 24.5$.

11.26 Utilizando el valor de s^2 que se halló en el ejercicio 11.19*a*), grafique la recta de regresión y las bandas de confianza de 95% para la respuesta media $\mu_{Y|x}$ con los datos del ejercicio 11.3 en la página 398.

11.27 Con el valor de s^2 que se obtuvo en el ejercicio 11.19*a*), construya un intervalo de confianza de 95% para la cantidad de azúcar convertida correspondiente a $x = 1.6$, en el ejercicio 11.3 de la página 398.

11.28 En relación con el ejercicio 11.5 de la página 398, utilice el valor de s^2 que se obtuvo en el ejercicio 11.21*a*) para calcular

a) un intervalo de confianza de 99% para la cantidad promedio del producto químico que se disolverá en 100 gramos de agua a 50 °C;

b) un intervalo de predicción de 99% para la cantidad de producto químico que se disolverá en 100 gramos de agua a 50 °C.

11.29 Considere la regresión del número de millas para ciertos automóviles, en millas por galón (mpg) y su peso en libras (wt). Los datos son del *Consumer Reports* (abril de 1997). En la figura 11.13 se presenta parte de la salida del *SAS* para el procedimiento.

a) Estime las millas para un vehículo que pesa 4000 libras.

b) Suponga que los ingenieros de Honda afirman que, en promedio, el Civic (o cualquier otro modelo de vehículo que pese 2440 libras) recorre más de 30 mpg. Con base en los resultados del análisis de regresión, ¿cree el lector dicha afirmación? ¿Por qué?

c) Los ingenieros de diseño para el Lexus ES300 tienen por objetivo lograr 18 mpg como el ideal para dicho modelo (o cualquier otro que pese 3390 libras), aunque se espera que haya cierta variación. ¿Es probable que sea realista ese objetivo de valor? Comente al respecto.

11.30 Demuestre que para el caso del ajuste por mínimos cuadrados para el modelo de regresión lineal simple

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

que $\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$.

11.31 Considere la situación del ejercicio 11.30; pero suponga que $n = 2$ (es decir, tan sólo se dispone de dos puntos de los datos). Proporcione un argumento sobre que la recta de regresión por mínimos cuadrados dará como resultado $(y_1 - \hat{y}_1) = (y_2 - \hat{y}_2) = 0$. También demuestre que en ese caso $R^2 = 1.0$.

11.32 Existen aplicaciones importantes en las que debido a restricciones científicas conocidas, la recta de regresión **debe pasar por el origen** (es decir, la intersección debe estar en el cero). En otras palabras, el modelo debe ser

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

y tan sólo se requiere estimar un parámetro. Con frecuencia, dicho modelo se denomina **modelo de regresión por el origen**.

a) Demuestre que el estimador de mínimos cuadrados para la pendiente es

$$b = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i=1}^{n} x_i^2\right).$$

b) Demuestre que $\sigma_B^2 = \sigma^2 / \left(\sum_{i=1}^{n} x_i^2\right)$.

c) Demuestre que b del inciso *a*) es un estimador insesgado para β. Es decir, demuestre que $E(B) = \beta$.

11.33 Dado el conjunto de datos

y	x	y	x
7	2	40	10
50	15	70	20
100	30		

a) Grafique los datos.

b) Ajuste una recta de regresión "por el origen".

c) Grafique la recta de regresión sobre la gráfica de los datos.

d) Dé una fórmula general (en términos de las y_i y la pendiente b) para el estimador de σ^2.

e) Para este caso, dé una fórmula para $Var(\hat{y})$; $i = 1, 2, \ldots, n$.

f) Grafique los límites de confianza de 95% para la respuesta media sobre la gráfica alrededor de la recta de regresión.

11.34 Para los datos del ejercicio 11.33, encuentre un intervalo de predicción de 95% en $x = 25$.

Root MSE 1.48794 R-Square 0.9509
Dependent Mean 21.50000 Adj R-Sq 0.9447

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	44.78018	1.92919	23.21	<.0001
WT	1	-0.00686	0.00055133	-12.44	<.0001

MODEL	WT	MPG	Predict	LMean	UMean	Lpred	Upred	Residual
GMC	4520	15	13.7720	11.9752	15.5688	9.8988	17.6451	1.22804
Geo	2065	29	30.6138	28.6063	32.6213	26.6385	34.5891	-1.61381
Honda	2440	31	28.0412	26.4143	29.6681	24.2439	31.8386	2.95877
Hyundai	2290	28	29.0703	27.2967	30.8438	25.2078	32.9327	-1.07026
Infiniti	3195	23	22.8618	21.7478	23.9758	19.2543	26.4693	0.13825
Isuzu	3480	21	20.9066	19.8160	21.9972	17.3062	24.5069	0.09341
Jeep	4090	15	16.7219	15.3213	18.1224	13.0158	20.4279	-1.72185
Land	4535	13	13.6691	11.8570	15.4811	9.7888	17.5493	-0.66905
Lexus	3390	22	21.5240	20.4390	22.6091	17.9253	25.1227	0.47599
Lincoln	3930	18	17.8195	16.5379	19.1011	14.1568	21.4822	0.18051

Figura 11.13: Salida del *SAS* para el ejercicio 11.29.

11.7 Selección de un modelo de regresión

Gran parte de lo que se ha presentado hasta aquí acerca de la regresión que involucra una sola variable independiente depende de la suposición de que el modelo elegido es correcto, la presunción de que $\mu_{Y|x}$ se relaciona con x linealmente en los parámetros. Es cierto que no se esperaría que la predicción de la respuesta fuera buena si hubiera diversas variables independientes que no se consideraran en el modelo, que afectaran la respuesta y variaran en el sistema. Además, la predicción seguramente sería inadecuada si la estructura verdadera que relacionara $\mu_{Y|x}$ con x fuera no lineal en extremo en el rango de las variables consideradas.

Es frecuente que el modelo de regresión lineal simple se utilice aun cuando se sepa que el modelo es algo distinto del lineal, o que se desconozca la estructura verdadera. Este enfoque con frecuencia es muy bueno, en particular cuando el rango de las x es estrecho. Entonces, el modelo que se utiliza se vuelve una función aproximadora, de la cual se espera sea una representación adecuada del panorama verdadero en la región de interés. Sin embargo, debe notarse el efecto que tendría un modelo inadecuado sobre los resultados presentados hasta este momento. Por ejemplo, si el modelo verdadero, desconocido para el experimentador, es lineal en más de una x, digamos,

$$\mu_{Y|x_1,x_2} = \alpha + \beta_1 x_1 + \beta_2 x_2,$$

entonces, el estimador $b = S_{xy}/S_{xx}$, de los mínimos cuadrados ordinarios calculado considerando tan sólo x_1 en el experimento es, en circunstancias generales, un estimado sesgado del coeficiente β_1 (véase el ejercicio 11.37 en la página 423). Asimismo, el estimador s^2 para σ^2 está sesgado debido a la variable adicional.

11.8 El enfoque del análisis de varianza

Con frecuencia, el problema de analizar la calidad de la recta de regresión estimada se maneja mediante el enfoque del **análisis de varianza** (ANOVA): procedimiento en el que la variación total de la variable dependiente se subdivide en componentes significativos, que luego se observan y se tratan en forma sistemática. El análisis de varianza, que se estudia en el capítulo 13, es un recurso poderoso que se emplea en muchas aplicaciones.

Suponga el lector que se tiene n puntos de datos experimentales en la forma usual (x_i, y_i), y que se obtiene la recta de regresión. En la sección 11.4 para la estimación de σ^2 se estableció la identidad

$$S_{yy} = bS_{xy} + SSE.$$

Una formulación alternativa y quizá más informativa es la siguiente:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

De modo que se hizo una partición de la **suma cuadrática corregida total de y** en dos componentes que deberían reflejar un significado particular para el experimentador. Esta partición se debería indicar en forma simbólica como

$$SST = SSR + SSE.$$

El primer componente de la derecha, SSR, se denomina **suma cuadrática de la regresión**, y refleja la cantidad de variación de los valores y que se **explica con el modelo**, que en este caso es la línea recta postulada. El segundo componente es la familiar suma de errores al cuadrado, que refleja la variación alrededor de la recta de regresión.

Suponga el lector que hay interés en las hipótesis

$$H_0\text{: } \beta = 0,$$
$$H_1\text{: } \beta \neq 0.$$

donde la hipótesis nula en esencia dice que el modelo es $\mu_{Y|x} = \alpha$. Es decir, la variación en los resultados Y debido a las fluctuaciones de probabilidad o aleatorias que son independientes de los valores de x. Esta condición se refleja en la figura 11.10*b*). En las condiciones de esta hipótesis nula se puede demostrar que SSR/σ^2, y SSE/σ^2 son valores de variables chi-cuadradas independientes con 1 y $n-2$ grados de libertad, respectivamente, por lo que según el teorema 7.12 se sigue que SST/σ^2 también es un valor de una variable chi-cuadrada con $n-1$ grados de libertad. Para probar la hipótesis anterior, calculamos

$$f = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{s^2}$$

y se rechaza H_0 al nivel de significancia α cuando $f > f_\alpha(1, n-2)$.

Por lo general, los cálculos se resumen mediante una **tabla de análisis de varianza**, como se indica en la tabla 11.2. Es costumbre referirse a las distintas sumas de los cuadrados divididos entre sus grados respectivos de libertad como **medias cuadráticas**.

Tabla 11.2: Análisis de varianza para la prueba de $\beta = 0$

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
Regresión	SSR	1	SSR	$\frac{SSR}{s^2}$
Error	SSE	$n-2$	$s^2 = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

Cuando se rechaza la hipótesis nula, es decir, cuando el estadístico F calculado excede el valor crítico $f_\alpha(1, n-2)$, concluimos que **hay una cantidad significativa de variación en la respuesta que da el modelo postulado, el cual es la función de una línea recta**. Si el estadístico F está en la región de rechazo, se concluye que los datos no reflejan evidencia suficiente para apoyar el modelo que se postula.

En la sección 11.5 se da un procedimiento donde se usa el estadístico

$$T = \frac{B - \beta_0}{S/\sqrt{S_{xx}}}$$

para probar la hipótesis

$$H_0\text{: } \beta = \beta_0,$$
$$H_1\text{: } \beta \neq \beta_0,$$

donde T sigue la distribución t con $n-2$ grados de libertad. La hipótesis se rechaza si $|t| > t_{\alpha/2}$, para un nivel de significancia de α. Es interesante observar que en el caso especial en que se prueba

$$H_0\text{: } \beta = 0,$$
$$H_1\text{: } \beta \neq 0,$$

el valor del estadístico T se convierte en

$$t = \frac{b}{s/\sqrt{S_{xx}}},$$

y la hipótesis en consideración es idéntica a la que se prueba en la tabla 11.2. Sobre todo, la hipótesis nula establece que la variación en la respuesta se debe tan sólo a la aleatoriedad. El análisis de varianza utiliza la distribución F en vez de la t. Para la alternativa bilateral, ambos enfoques son idénticos. Esto se observa si se escribe

$$t^2 = \frac{b^2 S_{xx}}{s^2} = \frac{bS_{xy}}{s^2} = \frac{SSR}{s^2},$$

que es idéntico al valor f utilizado en el análisis de varianza. La relación fundamental entre la distribución t con v grados de libertad y la distribución F con 1 y v grados de libertad es

$$t^2 = f(1, v).$$

Desde luego, la prueba t permite probar contra la alternativa unilateral, en tanto que la prueba F está restringida a probar contra una alternativa bilateral.

Salida por computadora comentada para la regresión lineal simple

Considere el lector otra vez los datos de la tabla 11.1, sobre la demanda de oxígeno químico. En las figuras 11.14 y 11.15 se presentan salidas por computadora más completas. De nuevo se ilustran con el software MINITAB para PC. La columna de la razón t indica pruebas para la hipótesis nula de valores de cero en el parámetro. El término "Fit" denota los valores de $\hat{y}$, que con frecuencia se denominan **valores ajustados**. El término "SE FIT" se emplea para calcular los intervalos de confianza sobre la respuesta media. El valor de R^2 se calcula como $(SSR/SST) \times 100$, y significa la proporción de variación de las y explicada por la regresión de la línea recta. Asimismo, muestra los intervalos de confianza sobre la respuesta media y los intervalos de predicción sobre una observación nueva.

11.9 Prueba para la linealidad de la regresión: Datos con observaciones repetidas

En ciertas clases de situaciones experimentales, el investigador tiene la capacidad de efectuar observaciones repetidas de la respuesta para cada valor de x. Aunque no es necesario tener dichas repeticiones para estimar α y β, las repeticiones permiten al experimentador obtener información cuantitativa acerca de lo apropiado que resulta el modelo. En realidad, si se generan observaciones repetidas, el investigador puede efectuar una prueba de significancia para determinar si el modelo es adecuado o no.

Seleccionemos una muestra aleatoria de n observaciones con k valores distintos de x, por ejemplo, $x_1, x_2, \ldots, x_n$, de manera que la muestra contenga n_1 valores observados de la variable aleatoria Y_1 correspondientes a los valores x_1, que también contenga n_2 valores observados de Y_2 correspondientes a $x_2, \ldots,$ n_k valores observados de Y_k correspondientes a x_k. Necesariamente, $n = \sum_{i=1}^{k} n_i$.

```
The regression equation is COD = 3.83 + 0.904 Per_Red
Predictor    Coef  SE Coef      T      P
 Constant   3.830    1.768   2.17  0.038
  Per_Red 0.90364  0.05012  18.03  0.000
S = 3.22954   R-Sq = 91.3%   R-Sq(adj) = 91.0%
            Analysis of Variance
Source          DF      SS      MS       F    P
Regression       1  3390.6  3390.6  325.08  0.000
Residual Error 31   323.3    10.4
Total            32  3713.9

Obs  Per_Red     COD     Fit  SE Fit  Residual  St Resid
  1      3.0   5.000   6.541   1.627    -1.541     -0.55
  2     36.0  34.000  36.361   0.576    -2.361     -0.74
  3      7.0  11.000  10.155   1.440     0.845      0.29
  4     37.0  36.000  37.264   0.590    -1.264     -0.40
  5     11.0  21.000  13.770   1.258     7.230      2.43
  6     38.0  38.000  38.168   0.607    -0.168     -0.05
  7     15.0  16.000  17.384   1.082    -1.384     -0.45
  8     39.0  37.000  39.072   0.627    -2.072     -0.65
  9     18.0  16.000  20.095   0.957    -4.095     -1.33
 10     39.0  36.000  39.072   0.627    -3.072     -0.97
 11     27.0  28.000  28.228   0.649    -0.228     -0.07
 12     39.0  45.000  39.072   0.627     5.928      1.87
 13     29.0  27.000  30.035   0.605    -3.035     -0.96
 14     40.0  39.000  39.975   0.651    -0.975     -0.31
 15     30.0  25.000  30.939   0.588    -5.939     -1.87
 16     41.0  41.000  40.879   0.678     0.121      0.04
 17     30.0  35.000  30.939   0.588     4.061      1.28
 18     42.0  40.000  41.783   0.707    -1.783     -0.57
 19     31.0  30.000  31.843   0.575    -1.843     -0.58
 20     42.0  44.000  41.783   0.707     2.217      0.70
 21     31.0  40.000  31.843   0.575     8.157      2.57
 22     43.0  37.000  42.686   0.738    -5.686     -1.81
 23     32.0  32.000  32.746   0.567    -0.746     -0.23
 24     44.0  44.000  43.590   0.772     0.410      0.13
 25     33.0  34.000  33.650   0.563     0.350      0.11
 26     45.0  46.000  44.494   0.807     1.506      0.48
 27     33.0  32.000  33.650   0.563    -1.650     -0.52
 28     46.0  46.000  45.397   0.843     0.603      0.19
 29     34.0  34.000  34.554   0.563    -0.554     -0.17
 30     47.0  49.000  46.301   0.881     2.699      0.87
 31     36.0  37.000  36.361   0.576     0.639      0.20
 32     50.0  51.000  49.012   1.002     1.988      0.65
 33     36.0  38.000  36.361   0.576     1.639      0.52
```

Figura 11.14: Salida de *MINITAB* de la regresión lineal simple para los datos de demanda de oxígeno químico; parte I.

Se define

$$y_{ij} = \text{el } j\text{-ésimo valor de la variable aletoria } Y_i,$$
$$y_{i.} = T_{i.} = \sum_{j=1}^{n_i} y_{ij},$$
$$\bar{y}_{i.} = \frac{T_{i.}}{n_i}.$$

Obs	Fit	SE Fit	95% CI	95% PI
1	6.541	1.627	(3.223, 9.858)	(-0.834, 13.916)
2	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
3	10.155	1.440	(7.218, 13.092)	(2.943, 17.367)
4	37.264	0.590	(36.062, 38.467)	(30.569, 43.960)
5	13.770	1.258	(11.204, 16.335)	(6.701, 20.838)
6	38.168	0.607	(36.931, 39.405)	(31.466, 44.870)
7	17.384	1.082	(15.177, 19.592)	(10.438, 24.331)
8	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
9	20.095	0.957	(18.143, 22.047)	(13.225, 26.965)
10	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
11	28.228	0.649	(26.905, 29.551)	(21.510, 34.946)
12	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
13	30.035	0.605	(28.802, 31.269)	(23.334, 36.737)
14	39.975	0.651	(38.648, 41.303)	(33.256, 46.694)
15	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
16	40.879	0.678	(39.497, 42.261)	(34.149, 47.609)
17	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
18	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
19	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
20	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
21	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
22	42.686	0.738	(41.181, 44.192)	(35.930, 49.443)
23	32.746	0.567	(31.590, 33.902)	(26.059, 39.434)
24	43.590	0.772	(42.016, 45.164)	(36.818, 50.362)
25	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
26	44.494	0.807	(42.848, 46.139)	(37.704, 51.283)
27	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
28	45.397	0.843	(43.677, 47.117)	(38.590, 52.205)
29	34.554	0.563	(33.406, 35.701)	(27.868, 41.239)
30	46.301	0.881	(44.503, 48.099)	(39.473, 53.128)
31	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
32	49.012	1.002	(46.969, 51.055)	(42.115, 55.908)
33	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)

Figura 11.15: Salida de *MINITAB* de la regresión lineal simple para los datos de demanda de oxígeno químico; parte II.

Entonces, si $n_4 = 3$, las mediciones de Y se efectúan correspondiendo a $x = x_4$, y se indicarían estas observaciones como y_{41}, y_{42} y y_{43}. Por lo tanto,

$$Ti = y_{41} + y_{42} + y_{43}.$$

El concepto de la falta de ajuste

La suma de errores cuadráticos consiste en dos partes: la cantidad debida a la variación entre los valores de Y dentro de valores dados de x, y un componente que normalmente se denomina contribución a la **falta de ajuste**. El primer componente refleja tan sólo la variación aleatoria, o **error experimental puro**; mientras que el segundo es una medida de la variación sistemática introducida por los términos de orden superior. En nuestro caso, éstos son términos de x distintos de la contribución lineal o de primer orden. Observe que al elegir un modelo lineal en esencia se supone que este segundo componente no existe y, por lo tanto, la suma cuadrática de errores se debe por completo a errores aleatorios. Si éste fuera el caso, entonces

$S^2 = \text{SSE}/(n-2)$ es un estimador insesgado de σ^2. Sin embargo, si el modelo no se ajusta a los datos en forma apropiada, entonces la suma cuadrática de errores estará inflada y producirá un estimador sesgado de σ^2. Sea que el modelo se ajuste o no a los datos, siempre que se tienen observaciones repetidas es posible obtener un estimador insesgado de σ^2 calculando

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n_i - 1}; \quad i = 1, 2, \ldots, k,$$

para cada uno de los k valores distintos de x y, después, al agrupar estas varianzas, tenemos

$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n-k}.$$

El numerador de s^2 es una **medida del error experimental puro**. A continuación se presenta un procedimiento computacional para separar la suma cuadrática de errores en los dos componentes que representan el error puro y la falta de ajuste:

Cálculo de la suma de cuadrados debida a la falta de ajuste

1. Calcule la suma de los cuadrados del error puro

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2.$$

Esta suma de cuadrados tiene $n-k$ grados de libertad asociados con ella, y la media cuadrática resultante es el estimador insesgado s^2 de σ^2.

2. Reste la suma de cuadrados del error puro de la suma de cuadrados del error, *SSE*, con lo que se obtiene la suma de cuadrados debida a la falta de ajuste. Los grados de libertad de la falta de ajuste también se obtienen con la sola resta de $(n-2) - (n-k) = k-2$.

En la tabla 11.3 se resumen los cálculos que se requieren para la prueba de hipótesis en un problema de regresión con mediciones repetidas de la respuesta.

Tabla 11.3: Análisis de varianza para la prueba de la linealidad de la regresión

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
Regresión	SSR	1	SSR	$\frac{\text{SSR}}{s^2}$
Error	SSE	$n-2$		
Falta de ajuste	SSE − SSE (puro)	$k-2$	$\frac{\text{SSE} - \text{SSE (puro)}}{k-2}$	$\frac{\text{SSE} - \text{SSE (puro)}}{s^2(k-2)}$
Error puro	SSE (puro)	$n-k$	$s^2 = \frac{\text{SSE (puro)}}{n-k}$	
Total	SST	$n-1$		

Las figuras 11.16 y 11.17 ilustran los puntos muestrales para las situaciones del "modelo correcto" y del "modelo incorrecto". En la figura 11.16, donde $\mu_{Y|x}$ cae

sobre una línea recta, no hay falta de ajuste cuando se acepta un modelo lineal, por lo que la variación muestral alrededor de la recta de regresión es un error puro, resultante de la variación que ocurre entre observaciones repetidas. En la figura 11.17, donde claramente $\mu_{Y|x}$ no cae sobre una línea recta, la falta de ajuste por seleccionar en forma errónea un modelo lineal es responsable de la mayoría de la variación alrededor de la recta de regresión, además del error puro.

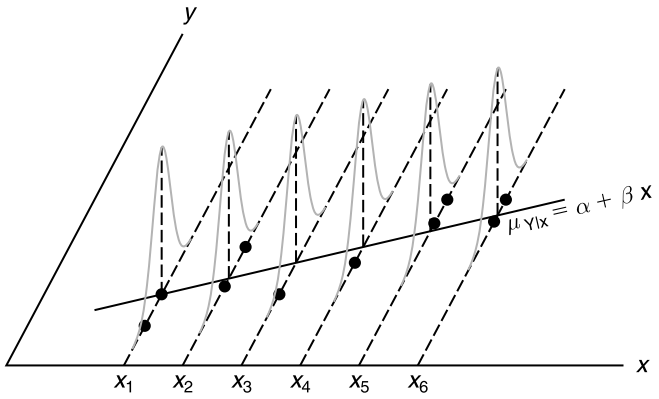


Figura 11.16: Modelo lineal correcto sin componente de falta de ajuste.

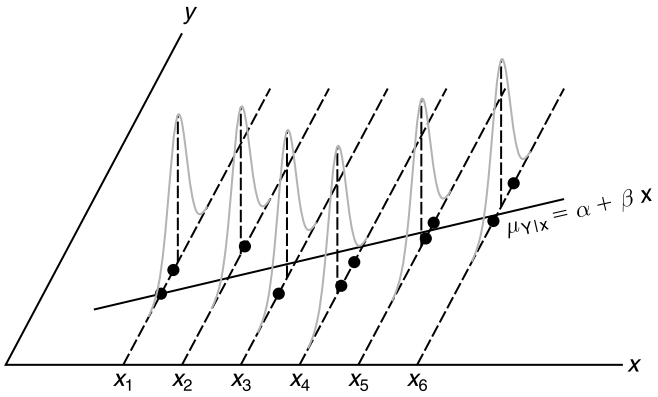


Figura 11.17: Modelo lineal incorrecto con componente de falta de ajuste.

¿Cuál es la importancia de detectar la falta de ajuste?

El concepto de falta de ajuste es importante en extremo en las aplicaciones del análisis de regresión. En realidad, la necesidad de construir o diseñar un experimento que tome en cuenta la falta de ajuste se vuelve más crítico que el problema mismo; en tanto que el mecanismo subyacente implicado se vuelve más complicado. Es seguro que no siempre se puede tener la certeza de que la estructura que se postula, en este

caso el modelo de regresión lineal, sea una representación correcta o incluso adecuada. El ejemplo siguiente muestra la manera en que se parte la suma de cuadrados del error en las dos componentes que representan el error puro y la falta de ajuste. Lo adecuado del modelo se prueba al nivel de significancia α, comparando la media cuadrática de la falta de ajuste dividida entre s^2 con $f_\alpha(k-2,\, n-k)$.

Ejemplo 11.8: En la tabla 11.4 se presenta el registro de las observaciones del producto de una reacción química, tomada a distintas temperaturas.

Tabla 11.4: Datos para el Ejemplo 11.8

y(%)	x(°C)	y(%)	x(°C)
77.4	150	88.9	250
76.7	150	89.2	250
78.2	150	89.7	250
84.1	200	94.8	300
84.5	200	94.7	300
83.7	200	95.9	300

Obtenga el modelo lineal $\mu_{Y|x} = \alpha + \beta x$ y pruebe la falta de ajuste.

Solución: En la tabla 11.5 se presentan los resultados de los cálculos.

Tabla 11.5: Análisis de varianza de los datos de producto-temperatura

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	f calculada	Valores P
Regresión	509.2507	1	509.2507	1531.2507	< 0.0001
Error	3.8660	10			
Falta de ajuste	1.2060	2	0.6030	1.81	0.2241
Error puro	2.6600	8	0.3325		
Total	513.1167	Total			

Conclusión: La partición de la variación total revela de esta manera una variación significativa en el modelo lineal, y una cantidad insignificante de variación debida a la falta de ajuste. De manera que los datos experimentales no parecen sugerir la necesidad de considerar en el modelo términos distintos de los de primer orden, y no se rechaza la hipótesis nula.

Salida de computadora comentada para probar la falta de ajuste

En la figura 11.18 se presenta una salida comentada de computadora para el análisis de los datos del ejemplo 11.8. El resultado es una salida SAS. Observe la "LOF" con 2 grados de libertad, que representa las contribuciones cuadrática y cúbica al modelo, y el valor P de 0.22, que sugiere que el modelo lineal (primer orden) es adecuado.

```
Dependent Variable: yield
                                        Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        3     510.4566667     170.1522222     511.74    <.0001
Error                        8       2.6600000       0.3325000
Corrected Total             11     513.1166667

              R-Square     Coeff Var      Root MSE    yield Mean
              0.994816      0.666751      0.576628      86.48333

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
temperature                  1     509.2506667     509.2506667    1531.58    <.0001
LOF                          2       1.2060000       0.6030000       1.81    0.2241
```

Figura 11.18: Salida *SAS* que muestra el análisis de los datos del ejemplo 11.8.

Ejercicios

11.35 *a*) Encuentre la estimación por mínimos cuadrados para el parámetro β, en la ecuación lineal $\mu_{Y|x} = \beta x$.

b) Estime la recta de regresión que pasa por el origen para los datos siguientes:

x	0.5	1.5	3.2	4.2	5.1	6.5
y	1.3	3.4	6.7	8.0	10.0	13.2

11.36 Suponga que en el ejercicio 11.35 no se sabe si la regresión real debe pasar por el origen. Estime el modelo lineal $\mu_{Y|x} = \alpha + \beta x$ y pruebe la hipótesis de que $\alpha = 0$ con el nivel de significancia de 0.10, contra la alternativa de que $\alpha \neq 0$.

11.37 Suponga que un experimentador propone un modelo del tipo

$$Y_i = \alpha + \beta x_{1i} + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

cuando en realidad una variable adicional, x_2, también contribuye linealmente a la respuesta. Entonces, el modelo verdadero está dado por

$$Y_i = \alpha + \beta x_{1i} + \gamma x_{2i} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Calcule el valor esperado del estimador

$$B = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)Y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}.$$

11.38 En el ejercicio 11.3 de la página 398, utilice el enfoque del análisis de varianza para probar la hipótesis de que $\beta = 0$, contra la hipótesis alternativa de que $\beta \neq 0$, con un nivel de significancia de 0.05.

11.39 En los pesticidas se utilizan compuestos de organofosfatos (OF). Sin embargo, es importante estudiar el efecto que tienen sobre las especies expuestos a ellos. Como parte del estudio de laboratorio *Some Effects of Organophosphate Pesticides on Wildlife Species*, elaborado por el Departamento de Pesca y Vida Silvestre del Instituto Politécnico y Universidad Estatal de Virginia, se realizó un experimento en el cual se suministraron distintas dosis de un pesticida OF en particular a 5 grupos de 5 ratones (*peromysius leucopus*). Los 25 ratones eran hembras de edad y condiciones similares. Un grupo no recibió el producto. La respuesta básica y consistió en medir la actividad cerebral. Se postuló que dicha actividad disminuiría con el incremento en la dosis de OF. A continuación se presentan los datos:

Animal	**Dosis, x (mg/kg de peso corporal)**	**Actividad, y (moles/litro/min)**
1	0.0	10.9
2	0.0	10.6
3	0.0	10.8
4	0.0	9.8
5	0.0	9.0
6	2.3	11.0
7	2.3	11.3
8	2.3	9.9
9	2.3	9.2
10	2.3	10.1
11	4.6	10.6
12	4.6	10.4
13	4.6	8.8
14	4.6	11.1
15	4.6	8.4
16	9.2	9.7
17	9.2	7.8
18	9.2	9.0
19	9.2	8.2
20	9.2	2.3
21	18.4	2.9
22	18.4	2.2
23	18.4	3.4
24	18.4	5.4
25	18.4	8.2

a) Con el modelo

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, 2, \ldots, 25,$$

encuentre los estimadores de mínimos cuadrados de α y β.

b) Construya una tabla de análisis de varianza en la cual aparezcan por separado los errores por falta de ajuste y puro. Determine si la falta de ajuste es significativa al nivel de 0.05. Interprete los resultados.

11.40 En el ejercicio 11.5 de la página 398, pruebe la linealidad de la regresión. Use un nivel de significancia de 0.05. Haga comentarios al respecto.

11.41 Para el ejercicio 11.6 de la página 398, pruebe la linealidad de la regresión. Comente sus resultados.

11.42 La ganancia de un transistor en un dispositivo de circuito integrado, entre el emisor y el colector (hFE), se relaciona con dos variables [Myers y Montgomery (2002)] que se controlan en el proceso de deposición, controlado por el emisor en el tiempo (x_1, en minutos) y la dosis del emisor (x_2, en iones $\times 10^{14}$). Se observaron 14 muestras después de la deposición, y los datos resultantes se presentan en la tabla siguiente. Consideraremos modelos de regresión lineal usando la ganancia como respuesta, y el control del emisor en el tiempo o la dosis del emisor, como las variables regresoras.

Obs.	x_1, (tiempo de control, min)	x_2, (dosis, iones $\times 10^{14}$)	y, (ganancia o hFE)
1	195	4.00	1004
2	255	4.00	1636
3	195	4.60	852
4	255	4.60	1506
5	255	4.20	1272
6	255	4.10	1270
7	255	4.60	1269
8	195	4.30	903
9	255	4.30	1555
10	255	4.00	1260
11	255	4.70	1146
12	255	4.30	1276
13	255	4.72	1225
14	340	4.30	1321

a) Determine si el tiempo de control del emisor influye en la ganancia en una relación lineal. Es decir, pruebe H_0: $\beta_1 = 0$, donde β_1 es la pendiente de la variable regresora.

b) Efectúe una prueba de falta de ajuste para determinar si es adecuada la relación lineal. Saque sus conclusiones.

c) Determine si la dosis del emisor influye en la ganancia en una relación lineal. ¿Cuál variable regresora es el mejor predictor de la ganancia?

11.43 Los siguientes datos son el resultado de una investigación sobre el efecto de la temperatura de reacción, x, sobre la conversión porcentual de un proceso químico, y. [*Véase* Myers y Montgomery (2002).] Haga un ajuste por regresión lineal simple y utilice pruebas de falta de ajuste para determinar si el modelo es adecuado. Analice los resultados.

Observación	Temperatura (°C), x	Conversión %, y
1	200	43
2	250	78
3	200	69
4	250	73
5	189.65	48
6	260.35	78
7	225	65
8	225	74
9	225	76
10	225	79
11	225	83
12	225	81

11.44 Es frecuente que se utilice el tratamiento con calor para carburar partes metálicas como los engranes. El espesor de la capa carburada se considera una característica importante del engrane, y contribuye a la confiabilidad conjunta de la parte. Debido a la naturaleza crítica de esta característica, se realiza una prueba de laboratorio para cada lote del horno. La prueba es destructiva, y consiste en que la parte real se corta en forma transversal y se sumerge en un producto químico durante cierto tiempo. Esta prueba requiere efectuar un análisis del carbono sobre la superficie tanto de la parte superior del engrane (arriba de los dientes) como de su raíz (entre los dientes). Los datos siguientes son los resultados de la prueba de análisis de carbono en las 19 partes.

Tiempo de inmersión	Grado	Tiempo de inmersión	Grado
0.58	0.013	1.17	0.021
0.66	0.016	1.17	0.019
0.66	0.015	1.17	0.021
0.66	0.016	1.20	0.025
0.66	0.015	2.00	0.025
0.66	0.016	2.00	0.026
1.00	0.014	2.20	0.024
1.17	0.021	2.20	0.025
1.17	0.018	2.20	0.024
1.17	0.019		

a) Haga un ajuste de regresión lineal simple que relacione el grado del análisis de carbono, y, contra el tiempo de inmersión. Pruebe H_0: $\beta_1 = 0$.

b) Si se rechaza la hipótesis del inciso *a*), determine si el modelo lineal es adecuado.

11.45 Se desea obtener un modelo de regresión que relacione la temperatura con la proporción de impurezas de una sustancia sólida que pasa a través de helio sólido. Se lista la temperatura en grados centígrados. A continuación se presentan los datos.

a) Ajuste un modelo de regresión lineal.

b) ¿Parece que la proporción de impurezas que pasan a través del helio incrementa la temperatura conforme ésta se acerca a -273 grados centígrados?

c) Encuentre R^2.

d) Con base en la información anterior, ¿parece adecuado el modelo lineal? ¿Qué información adicional necesitaría el lector para responder mejor a la pregunta?

Temperatura (C)	Proporción de impurezas
−260.5	.425
−255.7	.224
−264.6	.453
−265.0	.475
−270.0	.705
−272.0	.860
−272.5	.935
−272.6	.961
−272.8	.979
−272.9	.990

11.46 Es de interés estudiar el efecto que tiene el tamaño de la población de varias ciudades de Estados Unidos sobre las concentraciones de ozono. Los datos consisten en la población de 1999, en millones de habitantes y en la cantidad de ozono presente por hora en ppmm (partes por mil millones). Los datos son los siguientes:

Ozono (ppmm/hora), y	Población, x
126	0.6
135	4.9
124	0.2
128	0.5
130	1.1
128	0.1
126	1.1
128	2.3
128	0.6
129	2.3

a) Ajuste un modelo de regresión lineal que relacione la concentración del ozono con la población. Pruebe H_0: $\beta = 0$ usando el enfoque del ANOVA.

b) Realice una prueba de la falta de ajuste. Con base en los resultados de la prueba, ¿es apropiado el modelo lineal?

c) Pruebe la hipótesis del inciso *a*) utilizando la media cuadrática del error puro en la prueba F. ¿Cambian los resultados? Comente las ventajas de cada prueba.

11.47 Evaluar la deposición del nitrógeno de la atmósfera es una tarea importante de The National Atmospheric Deposition Program (NADP, asociación de muchas instituciones). La NAPD está estudiando la deposición atmosférica y su efecto sobre los cultivos agrícolas, las aguas superficiales de los bosques, y otros recursos. Los óxidos del nitrógeno pueden tener efectos sobre el ozono atmosférico y la cantidad de nitrógeno puro que se encuentra en el aire que respiramos. A continuación se presentan los datos:

Año	Óxido de nitrógeno
1978	0.73
1979	2.55
1980	2.90
1981	3.83
1982	2.53
1983	2.77
1984	3.93
1985	2.03
1986	4.39
1987	3.04
1988	3.41
1989	5.07
1990	3.95
1991	3.14
1992	3.44
1993	3.63
1994	4.50
1995	3.95
1996	5.24
1997	3.30
1998	4.36
1999	3.33

a) Grafique los datos.

b) Ajuste un modelo de regresión lineal y obtenga R^2.

c) ¿Qué puede decirse acerca de la tendencia de los óxidos con el paso del tiempo?

11.48 Para una variedad particular de planta, los investigadores desean desarrollar una fórmula para predecir la cantidad de semillas (gramos) como función de la densidad de las plantas. Efectuaron un estudio con cuatro niveles del factor X, el número de plantas por parcela. Se utilizaron cuatro réplicas para cada nivel de X. A continuación se muestran los datos:

Plantas por parcela x	Cantidad de semillas, y (gramos)			
10	12.6	11.0	12.1	10.9
20	15.3	16.1	14.9	15.6
30	17.9	18.3	18.6	17.8
40	19.2	19.6	18.9	20.0

¿Es adecuado un modelo de regresión lineal para analizar este conjunto de datos?

11.10 Gráficas de datos y transformaciones

En este capítulo se estudia la construcción de modelos de regresión en los que hay una variable independiente o regresora. Además, durante la construcción del modelo se supone que tanto x como y entran en el modelo en *forma lineal*. Con frecuencia,

es aconsejable trabajar con un modelo alternativo en el que x o y (o ambas) intervengan en una forma no lineal. Es posible que se prescriba una ***transformación*** de los datos debido a consideraciones teóricas inherentes al estudio científico, o a que una simple gráfica de los datos sugiera la necesidad de *reexpresar* las variables del modelo. La necesidad de llevar a cabo una transformación es muy fácil de diagnosticar en el caso de la regresión lineal simple, debido a que las gráficas en dos dimensiones brindan un panorama verdadero de la manera en que las variables se comportan en el modelo.

Un modelo en el que se transformen x o y no debería verse como un *modelo de regresión no lineal*. Por lo general, se denomina como lineal a un modelo de regresión cuando es **lineal en los parámetros**. En otras palabras, suponga que lo complejo de los datos u otra información científica sugiera que debe hacerse la **regresión de y^* contra x^***, donde cada una de ellas es una transformación de las variables naturales x y y. Entonces, el modelo de la forma

$$y_i^* = \alpha + \beta x_i^* + \epsilon_i$$

es lineal porque lo es en los parámetros α y β. El material que se estudió en las secciones 11.2 a 11.9 permanece sin cambio, con y_i^* y x_i^* que reemplazan a y_i y a x_i. Un ejemplo sencillo y útil es el modelo log-log:

$$\log y_i = \alpha + \beta \log x_i + \epsilon_i.$$

Aunque este modelo es no lineal en x y y, sí lo es en los parámetros y por ello recibe el tratamiento de un modelo lineal. Por otro lado, un ejemplo de modelo verdaderamente no lineal es:

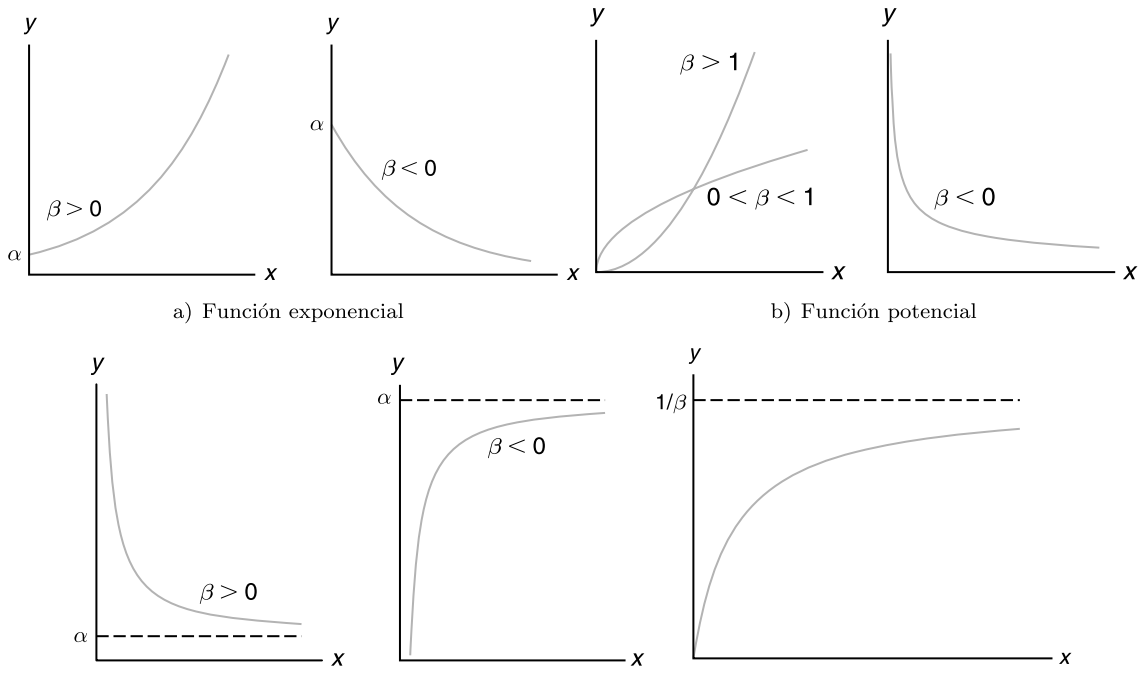
$$y_i = \beta_0 + \beta_1 x^{\beta_2} + \epsilon_i,$$

donde debe estimarse el parámetro β_2 (así como β_0 y β_1). El modelo es no lineal en β_2.

Las transformaciones susceptibles de mejorar el ajuste y la predictibilidad del modelo son muy numerosas. Para un análisis completo de las transformaciones, el lector puede consultar a Myers (1990, véase la bibliografía). Aquí indicamos algunas de ellas y mostraremos la apariencia de las gráficas que sirven como diagnóstico. Considere la taba 11.6. Ahí se dan varias funciones que describen relaciones entre y y x que producen una *regresión lineal* con la transformación indicada. Además, con la finalidad de dar todo completo, se presentan al lector las variables dependiente e independiente por utilizar en la *regresión lineal simple* resultante. La figura 11.19 ilustra las funciones que se listan en la tabla 11.6. Éstas sirven como guía para el análisis en la elección de una transformación a partir de la observación de la gráfica de y contra x.

Tabla 11.6: Algunas transformaciones útiles para linealizar

Forma funcional que relaciona y con x	**Transformación propia**	**Forma de la regresión lineal simple**
Exponencial: $y = \alpha e^{\beta x}$	$y^* = \ln y$	Hacer la regresión de y^* contra x
Potencia: $y = \alpha x^{\beta}$	$y^* = \log y; \quad x^* = \log x$	Hacer la regresión de y^* contra x^*
Recíproca: $y = \alpha + \beta\left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Hacer la regresión de y contra x^*
Función hiperbólica: $y = \frac{x}{\alpha+\beta x}$	$y^* = \frac{1}{y}; \quad x^* = \frac{1}{x}$	Hacer la regresión de y^* contra x^*



a) Función exponencial

b) Función potencial

c) Función recíproca

d) Función hiperbólica

Figura 11.19: Diagramas que ilustran las funciones enlistadas en la tabla 11.6.

¿Cuáles son las implicaciones de un modelo transformado?

Lo que sigue intenta ser una ayuda para el analista, cuando es evidente que una transformación producirá una mejoría. Sin embargo, antes de dar un ejemplo, deben mencionarse dos puntos importantes. El primero tiene que ver con la escritura formal del modelo, una vez que se hayan transformado los datos. Con mucha frecuencia el analista no piensa en esto: tan solo lleva a cabo la transformación sin ningún interés en la forma del modelo *antes* ni *después* de la transformación. El modelo exponencial sirve como una ilustración buena de ello. El modelo en las variables naturales (no transformadas) que produce un *modelo de error aditivo* en las variables transformadas está dado por

$$y_i = \alpha e^{\beta x_i} \cdot \epsilon_i,$$

que es un *modelo de error multiplicativo*. Al sacar logaritmos es claro que se obtiene

$$\ln y_i = \ln \alpha + \beta x_i + \ln \epsilon_i.$$

Como resultado, las suposiciones básicas se efectúan sobre $\ln \epsilon_i$. El propósito de esta presentación únicamente es recordar al lector que no debe verse una transformación tan sólo como una manipulación algebraica a la cual se suma un error. Con frecuencia, un modelo en las variables transformadas que tiene una adecuada *estructura de error aditivo* es resultado de un modelo en las variables naturales con una estructura de error diferente.

El segundo punto importante es sobre la noción de las medidas de mejoría. Las medidas evidentes para comparar son, por supuesto, el valor de R^2 y la media

cuadrática de los residuos, s^2. (En el capítulo 12 se dan otras mediciones de rendimiento de las comparaciones entre modelos que compiten.) Ahora, si la respuesta y no se transforma, entonces es claro que s^2 y R^2 se pueden usar para medir la utilidad de la transformación. Los residuos estarán en las mismas unidades para los dos modelos: el transformado y el que no lo está. Pero cuando y se transforma, los criterios de rendimiento para el modelo transformado deberían basarse en los valores de los residuos en las unidades de medida de la respuesta no transformada. De ese modo las comparaciones son más apropiadas. El siguiente ejemplo proporciona una ilustración de lo anterior.

Ejemplo 11.9: Se registra la presión P de un gas que corresponde a distintos volúmenes V, y los datos se presentan en la tabla 11.7.

Tabla 11.7: Datos para el ejemplo 11.9

V(cm^3)	50	60	70	90	100
P(kg/cm^2)	64.7	51.3	40.5	25.9	7.8

La ley del gas ideal está dada por la forma funcional $PV^\gamma = C$, donde γ y C son constantes. Estime las constantes C y γ.

Solución: Se tomarán logaritmos naturales en ambos lados del modelo

$$P_i V^\gamma = C \cdot \epsilon_i, \quad , i = 1, 2, 3, 4, 5.$$

Como resultado, es posible escribir el modelo lineal

$$\ln P_i = \ln C - \gamma \ln V_i + \epsilon_i^*, \quad i = 1, 2, 3, 4, 5,$$

donde $\epsilon_i^* = \ln \epsilon_i$. Los siguientes son los resultados de la regresión lineal simple:

Intersección: $\widehat{\ln C} = 14.7589$, $\widehat{C} = 2,568,862.88$, Pendiente: $\hat{\gamma} = 2.65347221$.

La siguiente tabla representa información tomada del análisis de regresión.

$\boldsymbol{P_i}$	$\boldsymbol{V_i}$	$\boldsymbol{\ln P_i}$	$\boldsymbol{\ln V_i}$	$\boldsymbol{\widehat{\ln P_i}}$	$\boldsymbol{\widehat{P_i}}$	$\boldsymbol{e_i = P_i - \widehat{P_i}}$
64.7	50	4.16976	3.91202	4.37853	79.7	−15.0
51.3	60	3.93769	4.09434	3.89474	49.1	2.2
40.5	70	3.70130	4.24850	3.48571	32.6	7.9
25.9	90	3.25424	4.49981	2.81885	16.8	9.1
7.8	100	2.05412	4.60517	2.53921	12.7	−4.9

Es instructivo graficar los datos y la ecuación de regresión. La figura 11.20 muestra una gráfica de los datos no transformados de presión y volumen; en tanto que la curva representa la ecuación de regresión.

Gráficas de diagnóstico de los residuos: Detección gráfica de la trasgresión de las suposiciones

Las gráficas de los datos crudos son de mucha ayuda para determinar la naturaleza del modelo que debe ajustárseles cuando tan sólo hay una variable independiente. En seguida se intenta ilustrar esto. Sin embargo, la detección de la forma del modelo adecuado no es el único beneficio que se obtiene con la gráfica de diagnóstico. Como

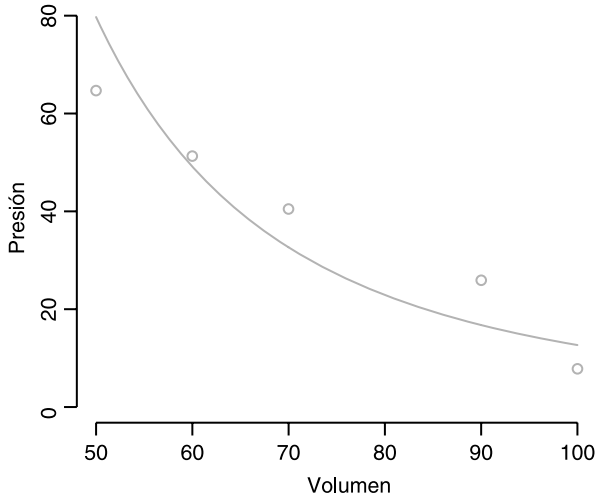


Figura 11.20: Datos de presión y volumen y la regresión ajustada.

en gran parte del material asociado con las pruebas de hipótesis que se trata en el capítulo 10, los métodos de graficación ilustran y detectan la trasgresión de las suposiciones. El lector debería recordar que muchos de los conceptos que se ilustran en el capítulo requieren de suposiciones sobre los errores del modelo, las ϵ_i. En los hechos, se supone que las ϵ_i son variables aleatorias independientes $N(0,\sigma)$. Por supuesto, las ϵ_i no son observadas al principio. Sin embargo, los $e_i = y_i - \hat{y}_i$, los *residuos*, son los errores en el ajuste de la recta de regresión, por lo que sirven para reproducir los ϵ_i. Así, la complejidad de estos residuos con frecuencia resalta las dificultades. La gráfica de los residuos, idealizada por supuesto, es como la que se aprecia en la figura 11.21. Es decir, deberían demostrar en verdad fluctuaciones aleatorias alrededor del valor de cero.

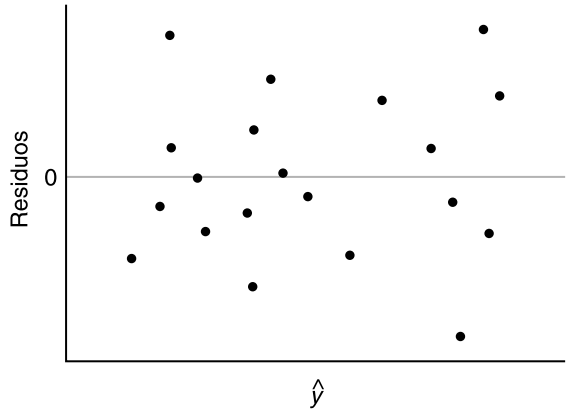


Figura 11.21: Gráfica ideal de los residuos.

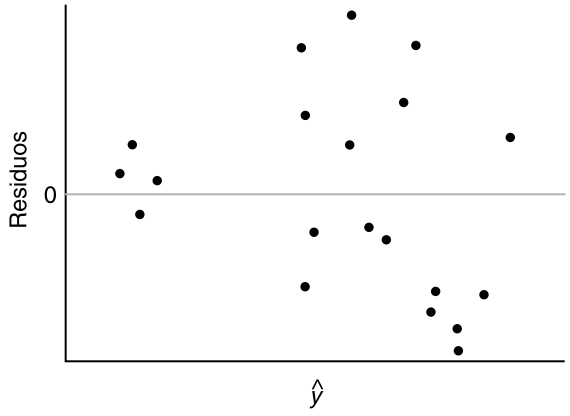


Figura 11.22: Gráfica de los residuos que ilustra una varianza heterogénea de los errores.

Varianza no homogénea

Una suposición importante que se hace en el análisis de regresión es la varianza homogénea. Con frecuencia, las trasgresiones se detectan con la apariencia de la gráfica de residuos. En los datos científicos, una condición común es que se incremente la varianza del error con el aumento de la variable regresora. Una varianza grande del error produce residuos grandes y, por ende, una gráfica de ellos como la que se presenta en la figura 11.22 es señal de varianza no homogénea. En el capítulo 12 se presenta un análisis más amplio acerca de las gráficas de los residuos e información acerca de los diferentes tipos de éstos.

Gráfica de la probabilidad normal

La suposición de que los errores del modelo son normales se hace cuando el analista de los datos aborda ya sea las pruebas de hipótesis o la estimación de intervalos de confianza. De nuevo, la contraparte numérica de los ϵ_i, los residuos, son sujetos de diagnosticarse mediante la graficación para detectar cualesquiera trasgresiones extremas. En el capítulo 8 se presentaron las gráficas normales cuantil-cuantil y se analizaron en forma breve las de probabilidad normal. En el estudio de caso que se desarrolla en la siguiente sección se ilustran las gráficas de residuos.

11.11 Caso de estudio de regresión lineal simple

En la manufactura de productos comerciales de madera es importante estimar la relación que hay entre la densidad de un producto de madera y su rigidez. Está en consideración un tipo relativamente nuevo de aglomerado que puede hacerse con mucha mayor facilidad que el producto comercial ya aceptado. Es necesario saber cuál es la densidad con que su rigidez es comparable con la del producto comercial bien conocido y documentado. El estudio lo realizó Terrance E. Conners, *Investigation of Certain Mechanical Properties of a Wood-Foam Composite* (M.S. Thesis, Departamento de Bosques y Vida Silvestre, University of Massachusetts). Se produjeron 30 tableros de aglomerado con densidades que variaban aproximadamente de 8 a 26 libras por pie cúbico, y se midió la rigidez en libras por pulgada cuadrada. En la tabla 11.8 se presentan los datos.

Es necesario que el analista de datos se centre en un ajuste apropiado para los datos, y que utilice los métodos de inferencia que se estudian en este capítulo. Pueden ser apropiadas tanto la prueba de hipótesis sobre la pendiente de la regresión, como la estimación de los intervalos de confianza o predicción. Se comenzará presentando un simple diagrama de dispersión de los datos crudos con una regresión lineal simple sobrepuesta. En la figura 11.23 se observa dicha gráfica.

El ajuste de regresión lineal simple a los datos produce el modelo ajustado

$$\hat{y} = -25,433.739 + 3,884.976x \quad (R^2 = 0.7975),$$

y ya es posible calcular los residuos. La figura 11.24 presenta los residuos graficados contra las mediciones de la densidad. Éste difícilmente es un conjunto de residuos ideal o satisfactorio, pues no muestran una distribución al azar alrededor del valor de cero. En realidad, los agrupamientos de valores positivos y negativos sugerirían que debe investigarse una tendencia curvilínea en los datos.

Para tener idea sobre la suposición de la distribución normal de los errores, se generó una gráfica de probabilidad normal de los residuos. Éste es el tipo de gráfica que se estudió en la sección 8.3, donde el eje vertical representa la función de distribución empírica en una escala que produce una línea recta cuando se grafica contra

Tabla 11.8: Densidad y rigidez de 30 tableros de aglomerado

Densidad, x	**Rigidez, y**	**Densidad, x**	**Rigidez, y**
9.50	14,814.00	8.40	17,502.00
9.80	14,007.00	11.00	19,443.00
8.30	7,573.00	9.90	14,191.00
8.60	9,714.00	6.40	8,076.00
7.00	5,304.00	8.20	10,728.00
17.40	43,243.00	15.00	25,319.00
15.20	28,028.00	16.40	41,792.00
16.70	49,499.00	15.40	25,312.00
15.00	26,222.00	14.50	22,148.00
14.80	26,751.00	13.60	18,036.00
25.60	96,305.00	23.40	104,170.00
24.40	72,594.00	23.30	49,512.00
19.50	32,207.00	21.20	48,218.00
22.80	70,453.00	21.70	47,661.00
19.80	38,138.00	21.30	53,045.00

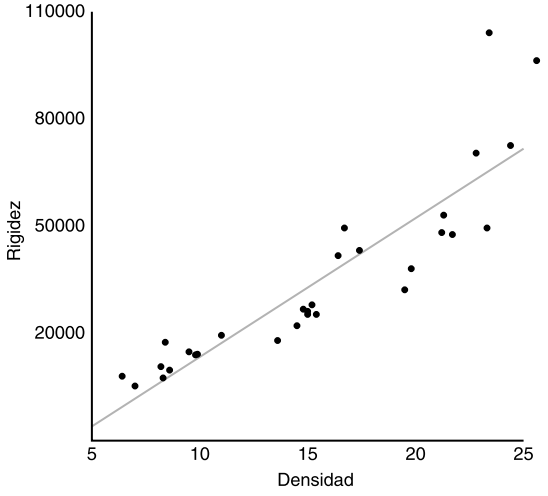


Figura 11.23: Diagrama de dispersión de los datos de densidad de la madera.

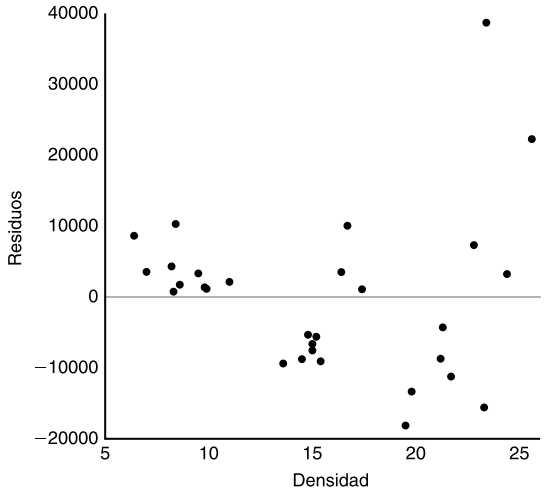


Figura 11.24: Gráfica de los residuos para los datos de densidad de la madera.

los residuos mismos. En la figura 11.25 se muestra la gráfica de probabilidad normal de los residuos. Esta gráfica no refleja la apariencia de recta que a uno le gustaría ver, lo cual es otro síntoma de una selección errónea, quizá sobresimplificada, de un modelo de regresión.

Ambos tipos de gráficas de residuo y, también, el diagrama de dispersión sugieren que sería adecuado un modelo algo más complicado. Una posibilidad es usar un modelo con transformación de logaritmos naturales. En otras palabras, hay que elegir hacer la regresión de $\ln y$ contra x. Esto produce la regresión

$$\widehat{\ln y} = 8.257 + 0.125x \quad (R^2 = 0.9016).$$

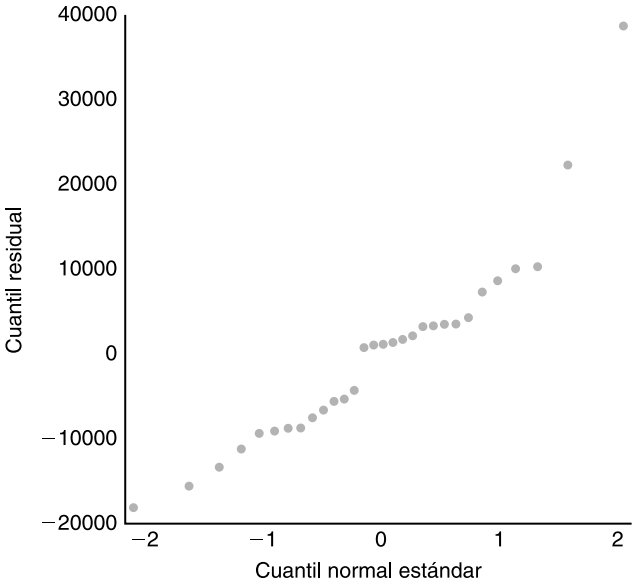


Figura 11.25: Gráfica de probabilidad normal de los residuos para los datos de densidad de la madera.

Para obtener alguna perspectiva de si el modelo transformado es más apropiado, considere las figuras 11.26 y 11.27, que muestran gráficas de los residuos de la rigidez [es decir, y_i -antilog $(\widehat{\ln y})$] contra la densidad. La figura 11.26 parece más cercana a un patrón aleatorio alrededor del cero, en tanto que la 11.27 con seguridad se acerca a una línea recta. Esto, además de un valor de R^2 más elevado, sugeriría que el modelo transformado es más apropiado.

11.12 Correlación

Hasta este momento se ha supuesto que la variable regresora independiente x es una variable científica o física, pero no aleatoria. En la realidad de este contexto es frecuente que x se denomine **variable matemática**, la cual, en el proceso de muestreo, se mide con un error despreciable. En muchas aplicaciones de las técnicas de regresión es más realista suponer que tanto X como Y son variables aleatorias y que las mediciones $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ son observaciones de una población que tiene la función densidad conjunta $f(x, y)$. Se debe tener en cuenta el problema de medir la relación entre las dos variables X y Y. Por ejemplo, si X y Y representaran la longitud y la circunferencia de una clase particular de hueso en el cuerpo de un adulto, se debe realizar un estudio antropológico para determinar si los valores grandes de X están asociados con valores grandes de Y, y viceversa.

Por otro lado, si X representa la edad de un automóvil usado y Y representa su valor en libros al menudeo, se esperaría que los valores grandes de X correspondieran a valores pequeños de Y, y los valores pequeños de X tuvieran correspondencia con los grandes de Y. El **análisis de correlación** intenta medir la intensidad de tales relaciones entre dos variables por medio de un solo número denominado **coeficiente de correlación**.

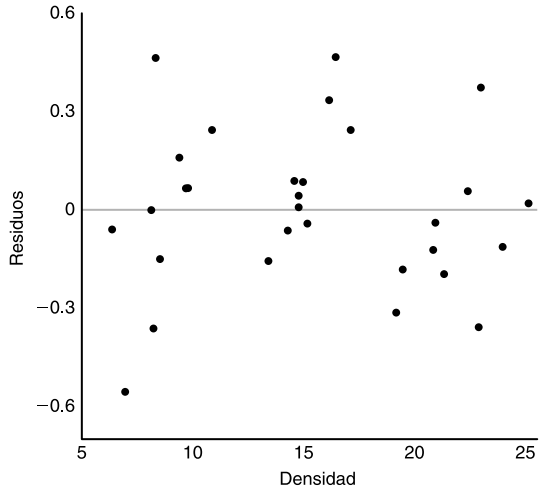


Figura 11.26: Gráfica de residuos donde se utiliza una transformación logarítimica para los datos de densidad de la madera.

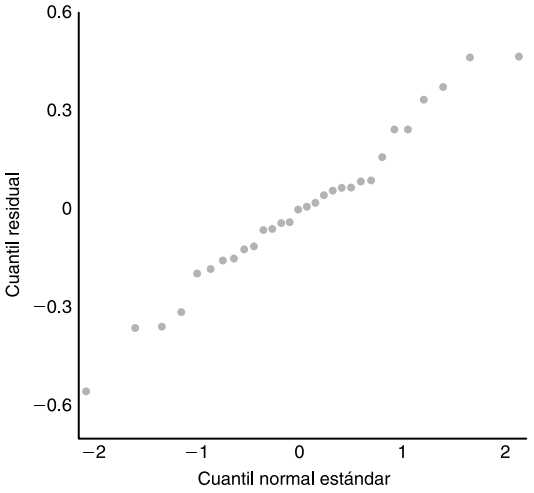


Figura 11.27: Gráfica de probabilidad normal de residuos en la cual se utiliza una transformación logarítmica para los datos de densidad de la madera.

En teoría, con frecuencia se supone que la distribución condicional $f(y|x)$ de Y, para valores fijos de X, es normal con media $\mu_{Y|x} = \alpha + \beta x$ y varianza $\sigma^2_{Y|x} = \sigma^2$, y que de igual manera, X está distribuida en forma normal con media μ y varianza σ^2_x. Entonces, la densidad conjunta de X y de Y es

$$\begin{aligned} f(x,y) &= n(y|x;\alpha+\beta x,\sigma)n(x;\mu_X,\sigma_X) \\ &= \frac{1}{2\pi\sigma_x\sigma}\exp\left\{-\frac{1}{2}\left[\left(\frac{y-\alpha-\beta x}{\sigma}\right)^2+\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]\right\}, \end{aligned}$$

para $-\infty < x < \infty$, y $-\infty < y < \infty$.

Escribamos la variable aleatoria Y en la forma

$$Y = \alpha + \beta X + \epsilon,$$

donde ahora X es una variable aleatoria independiente del error aleatorio ϵ. Como la media del error aleatorio es cero, se sigue que

$$\mu_Y = \alpha + \beta\mu_X \qquad \text{y} \qquad \sigma^2_Y = \sigma^2 + \beta^2\sigma^2_X.$$

Al sustituir α y σ^2 en la expresión anterior para $f(x, y)$, se obtiene la **distribución normal bivariada**

$$\begin{aligned} f(x,y) = &\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\ &\times\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2-2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}, \end{aligned}$$

para $-\infty < x < \infty$ y $-\infty < y < \infty$, donde

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} = \beta^2 \frac{\sigma_X^2}{\sigma_Y^2}.$$

La constante ρ (ro) se denomina **coeficiente de correlación de la población**, y juega un papel importante en muchos problemas de análisis de datos bivariados. Es importante que el lector entienda la interpretación física de este coeficiente de correlación y la diferencia entre correlación y regresión. El término *regresión* aún tiene algún significado aquí. De hecho, la línea recta dada por $\mu_{Y|x} = \alpha + \beta x$ aún se llama recta de regresión igual que antes, y los estimadores de a y β son idénticos a los que se dieron en la sección 11.3. El valor de ρ es 0 cuando $\beta = 0$, que resulta cuando en esencia no existe regresión lineal; es decir, la recta de regresión es horizontal y cualquier conocimiento de X es inútil para predecir Y. Como $\sigma_Y^2 \geq \sigma^2$, se debe tener $\rho^2 \leq 1$ y, por ello, $-1 \leq \rho \leq 1$. Los valores de $\rho = \pm 1$ sólo ocurren cuando $\sigma^2 = 0$, en cuyo caso se tiene una relación lineal perfecta entre las dos variables. Así, un valor de ρ igual a $+1$ implica una relación lineal perfecta con pendiente positiva, en tanto que un valor de ρ igual a -1 resulta de una relación lineal perfecta con pendiente negativa. Entonces, podría decirse que los estimadores muestrales de ρ con magnitud cercana a la unidad implican una buena correlación o **asociación lineal** entre X y Y; mientras que valores cerca de cero indican poca o ninguna correlación.

Para obtener una estimación muestral de ρ, hay que recordar, de la sección 11.4, que la suma cuadrática del error es

$$\text{SSE} = S_{yy} - bS_{xy}.$$

Al dividir ambos lados de esta ecuación entre S_{yy} y reemplazar S_{xy} con bS_{xx}, se obtiene la relación

$$b^2 \frac{S_{xx}}{S_{yy}} = 1 - \frac{\text{SSE}}{S_{yy}}.$$

El valor de $b^2 S_{xx}/S_{yy}$ es igual a cero cuando $b = 0$, lo que ocurrirá cuando los puntos muestrales no tengan relación lineal. Como $S_{yy} \geq$ SSE, se concluye que $b^2 S_{xx}/S_{xy}$ debe estar entre 0 y 1. En consecuencia, $b\sqrt{S_{xx}/S_{yy}}$ debe variar entre -1 y $+1$, y los valores negativos corresponden a rectas con pendientes positivas. Un valor de -1 o $+1$ sucederá cuando SSE $= 0$, pero éste es el caso en el que todos los puntos muestrales caen sobre una línea recta. Por lo tanto, una relación lineal perfecta se da en los datos muestrales cuando $b\sqrt{S_{xx}/S_{yy}} = \pm 1$. Es claro que la cantidad $b\sqrt{S_{xx}/S_{yy}}$, que se designará de aquí en adelante como r, puede usarse como un estimador del coeficiente de correlación ρ de la población. Se acostumbra hacer referencia al estimador r como **coeficiente de correlación producto-momento de Pearson**, o tan sólo **coeficiente de correlación muestral**.

Coeficiente de correlación

La medida ρ de la asociación lineal entre dos variables X y Y se estima por medio del **coeficiente de correlación muestral** r, donde

$$r = b\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Debe tenerse cuidado en la interpretación de valores de r entre -1 y $+1$. Por ejemplo, valores de r iguales a 0.3 y 0.6 significan sólo que se tiene dos correlaciones

positivas, una un poco más fuerte que la otra. Sería un error concluir que $r = 0.6$ indica una relación lineal dos veces mejor que la del valor $r = 0.3$. Por otro lado, si se escribe

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{\text{SSR}}{S_{yy}},$$

entonces, r^2, que por lo general se denomina **coeficiente muestral de determinación**, representa la proporción de la variación de S_{yy} explicada por la regresión de Y sobre x, que es SSR. Es decir, r^2 expresa la proporción de la variación total de los valores de la variable Y que son ocasionados o explicados por una relación lineal con los valores de la variable aleatoria X. Así, una correlación de 0.6 significa que 0.36 o 36% de la variación total de los valores de Y en la muestra se debe a la relación lineal con los valores de X.

Ejemplo 11.10: Es importante que los investigadores científicos del área de productos forestales sean capaces de estudiar la correlación entre la anatomía y las propiedades mecánicas de los árboles. De acuerdo con el estudio *Quantitative Anatomical Characteristics of Plantation Grown Loblolly Pine (Pinus Taeda L.) and Cottonwood (Populus deltoides Bart. Ex Marsh.) and Their Relationships to Mechanical Properties* conducido por el Departamento de Bosques y Productos Forestales del Instituto Politécnico y Universidad Estatal de Virginia, experimento en el cual se seleccionaron al azar 29 pinos de incienso para investigarlos, se obtuvieron los datos que se presentan en la tabla 11.9, sobre la gravedad específica en gramos/cm^3 y el modulo de ruptura en kilopascales (kPa). Calcule e interprete el coeficiente de correlación muestral.

Tabla 11.9: Datos de 29 pinos de incienso para el ejemplo 11.10

Gravedad específica, x (g/cm^3)	**Módulo de ruptura, y (kPa)**	**Gravedad específica, x (g/cm^3)**	**Módulo de ruptura, y (kPa)**
0.414	29,186	0.581	85,156
0.383	29,266	0.557	69,571
0.399	26,215	0.550	84,160
0.402	30,162	0.531	73,466
0.442	38,867	0.550	78,610
0.422	37,831	0.556	67,657
0.466	44,576	0.523	74,017
0.500	46,097	0.602	87,291
0.514	59,698	0.569	86,836
0.530	67,705	0.544	82,540
0.569	66,088	0.557	81,699
0.558	78,486	0.530	82,096
0.577	89,869	0.547	75,657
0.572	77,369	0.585	80,490
0.548	67,095		

Solución: Con los datos se encuentra que

$$S_{xx} = 0.11273, \qquad S_{yy} = 11,807,324,805, \qquad S_{xy} = 34,422.27572.$$

Por lo tanto,

$$r = \frac{34,422.27572}{\sqrt{(0.11273)(11,807,324,805)}} = 0.9435.$$

Un coeficiente de correlación de 0.9435 indica una buena relación lineal entre X y Y. Como $r^2 = 0.8902$, puede decirse que aproximadamente el 89% de la variación de los valores de Y es ocasionada por la relación lineal con X.

Una prueba de la hipótesis especial $\rho = 0$ contra una alternativa apropiada es equivalente a probar $\beta = 0$ para el modelo de regresión lineal simple y, por lo tanto, son aplicables los procedimientos de la sección 11.8 donde se usaban tanto la distribución t con $n - 2$ grados de libertad o la distribución F con 1 y $n - 2$ grados de libertad. Sin embargo, si se desea evitar el procedimiento del análisis de varianza y tan sólo calcular el coeficiente de correlación muestral, puede verificarse (véase el ejercicio 11.51 en la página 438) que el valor t

$$t = \frac{b}{s/\sqrt{S_{xx}}}$$

también puede escribirse como

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

que, como antes, es un valor del estadístico T que tiene distribución t con $n - 2$ grados de libertad.

Ejemplo 11.11: Para los datos del ejemplo 11.10, pruebe la hipótesis de que no existe asociación lineal entre las variables.

Solución:

1. H_0: $\rho = 0$.

2. H_1: $\rho \neq 0$.

3. $\alpha = 0.05$.

4. Región crítica: $t < -2.052$ o $t > 2.052$.

5. Cálculos: $t = \frac{0.9435\sqrt{27}}{\sqrt{1-0.9435^2}} = 14.79$, $P < 0.0001$.

6. Decisión: Rechace la hipótesis de que no existe asociación lineal.

A partir de la información muestral, es fácil efectuar una prueba de la hipótesis más general de que $\rho = \rho_0$ contra una hipótesis alternativa. Si X y Y siguen una distribución normal bivariada, la cantidad

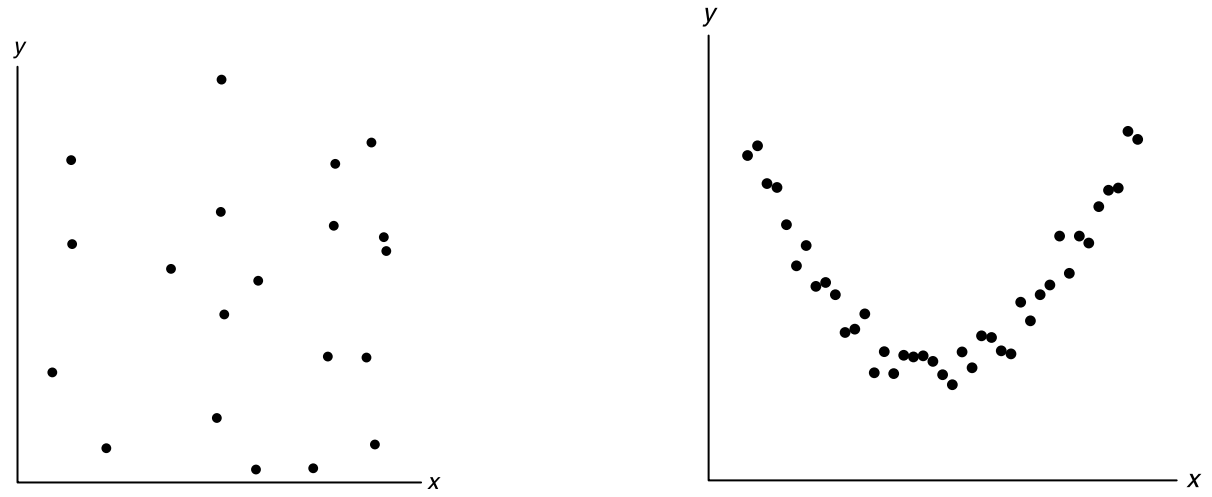
$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

es un valor de una variable aleatoria que sigue aproximadamente la distribución normal con media $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ y varianza $1/(n-3)$. Entonces, el procedimiento de prueba consiste en calcular

$$z = \frac{\sqrt{n-3}}{2}\left[\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right] = \frac{\sqrt{n-3}}{2}\ln\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right]$$

y compararlo con los puntos críticos de la distribución normal estándar.

Ejemplo 11.12: Para los datos del ejemplo 11.10, pruebe la hipótesis nula de que $\rho = 0.9$, contra la alternativa de que $\rho > 0.9$. Utilice un nivel de significancia de 0.05.



a) Sin asociación b) Relación causal

Figura 11.28: Diagrama de dispersión que muestra correlación de cero.

Solución: **1.** H_0: $\rho = 0.9$.

2. H_1: $\rho > 0.9$.

3. $\alpha = 0.05$.

4. Región crítica: $z > 1.645$.

5. Cálculos:

$$z = \frac{\sqrt{26}}{2} \ln \left[\frac{(1 + 0.9435)(0.1)}{(1 - 0.9435)(1.9)} \right] = 1.51, \quad P = 0.0655.$$

6. Decisión: Existe con certeza alguna evidencia de que el coeficiente de correlación no excede 0.9.

Debe precisarse que en los estudios de correlación, como en los problemas de regresión lineal, los resultados obtenidos sólo son tan buenos como el modelo que se adopte. En las técnicas de correlación estudiadas hasta aquí, se supone que las variables X y Y tienen una densidad normal bivariada, con el valor medio de Y para cada valor x relacionado en forma lineal con x. Con frecuencia es útil elaborar una gráfica preliminar de los datos experimentales para observar lo adecuado de la suposición de linealidad. Un valor del coeficiente de correlación muestral cercano a cero resultará de datos que muestren un efecto estrictamente aleatorio, como los de la figura 11.28*a*), lo que implica que hay poca o ninguna relación causal. Es importante recordar que el coeficiente de correlación entre dos variables es una medida de su relación lineal, y que un valor de $r = 0$ implica *falta de linealidad y no falta de asociación*. Por lo tanto, si existiera una relación cuadrática intensa entre X y Y, como la que se observa en la figura 11.28*b*), podría obtenerse una correlación de cero, que indicaría una relación no lineal.

Ejercicios

11.49 Calcule e interprete el coeficiente de correlación para las calificaciones siguientes de 6 estudiantes seleccionados al azar:

Calificación en matemáticas	70	92	80	74	65	83
Calificación en inglés	74	84	63	87	78	90

11.50 En el ejercicio 11.49 pruebe la hipótesis de que $\rho = 0$ contra la alternativa de que $\rho \neq 0$. Utilice un nivel de significancia de 0.05.

11.51 Demuestre los pasos necesarios para convertir la ecuación $r = \frac{b}{s/\sqrt{S_{xx}}}$ en la forma equivalente $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

11.52 Los datos siguientes se obtuvieron en un estudio de la relación entre el peso y el tamaño del pecho de niños recién nacidos:

Peso (kg)	Tamaño del pecho (kg)
2.75	29.5
2.15	26.3
4.41	32.2
5.52	36.5
3.21	27.2
4.32	27.7
2.31	28.3
4.30	30.3
3.71	28.7

a) Calcule r.

b) Pruebe la hipótesis nula de que $\rho = 0$ contra la alternativa de que $\rho > 0$, con un nivel de significancia de 0.01.

c) ¿Qué porcentaje de la variación de los tamaños del pecho de los niños está explicado por la diferencia de peso?

11.53 En relación con el ejercicio 11.1 de la página 397, suponga que x y y son variables aleatorias con distribución normal bivariada:

a) Calcule r.

b) Pruebe la hipótesis de que $\rho = 0$ contra la alternativa de que $\rho \neq 0$, con un nivel de significancia de 0.05.

11.54 Con relación al ejercicio 11.9 de la página 399, suponga una distribución normal bivariada para x y y.

a) Calcule r.

b) Pruebe la hipótesis nula de que $\rho = -0.5$, contra la alternativa de que $\rho < -0.5$, con un nivel de significancia de 0.025.

c) Determine el porcentaje de la variación en la cantidad de partículas removidas que se debe a cambios en la cantidad de lluvia diaria.

Ejercicios de repaso

11.55 Con referencia al ejercicio 11.6 de la página 398, construya

a) un intervalo de confianza de 95% para la calificación promedio en el curso de los estudiantes que obtuvieron 35 en el examen de colocación.

b) un intervalo de predicción de 95% para la calificación del curso de un estudiante que obtuvo 35 en el examen de colocación.

11.56 El Centro de Consulta Estadística del Instituto Politécnico y Universidad Estatal de Virginia analizó datos sobre las marmotas normales para el Departamento de Veterinaria. Las variables de interés fueron el peso corporal en gramos y el peso del corazón en gramos. También era de interés desarrollar una ecuación de regresión lineal, con la finalidad de determinar si había una relación lineal significativa entre el peso del corazón y el peso total del cuerpo. Utilice el peso del corazón como la variable independiente y el peso del cuerpo como la dependiente, y haga un ajuste de regresión lineal simple con los siguientes datos. Además, pruebe la hipótesis de que H_0: $\beta = 0$ contra H_1: $\beta \neq 0$. Saque conclusiones.

Peso corporal (gramos)	Peso del corazón (gramos)
4050	11.2
2465	12.4
3120	10.5
5700	13.2
2595	9.8
3640	11.0
2050	10.8
4235	10.4
2935	12.2
4975	11.2
3690	10.8
2800	14.2
2775	12.2
2170	10.0
2370	12.3
2055	12.5
2025	11.8
2645	16.0
2675	13.8

11.57 A continuación se presentan las cantidades de sólidos removidos de cierto material cuando se expone a periodos de secado de diferentes duraciones.

x (horas)	y (gramos)	
4.4	13.1	14.2
4.5	9.0	11.5
4.8	10.4	11.5
5.5	13.8	14.8
5.7	12.7	15.1
5.9	9.9	12.7
6.3	13.8	16.5
6.9	16.4	15.7
7.5	17.6	16.9
7.8	18.3	17.2

a) Estime la recta de regresión lineal.

b) Pruebe si es adecuado el modelo lineal, con un nivel de significancia de 0.05.

11.58 Con referencia al ejercicio 11.7 de la página 399, construya

a) un intervalo de confianza de 95% para las ventas semanales promedio, cuando se gastan \$45 en publicidad.

b) un intervalo de predicción para las ventas semanales cuando se gastan \$45 en publicidad.

11.59 Se diseñó un experimento para el Departamento de Ingeniería de Materiales del Instituto Politécnico y Universidad estatal de Virginia, para estudiar las propiedades de fragilidad del nitrógeno con base en las mediciones de la presión de hidrógeno electrolítico. Se utilizó una solución al 0.1 N NaOH, material que es un tipo de acero inoxidable. La densidad de corriente de carga catódica fue controlada y variada en cuatro niveles. Se observó la presión de hidrógeno efectiva, así como la respuesta. A continuación se presentan los datos.

Experimento	Densidad de corriente de carga, x (mA/cm^2)	Presión de hidrógeno efectiva, y (atm)
1	0.5	86.1
2	0.5	92.1
3	0.5	64.7
4	0.5	74.7
5	1.5	223.6
6	1.5	202.1
7	1.5	132.9
8	2.5	413.5
9	2.5	231.5
10	2.5	466.7
11	2.5	365.3
12	3.5	493.7
13	3.5	382.3
14	3.5	447.2
15	3.5	563.8

a) Efectúe un análisis de regresión lineal simple de y contra x.

b) Calcule la suma cuadrática del error puro y haga una prueba para la falta de ajuste.

c) ¿La información del inciso *b*) indica la necesidad de un modelo en x más allá del de la regresión de primer orden? Explique su respuesta.

11.60 Los datos siguientes representan la calificación en química de una muestra aleatoria de 12 personas de nuevo ingreso a cierta escuela, así como sus calificaciones en una prueba de inteligencia aplicada mientras aún no egresaban de la escuela:

Estudiante	Calificación en la prueba, x	Calificación en química, y
1	65	85
2	50	74
3	55	76
4	65	90
5	55	85
6	70	87
7	65	94
8	70	98
9	55	81
10	70	91
11	50	76
12	55	74

a) Calcule e interprete el coeficiente de correlación de la muestra.

b) Establezca las suposiciones necesarias acerca de las variables aleatorias.

c) Pruebe la hipótesis de que $\rho = 0.5$, contra la alternativa de que $\rho > 0.5$. En la conclusión use un valor P.

11.61 Para el modelo de regresión lineal simple, demuestre que $E(s^2) = \sigma^2$.

11.62 La sección de negocios del *Washington Times* de marzo de 1997 listaba 21 diferentes computadoras e impresoras usadas, así como sus precios de lista. También se listaba el ofrecimiento promedio. En la figura 11.29 de la página 440 se presentan los resultados parciales del análisis de regresión usando el software *SAS*.

a) Explique la diferencia entre el intervalo de confianza sobre la media y el intervalo de predicción.

b) Explique por qué los errores estándar de la predicción varían de una observación a otra.

c) ¿Cuál observación tiene el menor error estándar de la predicción? ¿Por qué?

11.63 Considere los datos de los vehículos de la figura 11.30 de *Consumer Reports*. También se indican el peso en toneladas, el rendimiento en millas por galón y la razón de manejo. Se ajustó un modelo de regresión que relacionaba el peso x con el rendimiento y. En la figura 11.30 de la página 441 se presenta una salida

parcial *SAS* que muestra algunos de los resultados de dicho análisis de regresión, y en la figura 11.31 de la página 442 se ilustra la gráfica de los residuos y el peso de cada vehículo.

a) Del análisis y la gráfica de los residuos, ¿parece que pudiera encontrarse un modelo mejorado si se usara una transformación? Explique su respuesta.

b) Ajuste el modelo por medio de reemplazar el peso con el logaritmo del peso. Comente los resultados.

c) Ajuste un modelo por medio de reemplazar mpg con los galones por 100 millas recorridas, que es como con frecuencia se reporta el rendimiento en otros países. ¿Cuál de los tres modelos es preferible? Explique su respuesta.

R-Square	Coeff Var	Root MSE	Price Mean
0.967472	7.923338	70.83841	894.0476

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	59.93749137	38.34195754	1.56	0.1345
Buyer	1.04731316	0.04405635	23.77	<.0001

product	Buyer	Price	Predict Value	Std Err Predict	Lower 95% Mean	Upper 95% Mean	Lower 95% Predict	Upper 95% Predict
IBM PS/1 486/66 420MB	325	375	400.31	25.8906	346.12	454.50	242.46	558.17
IBM ThinkPad 500	450	625	531.23	21.7232	485.76	576.70	376.15	686.31
IBM Think-Dad 755CX	1700	1850	1840.37	42.7041	1750.99	1929.75	1667.25	2013.49
AST Pentium 90 540MB	800	875	897.79	15.4590	865.43	930.14	746.03	1049.54
Dell Pentium 75 1GB	650	700	740.69	16.7503	705.63	775.75	588.34	893.05
Gateway 486/75 320MB	700	750	793.06	16.0314	759.50	826.61	641.04	945.07
Clone 586/133 1GB	500	600	583.59	20.2363	541.24	625.95	429.40	737.79
Compaq Contura 4/25 120MB	450	600	531.23	21.7232	485.76	576.70	376.15	686.31
Compaq Deskpro P90 1.2GB	800	850	897.79	15.4590	865.43	930.14	746.03	1049.54
Micron P75 810MB	800	675	897.79	15.4590	865.43	930.14	746.03	1049.54
Micron P100 1.2GB	900	975	1002.52	16.1176	968.78	1036.25	850.46	1154.58
Mac Quadra 840AV 500MB	450	575	531.23	21.7232	485.76	576.70	376.15	686.31
Mac Performer 6116 700MB	700	775	793.06	16.0314	759.50	826.61	641.04	945.07
PowerBook 540c 320MB	1400	1500	1526.18	30.7579	1461.80	1590.55	1364.54	1687.82
PowerBook 5300 500MB	1350	1575	1473.81	28.8747	1413.37	1534.25	1313.70	1633.92
Power Mac 7500/100 1GB	1150	1325	1264.35	21.9454	1218.42	1310.28	1109.13	1419.57
NEC Versa 486 340MB	800	900	897.79	15.4590	865.43	930.14	746.03	1049.54
Toshiba 1960CS 320MB	700	825	793.06	16.0314	759.50	826.61	641.04	945.07
Toshiba 4800VCT 500MB	1000	1150	1107.25	17.8715	1069.85	1144.66	954.34	1260.16
HP Laser jet III	350	475	426.50	25.0157	374.14	478.86	269.26	583.74
Apple Laser Writer Pro 63	750	800	845.42	15.5930	812.79	878.06	693.61	997.24

Figura 11.29: Salida *SAS* que muestra el análisis parcial de datos del ejercicio de repaso 11.62.

11.64 A continuación se presentan las observaciones registradas del producto de una reacción química tomadas a temperaturas diferentes:

x (°C)	y (%)	x (°C)	y (%)
150	75.4	150	77.7
150	81.2	200	84.4
200	85.5	200	85.7
250	89.0	250	89.4
250	90.5	300	94.8
300	96.7	300	95.3

a) Grafique los datos.

b) De la gráfica, ¿pareciera que la relación es lineal?

c) Haga un análisis de regresión lineal simple y pruebe la falta de ajuste.

d) Saque conclusiones con base en el resultado del inciso *c*).

11.65 La prueba de acondicionamiento físico es un aspecto importante del entrenamiento atlético. Una medida común de la magnitud de la aptitud cardiovascular es el volumen máximo de oxígeno inhalado durante un ejercicio extenuante. Se realizó un estudio a 24 hombres de edad madura para analizar la influencia del tiempo que les tomaba correr una distancia de dos millas. La medición del oxígeno inhalado se complementó con métodos estándar de laboratorio mientras los sujetos estaban en su rutina. El trabajo se publicó en "Maximal Oxygen Intake Prediction in Young and Middle Aged Males", *Journal of Sports Medicine* **9**, 1969, 17-22. A continuación se presentan los datos.

Sujeto	y, Volumen máximo de O_2	x, Tiempo en segundos
1	42.33	918
2	53.10	805
3	42.08	892
4	50.06	962
5	42.45	968

Obs	Model	WT	MPG	DR_RATIO
1	Buick Estate Wagon	4.360	16.9	2.73
2	Ford Country Squire Wagon	4.054	15.5	2.26
3	Chevy Ma libu Wagon	3.605	19.2	2.56
4	Chrysler LeBaron Wagon	3.940	18.5	2.45
5	Chevette	2.155	30.0	3.70
6	Toyota Corona	2.560	27.5	3.05
7	Datsun 510	2.300	27.2	3.54
8	Dodge Omni	2.230	30.9	3.37
9	Audi 5000	2.830	20.3	3.90
10	Volvo 240 CL	3.140	17.0	3.50
11	Saab 99 GLE	2.795	21.6	3.77
12	Peugeot 694 SL	3.410	16.2	3.58
13	Buick Century Special	3.380	20.6	2.73
14	Mercury Zephyr	3.070	20.8	3.08
15	Dodge Aspen	3.620	18.6	2.71
16	AMC Concord D/L	3.410	18.1	2.73
17	Chevy Caprice Classic	3.840	17.0	2.41
18	Ford LTP	3.725	17.6	2.26
19	Mercury Grand Marquis	3.955	16.5	2.26
20	Dodge St Regis	3.830	18.2	2.45
21	Ford Mustang 4	2.585	26.5	3.08
22	Ford Mustang Ghia	2.910	21.9	3.08
23	Macda GLC	1.975	34.1	3.73
24	Dodge Colt	1.915	35.1	2.97
25	AMC Spirit	2.670	27.4	3.08
26	VW Scirocco	1.990	31.5	3.78
27	Honda Accord LX	2.135	29.5	3.05
28	Buick Skylark	2.570	28.4	2.53
29	Chevy Citation	2.595	28.8	2.69
30	Olds Omega	2.700	26.8	2.84
31	Pontiac Phoenix	2.556	33.5	2.69
32	Plymouth Horizon	2.200	34.2	3.37
33	Datsun 210	2.020	31.8	3.70
34	Fiat Strada	2.130	37.3	3.10
35	VW Dasher	2.190	30.5	3.70
36	Datsun 810	2.815	22.0	3.70
37	BMW 320i	2.600	21.5	3.64
38	VW Rabbit	1.925	31.9	3.78

R-Square	Coeff Var	Root MSE	MPG Mean
0.817244	11.46010	2.837580	24.76053

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	48.67928080	1.94053995	25.09	<.0001
WT	-8.36243141	0.65908398	-12.69	<.0001

Figura 11.30: Salida *SAS* que muestra el análisis parcial de los datos del ejercicio de repaso 11.63.

Sujeto	y, Volumen máximo de O_2	x, Tiempo en segundos
6	42.46	907
7	47.82	770
8	49.92	743
9	36.23	1045
10	49.66	810
11	41.49	927
12	46.17	813
13	46.18	858
14	43.21	860
15	51.81	760
16	53.28	747
17	53.29	743
18	47.18	803
19	56.91	683
20	47.80	844
21	48.65	755
22	53.67	700
23	60.62	748
24	56.73	775

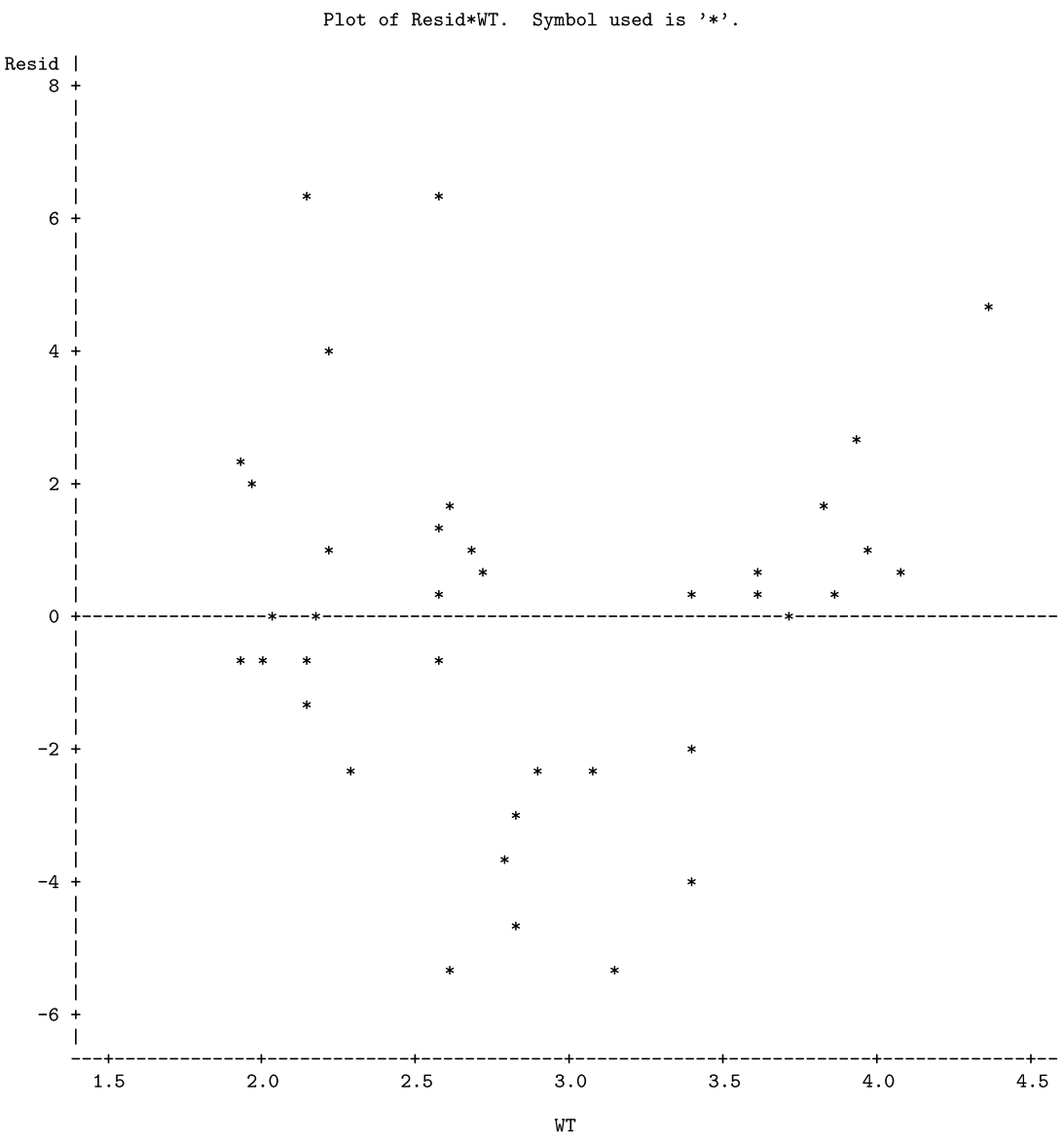


Figura 11.31: Salida *SAS* que muestra la gráfica de residuos del ejercicio de repaso 11.63.

a) Estime los parámetros en un modelo de regresión lineal simple.

b) ¿El tiempo que toma correr dos millas tiene influencia significativa sobre el máximo oxígeno inspirado? Utilice

$$H_0\text{: } \beta = 0,$$
$$H_1\text{: } \beta \neq 0.$$

c) Grafique los residuos en una gráfica contra x, y haga comentarios sobre lo apropiado del modelo lineal simple.

11.66 Suponga que cierto científico postula el modelo

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

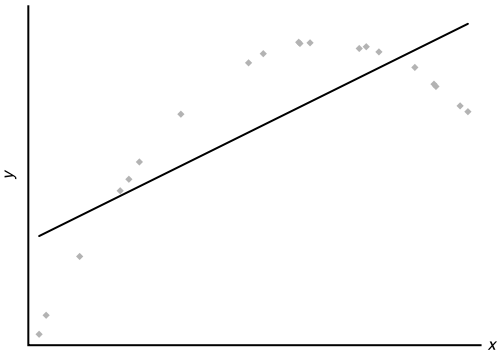
y α es un **valor conocido** no necesariamente igual a cero.

a) ¿Cuál es el estimador apropiado de mínimos cuadrados de β? Justifique su respuesta.

b) ¿Cuál es la varianza del estimador de la pendiente?

11.67 En el ejercicio 11.30 de la página 413 se pidió al estudiante que demostrara que $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$ para un modelo de regresión lineal simple. ¿Se cumple también para un modelo con intersección en el origen? Demuestre por qué.

11.68 Considere el conjunto de datos imaginarios que se muestra en seguida, donde la línea que pasa a través de los datos es la recta de regresión lineal simple. Dibuje una gráfica de residuos para el modelo ajustado anteriormente.



11.13 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

Siempre que se considere la utilización de la regresión lineal simple, elaborar una gráfica de los datos no sólo es recomendable, sino esencial. Siempre es edificante elaborar una gráfica de los residuos, tanto con la distribución t de Student, como la de probabilidad normal de ellos. Todas esas gráficas están diseñadas para detectar la trasgresión de las suposiciones.

El uso de los estadísticos t para las pruebas sobre los coeficientes de regresión es razonablemente robusto para la suposición de normalidad. La suposición de varianza homogénea es crucial, y las gráficas de los residuos están diseñadas para detectar la trasgresión de ella.

Capítulo 12

Regresión lineal múltiple y ciertos modelos de regresión no lineal

12.1 Introducción

En la mayoría de problemas de investigación donde se aplica el análisis de regresión, se necesita más de una variable independiente en el modelo de regresión. La complejidad de la mayoría de mecanismos científicos es tal que, con la finalidad de predecir una respuesta importante, se requiere un **modelo de regresión múltiple**. Cuando este modelo es lineal en los coeficientes se denomina **modelo de regresión lineal múltiple**. Para el caso de k variables independientes, x_1, $x_2, \ldots,$ x_k, la media de $Y|\, x_1, x_2, \ldots, x_k$ está dada por el modelo de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

y la respuesta estimada se obtiene a partir de la ecuación de regresión muestral

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k,$$

donde cada coeficiente de regresión β_i es estimado por b_i de los datos muestrales usando el método de los mínimos cuadrados. Como en el caso de una sola variable independiente, es frecuente que el modelo de regresión lineal sea una representación adecuada de una estructura más complicada dentro de ciertos rangos de las variables independientes.

También pueden aplicarse técnicas similares de mínimos cuadrados para estimar los coeficientes cuando el modelo lineal incluye, por ejemplo, potencias y productos de las variables independientes. Por ejemplo, cuando $k = 1$, el experimentador quizá sienta que la media $\mu_{Y|x}$ no cae sobre una línea recta, sino que queda descrita con más propiedad por el **modelo de regresión polinomial**

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r,$$

y la respuesta estimada se obtenga de la ecuación de regresión polinomial

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_r x^r.$$

En ocasiones se genera confusión al decir que un modelo polinomial es uno lineal. Sin embargo, los estadísticos normalmente se refieren a un modelo lineal como aquel donde los parámetros ocurren en forma lineal, sin importar la manera en que las variables independientes aparezcan en el modelo. Un ejemplo de modelo no lineal es la **relación exponencial**

$$\mu_{Y|x} = \alpha\beta^x,$$

que se estima mediante la ecuación de regresión

$$\hat{y} = ab^x.$$

En ciencias e ingeniería hay muchos fenómenos de naturaleza inherentemente no lineal y, cuando se conoce la estructura verdadera, no hay duda de que debería intentarse ajustar el modelo real. Es muy abundante la bibliografía acerca de la estimación con mínimos cuadrados de modelos no lineales. Aun cuando en este libro no se trata de cubrir en forma rigurosa la regresión no lineal, en la sección 12.12 estudiaremos ciertos tipos específicos de modelos no lineales. Los modelos no lineales que se analizan en este capítulo se refieren a condiciones no ideales, en las cuales el analista está seguro de que la respuesta y, por lo tanto, la respuesta del error del modelo, no tienen distribución normal, sino que más bien siguen una binomial o una de Poisson. Estas situaciones ocurren mucho en la práctica.

El estudiante que desee un estudio más general de la regresión no lineal debe consultar la obra de Myers *Classical and Modern Regression with Applications* (véase la bibliografía).

12.2 Estimación de los coeficientes

En esta sección se obtienen los estimadores de mínimos cuadrados de los parámetros β_0, $\beta_1, \ldots,$ β_k mediante el ajuste del modelo de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

a los puntos de los datos

$$\{(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i), \quad i = 1, 2, \ldots, n \text{ y } n > k\},$$

donde y_i, es la respuesta observada a los valores x_{1i}, $x_{2i}, \ldots,$ x_{ki} de las k variables independientes x_1, $x_2, \ldots,$ x_k. Se supone que cada observación (x_{1i}, $x_{2i}, \ldots,$ x_{ki}, y_i) satisface la siguiente ecuación:

Modelo de regresión lineal múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i,$$

o bien,

$$\boldsymbol{y}_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,$$

donde ϵ_i y e_i son los errores aleatorio y residual, respectivamente, asociados con la respuesta y_i y con el valor ajustado $\hat{y}_i$.

Como en el caso de la regresión lineal simple, se supone que los ϵ_i son independientes, y están distribuidos en forma idéntica con media cero y varianza común σ^2.

Tabla 12.1: Datos para el Ejemplo 12.1

Óxido nítrico, y	**Humedad, x_1**	**Temp., x_2**	**Presión x_3**	**Óxido nítrico, y**	**Humedad, x_1**	**Temp., x_2**	**Presión x_3**
0.90	72.4	76.3	29.18	1.07	23.2	76.8	29.38
0.91	41.6	70.3	29.35	0.94	47.4	86.6	29.35
0.96	34.3	77.1	29.24	1.10	31.5	76.9	29.63
0.89	35.1	68.0	29.27	1.10	10.6	86.3	29.56
1.00	10.7	79.0	29.78	1.10	11.2	86.0	29.48
1.10	12.9	67.4	29.39	0.91	73.3	76.3	29.40
1.15	8.3	66.8	29.69	0.87	75.4	77.9	29.28
1.03	20.1	76.9	29.48	0.78	96.6	78.7	29.29
0.77	72.2	77.7	29.09	0.82	107.4	86.8	29.03
1.07	24.0	67.7	29.60	0.95	54.9	70.9	29.37

Fuente: Charles T. Hare, "Light-Duty Diesel Emission Correction Factors for Ambient Conditions", EPA-600/2-77-116. U. S. Environmental Protection Agency.

Al usar el concepto de mínimos cuadrados para obtener los estimadores $b_0, b_1, \ldots, b_k$, minimizamos la expresión

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2.$$

Que se deriva con respecto de $b_0, b_1, \ldots, b_k$, para igualar el resultado a cero y generar el conjunto de $k + 1$ **ecuaciones normales de estimación para la regresión lineal múltiple**.

Ecuaciones normales de estimación para la regresión lineal múltiple

$$\begin{aligned}
nb_0 + b_1 \sum_{i=1}^{n} x_{1i} \quad &+b_2 \sum_{i=1}^{n} x_{2i} \quad + \cdots + b_k \sum_{i=1}^{n} x_{ki} \quad = \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 \quad &+b_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i \\
\vdots \qquad \vdots \qquad &\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
b_0 \sum_{i=1}^{n} x_{ki} + b_1 \sum_{i=1}^{n} x_{ki}x_{1i} &+b_2 \sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{ki}^2 \quad = \sum_{i=1}^{n} x_{ki}y_i
\end{aligned}$$

Para obtener los valores de $b_0, b_1, \ldots, b_k$ estas ecuaciones se resuelven con cualquier método apropiado para sistemas de ecuaciones lineales.

Ejemplo 12.1: Se realizó un estudio acerca de camiones ligeros movidos por diesel para saber si la humedad, la temperatura del aire y la presión barométrica influían en la emisión de óxido nitroso (en ppm). Las emisiones se midieron a distintas horas, en condiciones experimentales diversas. En la tabla 12.1 se presentan los datos. El modelo es

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

o, en forma equivalente,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \quad i = 1, 2, \ldots, 20.$$

Ajuste un modelo de regresión lineal múltiple a los datos dados y luego estime la cantidad de óxido nitroso para las condiciones en que la humedad es de 50%, temperatura de 76 °F y presión barométrica de 29.30.

Solución: La solución del conjunto de las ecuaciones de estimación da los estimadores únicos

$$b_0 = -3.507778,\ b_1 = -0.002625,\ b_2 = 0.000799,\ b_3 = 0.154155.$$

Por lo tanto, la ecuación de regresión es

$$\hat{y} = -3.507778 - 0.002625\, x_1 + 0.000799\, x_2 + 0.154155\, x_3.$$

Para una humedad de 50%, temperatura de 76 °F y presión barométrica de 29.30, la cantidad estimada de óxido nitroso es

$$\begin{aligned}\hat{y} &= -3.507778 - 0.002625(50.0) + 0.000799(76.0) + 0.1541553(29.30)\\ &= 0.9384 \text{ ppm}.\end{aligned}$$

Regresión polinomial

Ahora, suponga que se desea ajustar la ecuación polinomial

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r$$

para las n parejas de observaciones $\{(x_i,\ y_i);\ \ i = 1,\ 2, \ldots,\ n\}$. Cada observación, y_i, satisface la ecuación

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_r x_i^r + \epsilon_i$$

o bien,

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_r x_i^r + e_i,$$

donde r es el grado del polinomio, y ϵ_i y e_i son, de nuevo, los errores aleatorio y residual asociados con la respuesta y_i y con el valor ajustado $\hat{y}$, respectivamente. Aquí, el número de parejas, n, debe ser al menos $r + 1$, que es el número de parámetros por estimar.

Observe que el modelo polinomial puede considerarse un caso especial del modelo de regresión lineal más general, donde se hace $x_1 = x$, $x_2 = x^2, \ldots,\ x_r = x^r$. Las ecuaciones normales adoptan la misma forma que se da en la página 447. Luego se resuelven para b_0, b_1, $b_2, \ldots,$ b_r.

Ejemplo 12.2: Dados los datos

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2

ajuste una curva de regresión de la forma $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$ y, luego, estime $\mu_{Y|2}$.

Solución: De los datos, se encuentra que

$$\begin{aligned} 10\,b_0 + 45\,b_1 \quad &+ 285\,b_2 \quad &= 63.7, \\ 45\,b_0 + 285\,b_1 \quad &+ 2,025\,b_2 \quad &= 307.3, \\ 285\,b_0 + 2{\cdot}025\,b_1 &+ 15,333\,b_2 &= 2153.3. \end{aligned}$$

Al resolver las ecuaciones normales se obtiene

$$b_0 = 8.698, \quad b_1 = -2.341, \quad b_2 = 0.288.$$

Por lo tanto,

$$\hat{y} = 8.698 - 2.341\,x + 0.288\,x^2.$$

Cuando $x = 2$, la estimación de $\mu_{Y|2}$, es

$$\hat{y} = 8.698 - (2.341)(2) + (0.288)(2^2) = 5.168.$$

12.3 Modelo de regresión lineal con el empleo de matrices (opcional)

Al ajustar un modelo de regresión lineal múltiple, en particular cuando el número de variables es mayor que dos, el dominio de la teoría de matrices facilita en forma considerable las manipulaciones matemáticas. Suponga que el experimentador tiene k variables independientes $x_1, x_2, \ldots, x_k$ y n observaciones $y_1, y_2, \ldots, y_n$, cada una de las cuales puede expresarse con la ecuación

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i.$$

Este modelo representa en esencia a n ecuaciones que describen cómo se generan los valores de la respuesta durante el proceso científico. Con notación matricial, se escribe la ecuación siguiente

Modelo lineal general

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Después, la solución por mínimos cuadrados para la estimación de β que se ilustró en la sección 12.2, implica el cálculo de $\mathbf{b}$ para la que

$$SSE = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

es mínima. Este proceso de minimización implica resolver para $\mathbf{b}$ la ecuación

$$\frac{\partial}{\partial \mathbf{b}}(SSE) = \mathbf{0}.$$

Aquí no presentaremos los detalles de las soluciones de las ecuaciones anteriores. El resultado se reduce a la solución de $\mathbf{b}$ en

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}.$$

Observe la naturaleza de la matriz $\mathbf{X}$. Además del elemento inicial, el i-ésimo renglón representa los valores de x que dan lugar a la respuesta y_i. Queda:

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}$$

y

$$\mathbf{g} = \mathbf{X}'\mathbf{y} = \begin{bmatrix} g_0 = \sum_{i=1}^{n} y_i \\ g_1 = \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ g_k = \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix},$$

las ecuaciones normales pueden escribirse en forma matricial como

$$\mathbf{A}\mathbf{b} = \mathbf{g}.$$

Si la matriz $\mathbf{A}$ es no singular, la solución para los coeficientes de regresión se escribe como

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{g} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Así, obtenemos la ecuación de predicción o regresión al resolver un conjunto de $k + 1$ ecuaciones con un número igual de incógnitas. Esto implica que se invierta la matriz $\mathbf{X}'\mathbf{X}$ de orden $k + 1$ por $k + 1$. En la mayoría de libros sobre determinantes y matrices elementales se explican las técnicas para invertir matrices. Por supuesto, hay disponibles muchos paquetes rápidos de computadora para resolver problemas de regresión múltiple, los cuales no sólo dan salida de las estimaciones de los coeficientes de regresión, sino que también ofrecen otra clase de información relevante para hacer inferencias acerca de la ecuación de regresión.

Ejemplo 12.3: Se midió el porcentaje de supervivencia de cierto tipo de semen animal, después de almacenarlo, con distintas combinaciones de concentraciones de tres materiales que se emplean para incrementar su probabilidad de sobrevivir. En la tabla 12.2 se presentan los datos. Obtenga el modelo de regresión lineal múltiple para los datos.

Tabla 12.2: Datos para el ejemplo 12.3

y (porcentaje de supervivencia)	x_1 (porcentaje de peso)	x_2 (porcentaje de peso)	x_3 (porcentaje de peso)
25.5	1.74	5.30	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.20
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90
26.4	4.10	6.62	8.00
25.9	6.32	8.72	9.10
32.0	4.08	4.42	8.70
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.7	1.72	3.12	7.60
26.5	1.70	5.30	8.20

Solución: Las ecuaciones de estimación por mínimos cuadrados, $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$, son

$$\begin{bmatrix} 13 & 59.43 & 81.82 & 115.40 \\ 59.43 & 394.7255 & 360.6621 & 522.0780 \\ 81.82 & 360.6621 & 576.7264 & 728.3100 \\ 115.40 & 522.0780 & 728.3100 & 1035.9600 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}.$$

Con una computadora se obtienen los elementos de la matriz inversa

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0037 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0037 & -0.0021 & 0.0886 \end{bmatrix},$$

y, luego, con la relación $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ se llega a que los coeficientes de regresión estimados son

$$b_0 = 39.1574,\ b_1 = 1.0161,\ b_2 = -1.8616,\ b_3 = -0.3433.$$

Entonces, la ecuación de regresión estimada es

$$\hat{y} = 39.1574 + 1.0161\,x_1 - 1.8616\,x_2 - 0.3433\,x_3.$$

Ejemplo 12.4: Los datos de la tabla 12.3 representan el porcentaje de impurezas que ocurren a distintas temperaturas y tiempos de esterilización durante una reacción asociada con la manufactura de cierta bebida.

Estime los coeficientes de regresión en el modelo polinomial

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

para $i = 1, 2, \ldots, 18$.

Tabla 12.3: Datos para el ejemplo 12.4

Tiempo de esterilización, x_2 (min)	**Temperatura, x_1 (°C)** 75	100	125
15	14.05	10.55	7.55
	14.93	9.48	6.59
20	16.56	13.63	9.23
	15.85	11.75	8.78
25	22.41	18.55	15.93
	21.66	17.98	16.44

Solución: $b_0 = 56.4411, \quad b_1 = -0.36190, \quad b_2 = -2.75299,$
$b_{11} = 0.00081, \quad b_{22} = 0.08173, \quad b_{12} = 0.00314,$

y la ecuación estimada de regresión es

$$\hat{y} = 56.4411 - 0.36190x_1 - 2.75299x_2 + 0.00081x_1^2 + 0.08173x_2^2 + 0.00314x_1x_2.$$

Muchos de los principios y procedimientos asociados con la estimación de funciones de regresión polinomiales caen en la categoría de la **metodología de respuesta superficial**, que es un conjunto de técnicas que emplean con éxito los científicos e ingenieros de muchos campos. **Las x_i^2 se denominan términos cuadráticos puros, y las x_ix_j $(i \neq j)$ se llaman términos de interacción**. Problemas tales como seleccionar un diseño experimental adecuado, en particular en casos donde un número muy grande de variables entra en el modelo, y elegir condiciones "óptimas" de operación sobre $x_1, x_2, \ldots, x_k$ con frecuencia se abordan a través de dichos métodos. Para un estudio más amplio, se recomienda al lector recurrir a la obra de Myers y Montgomery *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (véase la bibliografía).

Ejercicios

12.1 Para el ejercicio de repaso 11.60 de la página 439, suponga que también se da el número de periodos de clase perdidos por los 12 estudiantes que toman el curso de química. A continuación se presentan los datos completos.

Estudiante	**Calificación en química, y**	**Calificación en el examen, x_1**	**Clases perdidas, x_2**
1	85	65	1
2	74	50	7
3	76	55	5
4	90	65	2
5	85	55	6
6	87	70	3
7	94	65	2
8	98	70	5
9	81	55	4
10	91	70	3
11	76	50	1
12	74	55	4

a) Ajuste una ecuación de regresión lineal múltiple de la forma $\hat{y} = b_0 + b_1x_1 + b_2x_2$.

b) Estime la calificación en química para un estudiante que obtuvo una calificación de 60 en el examen de inteligencia y perdió 4 clases.

12.2 En *Applied Spectroscopy* se estudian las propiedades de reflectancia infrarroja de un líquido viscoso

utilizado en la industria electrónica como lubricante. El experimento que se diseñó consistió en el efecto de frecuencia de banda x_1 y espesor de película x_2 sobre la densidad óptica y usando un espectrómetro infrarrojo Perkin-Elmer Modelo 621. [*Fuente:* Pachansky, J., England, C. D., y Wattman, R. "Infrarred spectroscopic studies of poly (perflouropropyleneoxide) on gold substrate. A classical dispersion analysis for the refractive index." *Applied Spectroscopy*, vol. 40, núm. 1, enero de 1986, p. 9, table 1.]

y	x_1	x_2
0.231	740	1.10
0.107	740	0.62
0.053	740	0.31
0.129	805	1.10
0.069	805	0.62
0.030	805	0.31
1.005	980	1.10
0.559	980	0.62
0.321	980	0.31
2.948	1,235	1.10
1.633	1,235	0.62
0.934	1,235	0.31

Estime la ecuación de regresión lineal múltiple

$$\hat{y} = b_0 + b_1x_1 + b_2x_2.$$

12.3 Se efectuó un conjunto de ensayos experimentales para determinar una forma de predecir el tiempo de cocción y a diferentes niveles del ancho de horno x_1 y temperaturas de la chimenea x_2. Los siguientes son los datos registrados:

y	x_1	x_2
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.35	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

Estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1x_1 + \beta_2x_2.$$

12.4 Se realizó un experimento para determinar si podía predecirse el peso de un animal después de un periodo dado, sobre la base de su peso inicial y la cantidad de alimento que había consumido. Se registraron los siguientes datos, en kilogramos:

Peso final, y	**Peso inicial, x_1**	**Peso del alimento, x_2**
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

a) Ajuste una ecuación de regresión múltiple de la forma

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1x_1 + \beta_2x_2.$$

b) Prediga el peso final de un animal que tenía un peso inicial de 35 kilogramos y consumió 250 kilogramos de alimento.

12.5 *a*) Ajuste una ecuación de regresión múltiple de la forma $\mu_{Y|x} = \beta_0 + \beta_1x_1 + \beta_2x^2$ a los datos del ejemplo 11.8.

b) Estime el producto de la reacción química para una temperatura de 225 °C.

12.6 Se efectuó un experimento sobre un modelo nuevo de una marca de automóvil específica, para determinar la distancia de frenado a distintas velocidades. A continuación se presentan los datos registrados.

Velocidad, v (km/h)	35	50	65	80	95	110
Distancia de frenado, d(m)	16	26	41	62	88	119

a) Ajuste una curva de regresión múltiple de la forma $\mu_{D|v} = \beta_0 + \beta_1v + \beta_2v^2$.

b) Estime la distancia de frenado cuando el carro viaje a 70 kilómetros por hora.

12.7 Se efectuó un experimento con la finalidad de determinar si el flujo sanguíneo cerebral de los seres humanos podía predecirse a partir de la tensión arterial del oxígeno (milímetros de mercurio). En el estudio se utilizaron 15 pacientes y se observaron los siguientes datos:

Flujo sanguíneo, y	**Tensión arterial del oxígeno, x**
84.33	603.40
87.80	582.50
82.20	556.20
78.21	594.60
78.44	558.90
80.01	575.20
83.53	580.10
79.46	451.20
75.22	404.00
76.58	484.00
77.90	452.40
78.80	448.40
80.67	334.80
86.60	320.30
78.20	350.30

Estime la ecuación de regresión cuadrática

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2.$$

12.8 El siguiente es un conjunto de datos experimentales codificados acerca de la resistencia a la compresión de una aleación específica, para valores distintos de la concentración de cierto aditivo:

Concentración, x	**Resistencia a la compresión, y**		
10.0	25.2	27.3	28.7
15.0	29.8	31.1	27.8
20.0	31.2	32.6	29.7
25.0	31.7	30.1	32.3
30.0	29.4	30.8	32.8

a) Estime la ecuación de regresión cuadrática $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$.

b) Pruebe la falta de ajuste del modelo.

12.9 Se cree que la energía eléctrica consumida cada mes por una planta química está relacionada con la temperatura ambiental promedio x_1, el número de días del mes x_2, la pureza promedio del producto x_3 y las toneladas fabricadas del producto x_4. Se dispone de datos históricos que se presentan en la siguiente tabla.

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

a) Ajuste un modelo de regresión lineal múltiple usando el conjunto de los datos anteriores.

b) Prediga el consumo de energía para un mes en que $x_1 = 75$ °F, $x_2 = 24$ días, $x_3 = 90\%$ y $x_4 = 98$ toneladas.

12.10 Para los datos siguientes

x	0	1	2	3	4	5	6
y	1	4	5	3	2	3	4

a) Ajuste el modelo cúbico $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

b) Prediga el valor de Y cuando $x = 2$.

12.11 El departamento de personal de cierta compañía industrial utilizó a 15 sujetos en un estudio, con la finalidad de determinar la relación entre la calificación de su desempeño en el trabajo (y) y las calificaciones de cuatro exámenes. Los datos son los siguientes:

y	x_1	x_2	x_3	x_4
11.2	56.5	71.0	38.5	43.0
14.5	59.5	72.5	38.2	44.8
17.2	69.2	76.0	42.5	49.0
17.8	74.5	79.5	43.4	56.3
19.3	81.2	84.0	47.5	60.2
24.5	88.0	86.2	47.4	62.0
21.2	78.2	80.5	44.5	58.1
16.9	69.0	72.0	41.8	48.1
14.8	58.1	68.0	42.1	46.0
20.0	80.5	85.0	48.1	60.3
13.2	58.3	71.0	37.5	47.1
22.5	84.0	87.2	51.0	65.2

Estime los coeficientes de regresión del modelo

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4.$$

12.12 Los siguientes datos reflejan información obtenida en 17 hospitales de la marina estadounidense en varios sitios del mundo. Los regresores son variables de la carga de trabajo, es decir, conceptos que daban como resultado la necesidad de personal en una instalación hospitalaria. A continuación se presenta una descripción breve de las variables:

y = horas-trabajo mensuales,
x_1 = carga diaria promedio de pacientes,
x_2 = exposiciones de rayos X mensuales,
x_3 = días-cama ocupados mensuales,
x_4 = población elegible en el área/1000,
x_5 = duración promedio de la permanencia de un paciente, en días.

Sitio	x_1	x_2	x_3	x_4	x_5	y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1003.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20106	3655.08	180.5	6.15	3503.93
11	96.00	13313	2912.00	60.9	5.88	3571.59
12	131.42	10771	3921.00	103.7	4.88	3741.40
13	127.21	15543	3865.67	126.8	5.50	4026.52
14	252.90	36194	7684.10	157.7	7.00	10343.81
15	409.20	34703	12446.33	169.4	10.75	11732.17
16	463.70	39204	14098.40	331.4	7.05	15414.94
17	510.22	86533	15524.00	371.6	6.35	18854.45

El objetivo es generar una ecuación empírica para estimar (o predecir) las necesidades de personal en los hospitales de la marina. Estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,x_3,x_4,x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 + \beta_5 x_5.$$

12.13 Se realizó un experimento para estudiar el tamaño de los calamares consumidos por tiburones y atunes. Las variables regresoras son características del pico o la boca del calamar. Las variables regresoras y la respuesta considerada para el estudio son las siguientes:

x_1 = longitud del morro, en pulgadas,
x_2 = longitud de aleta, en pulgadas,
x_3 = longitud del morro a la cola, en pulgadas,
x_4 = longitud de la cola a la aleta, en pulgadas,
x_5 = ancho, en pulgadas,
y = peso, en libras.

x_1	x_2	x_3	x_4	x_5	y
1.31	1.07	0.44	0.75	0.35	1.95
1.55	1.49	0.53	0.90	0.47	2.90
0.99	0.84	0.34	0.57	0.32	0.72
0.99	0.83	0.34	0.54	0.27	0.81
1.01	0.90	0.36	0.64	0.30	1.09
1.09	0.93	0.42	0.61	0.31	1.22
1.08	0.90	0.40	0.51	0.31	1.02
1.27	1.08	0.44	0.77	0.34	1.93
0.99	0.85	0.36	0.56	0.29	0.64
1.34	1.13	0.45	0.77	0.37	2.08
1.30	1.10	0.45	0.76	0.38	1.98
1.33	1.10	0.48	0.77	0.38	1.90
1.86	1.47	0.60	1.01	0.65	8.56
1.58	1.34	0.52	0.95	0.50	4.49
1.97	1.59	0.67	1.20	0.59	8.49
1.80	1.56	0.66	1.02	0.59	6.17
1.75	1.58	0.63	1.09	0.59	7.54
1.72	1.43	0.64	1.02	0.63	6.36
1.68	1.57	0.72	0.96	0.68	7.63
1.75	1.59	0.68	1.08	0.62	7.78
2.19	1.86	0.75	1.24	0.72	10.15
1.73	1.67	0.64	1.14	0.55	6.88

Estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,x_3,x_4,x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 + \beta_5 x_5.$$

12.14 Veintitrés estudiantes de pedagogía tomaron parte en un programa de evaluación diseñado para medir la eficacia de los profesores y determinar qué factores son importantes. Participaron 11 instructoras. La medición de la respuesta fue una evaluación cuantitativa del maestro colaborador. Las variables regresoras fueron las calificaciones de cuatro pruebas estandarizadas entregadas a cada instructor. Los datos son los siguientes:

y	x_1	x_2	x_3	x_4
410	69	125	59.00	55.66
569	57	131	31.75	63.97
425	77	141	80.50	45.32
344	81	122	75.00	46.67
324	0	141	49.00	41.21
505	53	152	49.35	43.83
235	77	141	60.75	41.61
501	76	132	41.25	64.57
400	65	157	50.75	42.41
584	97	166	32.25	57.95
434	76	141	54.50	57.90

Estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,x_3,x_4} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

12.15 Se llevó a cabo un estudio sobre el uso de cierto rodamiento y y su relación con x_1 = viscosidad del aceite y x_2 = carga. Se obtuvieron los datos siguientes. [De *Response Surface Methodology*, Myers y Montgomery (2002).]

y	x_1	x_2	y	x_1	x_2
193	1.6	851	230	15.5	816
172	22.0	1058	91	43.0	1201
113	33.0	1357	125	40.0	1115

a) Estime los parámetros desconocidos de la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

b) Prediga el uso para una viscosidad del aceite de 20 y una carga de 1200.

12.16 Un ingeniero de una compañía de semiconductores desea modelar la relación entre la ganancia del dispositivo o hFE(y) y tres parámetros: RS del emisor (x_1), RS de la base (x_2) y RS del emisor a la base (x_3). A continuación se muestran los datos:

x_1, RS del emisor	x_2, RS de la cBase	x_3, E-B-RS	y, hFE-1M-5V
14.62	226.0	7.000	128.40
15.63	220.0	3.375	52.62
14.62	217.4	6.375	113.90
15.00	220.0	6.000	98.01
14.50	226.5	7.625	139.90
15.25	224.1	6.000	102.60
16.12	220.5	3.375	48.14
15.13	223.5	6.125	109.60
15.50	217.6	5.000	82.68
15.13	228.5	6.625	112.60
15.50	230.2	5.750	97.52
16.12	226.5	3.750	59.06
15.13	226.6	6.125	111.80
15.63	225.6	5.375	89.09
15.38	234.0	8.875	171.90
15.50	230.0	4.000	66.80

x_1, RS del emisor	x_2, RS de la cBase	x_3, E-B-RS	y, hFE-1M-5V
14.25	224.3	8.000	157.10
14.50	240.5	10.870	208.40
14.62	223.7	7.375	133.40

a) Haga un ajuste de regresión lineal múltiple para los datos.

b) Prediga hFE cuando $x_1 = 14$, $x_2 = 220$ y $x_3 = 5$. [Datos tomados de Myers y Montgomery (2002)].

12.4 Propiedades de los estimadores de mínimos cuadrados

Las medias y varianzas de los estimadores $b_0, b_1, \ldots, b_k$ se obtienen con facilidad si se hacen ciertas suposiciones sobre los errores aleatorios $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$ que son idénticas a las que se hicieron en el caso de la regresión lineal simple. Si se supone que dichos errores son independientes, con media igual a cero y varianza σ^2, entonces puede demostrarse que $b_0, b_1, \ldots, b_k$ son, respectivamente, estimadores insesgados de los coeficientes de regresión $\beta_0, \beta_1, \ldots, \beta_k$. Además, las varianzas de las b se obtienen por medio de los elementos de la inversa de la matriz $\mathbf{A}$. Observe que los elementos fuera de la diagonal principal de $\mathbf{A} = \mathbf{X}'\mathbf{X}$ representan sumas de productos de los elementos en las columnas de $\mathbf{X}$; mientras que los elementos en dicha diagonal de $\mathbf{A}$ son las sumas de los cuadrados de los elementos en las columnas de $\mathbf{X}$. La matriz inversa, $\mathbf{A}^{-1}$, aparte del multiplicador σ^2, representa la **matriz de varianza-covarianza** de los coeficientes de regresión estimados. Es decir, los elementos de la matriz $\mathbf{A}^{-1}\sigma^2$ muestran, en la diagonal principal, las varianzas de $b_0, b_1, \ldots, b_k$; y fuera de la diagonal principal están las covarianzas. Por ejemplo, en un problema de regresión lineal múltiple con $k = 2$, se escribiría

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{bmatrix}$$

con los elementos debajo de la diagonal principal determinados por la simetría de la matriz. Entonces, se escribe

$$\sigma_{b_i}^2 = c_{ii}\sigma^2, \qquad i = 0, 1, 2,$$
$$\sigma_{b_i b_j} = Cov(b_i, b_j) = c_{ij}\sigma^2, \ i \neq j.$$

Desde luego, los estimadores de las varianzas y también los errores estándar de ellos se obtienen con el reemplazo de σ^2 con el estimador apropiado obtenido de los datos experimentales. Un estimador no sesgado de σ^2 de nuevo está definido por la suma de los errores al cuadrado, que se calcula con la fórmula establecida en el teorema 12.1. En el teorema se hacían las suposiciones sobre los ϵ_i descritas con anterioridad.

Teorema 12.1: Para la ecuación de regresión lineal

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

un estimador insesgado de σ^2 está dado por el error o media cuadrática residual

$$s^2 = \frac{SSE}{n - k - 1}, \quad \text{donde} \quad SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Puede verse que para el caso de la regresión lineal simple, el teorema 12.1 representa una generalización del teorema 11.1. La prueba se deja como ejercicio para

el lector. Al igual que en el caso de la regresión lineal más simple, la estimación de s^2 es una medida de la variación de los errores de la predicción, o residuos. En las secciones 12.10 y 12.11 se presentan otras inferencias importantes que tienen que ver con la ecuación ajustada de regresión, con base en los valores de los residuos individuales $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$.

El error y la suma de los cuadrados de la regresión adoptan la misma forma y juegan el mismo papel que para el caso de la regresión lineal simple. De hecho, la identidad de la suma de cuadrados

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

se sigue cumpliendo, y se conserva la notación anterior, que es,

$$SST = SSR + SSE$$

con

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \text{suma total de cuadrados,}$$

y

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \text{suma de cuadrados de regresión.}$$

Hay k grados de libertad asociados con SSR, y, como siempre, SST tiene $n - 1$ grados de libertad. Por lo tanto, después de restar, SSE tiene $n - k - 1$ grados de libertad. Así, nuestro estimador de σ^2 de nuevo está dado por la suma de errores al cuadrado dividida entre sus grados de libertad. En la salida de la mayoría de paquetes de cómputo de regresión múltiple aparecerán estas tres sumas de cuadrados.

Análisis de varianza en la regresión múltiple

La partición de la suma total de cuadrados en sus componentes, la suma de la regresión y de los cuadrados de los errores juegan un papel importante. Puede efectuarse un **análisis de varianza** que arroje luz sobre la calidad de la ecuación de regresión. Una hipótesis útil, que determina si el modelo explica una cantidad significativa de variación, es la siguiente:

$$H_0 : \ \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0.$$

El análisis de varianza implica una prueba F, mediante una tabla como la siguiente:

Fuente	**Suma de los cuadrados**	**Grados de libertad**	**Media cuadrática**	***F***
Regresión	SSR	k	$MSR = \frac{SSR}{k}$	$f = \frac{MSR}{MSE}$
Error	SSE	$n - (k+1)$	$MSE = \frac{SSE}{n-(k+1)}$	
Total	SST	$n - 1$		

La prueba que se relaciona es la **prueba de cola superior**. El rechazo de H_0 significa que la **ecuación de regresión difiere en una constante**. Es decir, al menos una variable regresora es importante. En las secciones que siguen se estudia más el uso del análisis de varianza.

Una utilidad adicional del error cuadrático medio (o media cuadrática residual) estriba en su uso en la prueba de hipótesis y en la estimación de intervalos de confianza, que se estudian en la sección 12.5. Además, el error cuadrático medio juega un papel importante en las situaciones en que el científico busca el mejor modelo entre un conjunto de ellos que están en competencia. Muchos criterios de construcción de modelos incluyen el estadístico s^2. En la sección 12.11 se presentan criterios para comparar modelos en competencia.

12.5 Inferencias en la regresión lineal múltiple

Una de las inferencias más útiles que se pueden hacer respecto de la calidad de la respuesta predicha y_0 correspondiente a los valores x_{10}, $x_{20}, \ldots,$ x_{k0} es el intervalo de confianza sobre la respuesta media $\mu_{Y|x_{10},x_{20},\ldots,x_{k0}}$. Estamos interesados en construir un intervalo de confianza sobre la respuesta media para el conjunto de condiciones dadas por

$$\mathbf{x}_0' = [1, x_{10}, x_{20}, \ldots, x_{k0}].$$

Se aumentan en 1 las condiciones sobre las x para facilitar la notación matricial. Normalmente, los ϵ_i producen normalidad en las b_j, y la media, varianzas y covarianzas aún son las mismas, como se indica en la sección 12.4. Por lo tanto,

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_{j0}$$

que, igualmente, está distribuida en forma normal y es, de hecho, un estimador insesgado para la **respuesta media** sobre la que se intenta trazar intervalos de confianza. La varianza de $\hat{y}_0$, escrita con notación matricial simplemente como función de σ^2, $(\mathbf{X}'\mathbf{X})^{-1}$, y el vector de condiciones, $\mathbf{x}_0'$, es

$$\sigma^2_{\hat{y}_0} = \sigma^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0.$$

Si se expandiera esta expresión para un caso dado, por ejemplo, para $k = 2$, es fácil observar que en forma apropiada es responsable de las varianzas y covarianzas de las b_j. Después de sustituir σ^2 con s^2, según se da en el teorema 12.1, puede construirse el intervalo de confianza de $(1 - \alpha)100\%$, sobre $\mu_{Y|x_{10},x_{20},\ldots,x_{k0}}$ a partir del estadístico

$$T = \frac{\hat{y}_0 - \mu_{Y|x_{10},x_{20},\ldots,x_{k0}}}{s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}},$$

que tiene distribución t con $n - k - 1$ grados de libertad.

Intervalo de confianza para $\mu_{Y|x_{10},x_{20},\ldots,x_{k0}}$

Un intervalo de confianza de $(1-\alpha)100\%$ para la **respuesta media** $\mu_{Y|x_{10},x_{20},\ldots,x_{k0}}$ es

$$\hat{y}_0 - t_{\alpha/2} s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < \mu_{Y|x_{10},x_{20},\ldots,x_{k0}} < \hat{y}_0 + t_{\alpha/2} s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - k - 1$ grados de libertad.

Es frecuente que la cantidad $s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$ se denomine **error estándar de la predicción** y, por lo general, aparece en la salida de muchos paquetes de cómputo para regresión.

Ejemplo 12.5: Con los datos del ejemplo 12.3, construya un intervalo de confianza de 95% para la respuesta media, cuando $x_1 = 3\%$, $x_2 = 8\%$ y $x_3 = 9\%$.

Solución: De la ecuación de regresión del ejemplo 12.3, el porcentaje estimado de supervivencia cuando $x_1 = 3\%$, $x_2 = 8\%$ y $x_3 = 9\%$, es:

$$\hat{y} = 39.1574 + (1.0161)(3) - (1.8616)(8) - (0.3433)(9) = 24.2232.$$

Y luego se determina que

$$\begin{aligned}\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 &= [1,3,8,9]\begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0037 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0037 & -0.0021 & 0.0886 \end{bmatrix}\begin{bmatrix}1\\3\\8\\9\end{bmatrix} \\ &= 0.1267.\end{aligned}$$

Usando el error cuadrático medio, $s^2 = 4.298$ o $s = 2.073$, y la tabla A.4, se observa que $t_{0.025} = 2.262$ para 9 grados de libertad. Por lo tanto, un intervalo de confianza de 95% para el porcentaje medio de supervivencia para $x_1 = 3\%$, $x_2 = 8\%$ y $x_3 = 9\%$, está dado por

$$\begin{aligned}24.2232 - (2.262)(2.073)\sqrt{0.1267} &< \mu_{Y|3,8,9} \\ &< 24.2232 + (2.262)(2.073)\sqrt{0.1267},\end{aligned}$$

o simplemente $22{,}5541 < \mu_{Y|3,8,9} < 25.8923$.

Igual que en el caso de la regresión lineal simple, se necesita distinguir con claridad entre el intervalo de confianza sobre la respuesta media y el intervalo de predicción sobre una *respuesta observada*. Esta última proporciona una frontera dentro de la cual puede decirse que caerá una respuesta nueva observada, con el grado preseleccionado de certidumbre.

Un intervalo para una sola respuesta predicha y_0 de nuevo se establece al considerar la diferencia $\hat{y}_0 - y_0$. Puede demostrarse que la distribución del muestreo es normal con media

$$\mu_{\hat{y}_0 - y_0} = 0,$$

y varianza

$$\sigma^2_{\hat{y}_0 - y_0} = \sigma^2[1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0].$$

Así, puede construirse un intervalo de predicción de $(1-\alpha)100\%$ para un solo valor de predicción y_0 a partir del estadístico

$$T = \frac{\hat{y}_0 - y_0}{s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}},$$

el cual tiene una distribución t con $n - k - 1$ grados de libertad.

Intervalo de predicción para y_0

Un intervalo de $(1 - \alpha)100\%$ de predicción para una **sola respuesta y_0** está dado por

$$\hat{y}_0 - t_{\alpha/2} s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < y_0 < \hat{y}_0 + t_{\alpha/2} s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - k - 1$ grados de libertad.

Ejemplo 12.6: Con los datos del ejemplo 12.3, construya un intervalo de predicción de 95% para una respuesta individual de porcentaje de supervivencia, cuando $x_1 = 3\%$, $x_2 = 8\%$ y $x_3 = 9\%$.

Solución: En relación con los resultados del ejemplo 12.5, encontramos que el intervalo de predicción de 95% para la respuesta y_0, cuando $x_1 = 3$, $x_2 = 8\%$ y $x_3 = 9\%$, es

$$24.2232 - (2.262)(2.073)\sqrt{1.1267} < y_0 < 24.2232 + (2.262)(2.073)\sqrt{1.1267},$$

que se reduce a $19.2459 < y_0 < 29.2005$. Observe que, como se esperaba, el intervalo de predicción es considerablemente más ancho que el intervalo de confianza para el porcentaje medio de supervivencia del ejemplo 12.5.

El conocimiento de las distribuciones de los estimadores de los coeficientes individuales permite al experimentador construir intervalos de confianza para los coeficientes y, de ese modo, hacer pruebas de hipótesis sobre ellos. Recuerde que en la sección 12.4 se vio que las b_j $(j = 0, 1, 2, \ldots, k)$ están distribuidas en forma normal con media β_j y varianza $c_{jj}\sigma^2$. Por lo que se puede usar el estadístico

$$t = \frac{b_j - \beta_{j0}}{s\sqrt{c_{jj}}}$$

con $n - k - 1$ grados de libertad para probar la hipótesis y construir intervalos de confianza sobre β_j. Por ejemplo, si se desea probar:

$$H_0\text{: } \beta_j = \beta_{j0},$$
$$H_1\text{: } \beta_j \neq \beta_{j0},$$

se calcula el estadístico t anterior y no se rechaza H_0 si $-t_{\alpha/2} < t < t_{\alpha/2}$, donde $t_{\alpha/2}$ tiene $n - k - 1$ grados de libertad.

Ejemplo 12.7: Para el modelo del ejemplo 12.3, pruebe la hipótesis de que $\beta_2 = -2.5$ contra la alternativa de que $\beta_2 > -2.5$, con un nivel de significancia de 0.05.

Solución:

$$H_0\text{: } \beta_2 = -2.5,$$
$$H_1\text{: } \beta_2 > -2.5.$$

Cálculos:

$$t = \frac{b_2 - \beta_{20}}{s\sqrt{c_{22}}} = \frac{-1.8616 + 2.5}{2.073\sqrt{0.0166}} = 2.390,$$
$$P = P(T > 2.390) = 0.04.$$

Decisión: Rechace H_0 y concluya que $\beta_2 > -2.5$.

Pruebas T individuales para comparar variables

La prueba T que se utiliza con más frecuencia en la regresión múltiple es aquella que prueba la importancia de los coeficientes individuales (es decir, $H_0 : \beta_j = 0$ contra la alternativa $H_0 : \beta_j \neq 0$). Es frecuente que estas pruebas contribuyan a lo que se denomina **comparación de variables**, con lo cual el analista intenta llegar al modelo más útil (la selección de qué regresor utilizar). Aquí debe destacar en que si se encuentra que un coeficiente es insignificante (es decir, **no se rechaza** la hipótesis $H_0 : \beta_j = 0$), la conclusión que se obtiene es que la **variable** es insignificante (explica una cantidad insignificante de la variación de y), **en la presencia de los demás regresores del modelo**. Se profundizará en este punto más adelante.

Salida anotada para los datos del ejemplo 12.3

La figura 12.1 muestra una salida anotada por computadora para el ajuste de regresión lineal múltiple de los datos del ejemplo 12.3. Se empleó el paquete *SAS*.

Observe los estimadores de los parámetros del modelo, los errores estándar y los estadísticos t que aparecen en la salida. Los errores estándar se calculan a partir de las raíces cuadradas de los elementos de la diagonal $(\mathbf{X}'\mathbf{X})^{-1}s^2$. En dicha ilustración, la variable x_3 es insignificante en presencia de x_1 y x_2 con base en la prueba t y el valor P correspondiente $= 0.5916$. Los términos CLM y CLI son intervalos de confianza sobre la respuesta media y los límites de la predicción sobre una observación individual, respectivamente. En el análisis de varianza la prueba f indica que queda explicada una cantidad significativa de variabilidad. Como ejemplo de las interpretaciones de CLM y CLI , considere la observación 10. Con una observación de 25.2 y un valor predicho de 26.068 se tiene una confianza de 95% de que la respuesta media estará entre 24.502 y 27.633, y de que una observación nueva caerá entre 21.124 y 31.011 con una probabilidad de 0.95. El valor R^2 de 0.9117 implica que el modelo explica el 91.17% de la variabilidad de la respuesta. En la sección 12.6 se analiza más la R^2.

Más sobre el análisis de varianza en la regresión múltiple (opcional)

En la sección 12.4 se estudió brevemente la partición de la suma total de los cuadrados $\sum_{i=1}^{n}(y_i - \bar{y})^2$ en sus dos componentes, el modelo de regresión y la suma de errores al cuadrado (que se ilustran en la figura 12.1). El análisis de varianza lleva a la prueba de

$$H_0 : \ \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0.$$

El rechazo de la hipótesis nula tiene una interpretación importante para el científico o el ingeniero. (Para aquellos que estén interesados en un tratamiento más profundo de este tema por medio de matrices, es útil estudiar el desarrollo de las sumas de los cuadrados que se usan en el ANOVA.)

En primer lugar, hay que recordar la definición de $\mathbf{y}$, $\mathbf{X}$, y β que se dio en la sección 12.3, así como la de $\mathbf{b}$, el vector de los estimadores de mínimos cuadrados dados por

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

```
                          Sum of       Mean
Source           DF      Squares     Square  F Value  Pr > F
Model             3    399.45437  133.15146    30.98  <.0001
Error             9     38.67640    4.29738
Corrected Total  12    438.13077

Root MSE         2.07301    R-Square     0.9117
Dependent Mean  29.03846    Adj R-Sq     0.8823
Coeff Var        7.13885

                      Parameter       Standard
Variable     DF        Estimate          Error    t Value    Pr > |t|
Intercept     1        39.15735        5.88706       6.65      <.0001
x1            1         1.01610        0.19090       5.32      0.0005
x2            1        -1.86165        0.26733      -6.96      <.0001
x3            1        -0.34326        0.61705      -0.56      0.5916

    Dependent Predicted    Std Error
Obs  Variable     Value Mean Predict    95% CL Mean       95% CL Predict   Residual
  1   25.5000   27.3514       1.4152 24.1500   30.5528   21.6734   33.0294  -1.8514
  2   31.2000   32.2623       0.7846 30.4875   34.0371   27.2482   37.2764  -1.0623
  3   25.9000   27.3495       1.3588 24.2757   30.4234   21.7425   32.9566  -1.4495
  4   38.4000   38.3096       1.2818 35.4099   41.2093   32.7960   43.8232   0.0904
  5   18.4000   15.5447       1.5789 11.9730   19.1165    9.6499   21.4395   2.8553
  6   26.7000   26.1081       1.0358 23.7649   28.4512   20.8658   31.3503   0.5919
  7   26.4000   28.2532       0.8094 26.4222   30.0841   23.2189   33.2874  -1.8532
  8   25.9000   26.2219       0.9732 24.0204   28.4233   21.0414   31.4023  -0.3219
  9   32.0000   32.0882       0.7828 30.3175   33.8589   27.0755   37.1008  -0.0882
 10   25.2000   26.0676       0.6919 24.5024   27.6329   21.1238   31.0114  -0.8676
 11   39.7000   37.2524       1.3070 34.2957   40.2090   31.7086   42.7961   2.4476
 12   35.7000   32.4879       1.4648 29.1743   35.8015   26.7459   38.2300   3.2121
 13   26.5000   28.2032       0.9841 25.9771   30.4294   23.0122   33.3943  -1.7032
```

Figura 12.1: Salida de *SAS* para los datos del ejemplo 12.3.

Una partición de la **suma de cuadrados no corregida**,

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^{n} \mathbf{y}_i^2$$

en dos componentes está dada por

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= \mathbf{b}'\mathbf{X}'\mathbf{y} + (\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}) \\ &= \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + [\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}].\end{aligned}$$

El segundo término (entre corchetes) en el lado derecho es tan sólo la suma de errores al cuadrado $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. El lector debería observar que una expresión alternativa para la suma de errores al cuadrado es

$$SSE = \mathbf{y}'[\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}.$$

El término $\mathbf{y'X(X'X)^{-1}X'y}$ se denomina la **suma de cuadrados de la regresión**. Sin embargo, no se trata de la expresión $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ que se usó para probar la "importancia" de los términos b_1, $b_2, \ldots,$ b_k, sino de

$$\mathbf{y'X(X'X)^{-1}X'y} = \sum_{i=1}^{n} \hat{y}_i^2,$$

que es la suma de cuadrados de la regresión no corregida para la media. Como tal, sólo puede usarse para probar si la "ecuación de regresión difiere significativamente de cero". Es decir,

$$H_0 : \ \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

En general, esto no es tan importante como probar

$$H_0 : \ \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

dado que esto plantea que la respuesta media es una constante, no necesariamente vale cero.

Grados de libertad

Así, la partición de las sumas de los cuadrados y los grados de libertad se reduce a

Fuente	Suma de cuadrados	g.l.
Regresión	$\sum_{i=1}^{n} \hat{y}_i^2 = \mathbf{y'X(X'X)^{-1}X'y}$	$k+1$
Error	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{y'[I_n - X(X'X)^{-1}X']y}$	$n-(k+1)$
Total	$\sum_{i=1}^{n} y_i^2 = \mathbf{y'y}$	n

Hipótesis de interés

Ahora, por supuesto, la hipótesis de interés para un ANOVA debe eliminar el papel de la intersección según se describió en forma previa. Si se habla estrictamente, si $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, entonces la recta de regresión estimada es tan sólo $\hat{y}_i = \bar{y}$. Como resultado, en realidad se busca evidencia de que la ecuación de regresión "no es una constante". Entonces, las sumas total y de los cuadrados de la regresión deben "corregirse para la media".
Como resultado, tenemos

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

En notación matricial es simplemente

$$\begin{aligned}\mathbf{y'[I_n - 1(1'1)^{-1}1']y} =& \mathbf{y'[X(X'X)^{-1}X' - 1(1'1)^{-1}1']y} \\ &+ \mathbf{y'[I_n - X(X'X)^{-1}X']y}.\end{aligned}$$

En esta expresión, el **1** sólo es un vector de n unos. Como resultado, simplemente restamos

$$\mathbf{y'1(1'1)^{-1}1'y} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

de $\mathbf{y'y}$ y de $\mathbf{y'X(X'X)^{-1}X'y}$ (es decir, corrigiendo las sumas total y de los cuadrados de la regresión para la media).

Por último, la partición apropiada de las sumas de los cuadrados con grados de libertad es como sigue:

Fuente	**Suma de cuadrados**	**g.l.**
Regresión	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \mathbf{y'[X(X'X)^{-1}X' - 1(1'1)^{-1}1]y}$	k
Error	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y'[I_n - X(X'X)^{-1}X']y}$	$n-(k+1)$
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y'[I_n - 1(1'1)^{-1}1']y}$	$n-1$

Ésta es la tabla ANOVA que aparece en la salida por computadora de la figura 12.1. Es frecuente denominar a la expresión $\mathbf{y'[1(1'1)^{-1}1']y}$ como la **suma de cuadrados de la regresión asociada con la media**, y a ella se asigna 1 grado de libertad.

Ejercicios

12.17 Para los datos del ejercicio 12.2 de la página 452, estime σ^2.

12.18 Para los datos del ejercicio 12.3 de la página 453, estime σ^2.

12.19 Para los datos del ejercicio 12.9 de la página 454, estime σ^2.

12.20 Obtenga estimadores de las varianzas y la covarianza de los estimadores b_1 y b_2 del ejercicio 12.2 de la página 454.

12.21 En relación con el ejercicio 12.9 de la página 454, encuentre el estimador de

a) $\sigma^2_{b_2}$,

b) $Cov(b_1, b_4)$.

12.22 Utilizando los datos del ejercicio 12.2 de la página 452, y el estimador σ^2 del ejercicio 12.17, calcule intervalos de confianza de 95% para la respuesta predicha y la respuesta media cuando $x_1 = 900$ y $x_2 = 1.00$.

12.23 Para el ejercicio 12.8 de la página 454, construya un intervalo de confianza de 90% para la resistencia media a la compresión cuando la concentración es $x = 19.5$ y se utiliza un modelo cuadrático.

12.24 Con los datos del ejercicio 12.9 de la página 454 y el estimador de σ^2 del ejercicio 12.19, calcule intervalos de confianza de 95% para la respuesta predicha y la respuesta media cuando $x_1 = 75$, $x_2 = 24$, $x_3 = 90$ y $x_4 = 98$.

12.25 Para el modelo del ejercicio 12.7 de la página 453, pruebe la hipótesis de que $\beta_2 = 0$, con un nivel de significancia de 0.05, contra la alternativa de que $\beta_2 \neq 0$.

12.26 Para el modelo del ejercicio 12.2 de la página 452, pruebe la hipótesis de que $\beta_1 = 0$, con un nivel de significancia de 0.05, contra la alternativa de que $\beta_1 \neq 0$.

12.27 Para el modelo del ejercicio 12.3 de la página 453, pruebe la hipótesis de que $\beta_1 = 2$ contra la alternativa de que $\beta_1 \neq 2$. En su conclusión, use un valor *P*.

12.28 Considere los datos siguientes, que se listan en el ejercicio 12.15 de la página 455.

y (uso)	x_1 (viscosidad del aceite)	x_2 (carga)
193	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

a) Estime σ^2 usando regresión múltiple de y sobre x_1 y x_2.

b) Calcule valores pronosticados, un intervalo de confianza de 95% para la media del uso, y un intervalo de predicción de 95% para el uso observado, si $x_1 = 20$ y $x_2 = 1000$.

12.29 Con los datos del ejercicio 12.28, y con un nivel de 0.05, pruebe:

a) H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$;

b) H_0: $\beta_2 = 0$ contra H_1: $\beta_2 \neq 0$.

c) ¿Tiene usted alguna razón para creer que deba cambiarse el modelo del ejercicio 12.28? ¿Por qué?

12.30 Con los datos del ejercicio 12.16 de la página 455,

a) Estime σ^2 usando regresión múltiple de y sobre x_1, x_2 y x_3;

b) Calcule un intervalo de predicción de 95% para la ganancia del dispositivo con tres regresores en $x_1 = 15.0$, $x_2 = 220.0$ y $x_3 = 6.0$.

12.6 Selección de un modelo ajustado mediante la prueba de hipótesis

En muchas situaciones de regresión, los coeficientes individuales revisten importancia para el experimentador. Por ejemplo, en una aplicación de economía, β_1, $\beta_2, \ldots$ podrían tener algún significado específico, por lo que los intervalos de confianza y las pruebas de hipótesis sobre dichos parámetros tendrían interés para el economista. Sin embargo, considere una situación de química industrial en la que el modelo propuesto supone que la reacción que ocurre es linealmente dependiente de la temperatura y concentración de la reacción de cierto catalizador. Es probable que se sepa que éste no es el verdadero modelo, sino una aproximación adecuada; de manera que el interés no estribaría en los parámetros individuales, sino en la capacidad de la función en su conjunto para predecir la respuesta verdadera en el rango de las variables consideradas. Por lo tanto, en esta situación, se pondría más énfasis en los intervalos de confianza $\sigma^2_{\hat{Y}}$ sobre la respuesta media, y otros parecidos, y se disminuiría el interés en las inferencias sobre los parámetros individuales.

El experimentador que utiliza análisis de regresión también está interesado en la eliminación de variables cuando la situación impone que, además de llegar a una ecuación de pronóstico funcional, debe encontrar la "mejor regresión" que implique sólo a variables que son predictores útiles. Se dispone de cierto número de programas de cómputo que llegan en secuencia a la denominada mejor ecuación de regresión, según ciertos criterios. En la sección 12.9 estudiaremos esto con mayor profundidad.

Un criterio que es común utilizar para ilustrar lo adecuado de un modelo ajustado de regresión es el **coeficiente de determinación múltiple**:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}.$$

Note que ésta se parece a la descripción de R^2 que se hizo en el capítulo 11. En este punto, la explicación podría ser más clara toda vez que ahora nos centramos en *SSR* como la **variabilidad explicada**. La cantidad R^2 tan sólo indica qué proporción de la variación total de la respuesta Y es explicada por el modelo ajustado. Es frecuente que un experimentador informe $R^2 \times 100\%$ e interprete el resultado como el porcentaje de variación explicado con el modelo propuesto. La raíz cuadrada de R^2 se

denomina **coeficiente de correlación múltiple** entre Y y el conjunto $x_1, x_2, \ldots, x_k$. En el ejemplo 12.3, el valor de R^2 que indica la proporción de variación explicada por las tres variables independientes x_1, x_2 y x_3 se encuentra mediante

$$R^2 = \frac{SSR}{SST} = \frac{399.45}{438.13} = 0.9117.$$

lo cual significa que 91.17% de la variación en porcentaje de supervivencia queda explicada por el modelo de regresión lineal.

La suma de cuadrados de la regresión puede emplearse para obtener algún indicio acerca de si el modelo es o no una explicación adecuada de la situación verdadera. Podemos probar la hipótesis H_0 de que la **regresión no es significativa** con únicamente plantear la razón

$$f = \frac{SSR/k}{SSE/(n-k-1)} = \frac{SSR/k}{s^2}$$

y rechazar H_0 con el nivel de significancia de α cuando $f > f_\alpha(k, n-k-1)$. Para los datos del ejemplo 12.3, se obtiene

$$f = \frac{399.45/3}{4.298} = 30.98.$$

De la salida que aparece en la figura 12.1, el valor P es menor que 0.0001. Esto no debe malinterpretarse. Aunque indica que la regresión explicada por el modelo es significativa, no descarta la posibilidad de que

1. El modelo de regresión lineal en este conjunto de x no sea el único que puede usarse para explicar los datos; en efecto, quizás haya otros modelos con transformaciones sobre las x que arrojen un valor mayor del estadístico F.

2. El modelo hubiera podido ser más eficaz con la inclusión de otras variables, además de x_1, x_2 y x_3, o quizá con la eliminación de una o más de las variables del modelo, por ejemplo x_3, que muestre un valor de $P = 0.5916$.

El lector debería recordar el análisis que se hizo en la sección 11.5 sobre las desventajas de utilizar R^2 como criterio para comparar modelos en competencia. Es claro que dichas desventajas son relevantes en la regresión lineal múltiple. En realidad, los riesgos de su empleo en la regresión múltiple son aún mayores debido a que es muy grande la tentación de sobreajustar. Siempre debe recordarse que el hecho de que un valor de $R^2 \approx 1.0$ puede obtenerse a expensas de los grados de libertad del error cuando se emplea un exceso de términos en el modelo. Sin embargo, un valor de $R^2 = 1$, que describa un modelo con ajuste casi perfecto, no siempre genera un modelo que haga buenas predicciones.

El coeficiente de determinación ajustado (R^2 ajustado)

En el capítulo 11 se presentan varias figuras que muestran salidas por computadora, tanto de *SAS* como de *MINITAB*, donde aparece un estadístico llamado R^2 *ajustado*, o coeficiente de determinación ajustado. R^2 ajustado es una variación de R^2 que proporciona un **ajuste para los grados de libertad**. El coeficiente de determinación, según se definió en la página 407, no puede disminuir conforme se agregan términos al modelo. En otras palabras, R^2 no disminuye conforme se reducen los grados de

libertad del error $n - k - 1$, y el último resultado se produce por un incremento de k, el número de términos en el modelo. R^2 ajustado se calcula con la división de *SSE* y *SST* entre sus valores respectivos de grados de libertad (es decir, R^2 ajustado es como sigue).

R^2 ajustado

$$R^2_{\text{aju}} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}.$$

Para ilustrar el uso de R^2_{aju} se revisará el ejemplo 12.3.

¿Cómo la remoción de x_3 afecta R^2 y R^2_{aju}?

La prueba t (o la F correspondiente) para x_3, el porcentaje ponderado del ingrediente 3, sugiere con claridad que un modelo más sencillo que sólo implique x_1 y x_2 bien podría ser una mejoría. En otras palabras, el modelo completo con todos los regresores podría estar sobreajustado. Por supuesto que es de interés investigar R^2 y R^2_{aju} tanto para el modelo completo (x_1, x_2 y x_3) como para el restringido (x_1, x_2). Por la figura 12.1, ya sabemos que $R^2_{\text{completo}} = 0.9117$. La *SSE* para el modelo reducido es 40.01, por lo que $R^2_{\text{restringido}} = 1 - \frac{40.01}{438.13} = \mathbf{0.9087}$. Así, con x_3 dentro del modelo se explica más variabilidad. No obstante, como ya se dijo, esto ocurriría aun si el modelo estuviera sobreajustado. Ahora, por supuesto, R^2_{aju} está diseñado para proporcionar un estadístico que castigue un modelo sobreajustado, de manera que podríamos esperar que se favorezca al modelo restringido. Entonces, para el modelo completo

$$R^2_{\text{aju}} = 1 - \frac{38.6764/9}{438.1308/12} = 1 - \frac{4.2974}{36.5109} = 0.8823,$$

mientras que para el modelo reducido (eliminación de x_3)

$$R^2_{\text{adj}} = 1 - \frac{40.01/10}{438.1308/12} = 1 - \frac{4.001}{36.5109} = 0.8904.$$

Así, R^2_{aju} en verdad favorece el modelo restringido y, por ello, confirma la evidencia producida por las pruebas t y F que sugieren que el modelo reducido es preferible sobre el que contiene los tres regresores. El lector quizás espere que otros estadísticos sugieran el rechazo del modelo sobreajustado. Véase el ejercicio 12.40 de la página 474.

Pruebas sobre subconjuntos y coeficientes individuales

Agregar cualquier variable única a un sistema de regresión *incrementará la suma de cuadrados de la regresión*, y con ello *se reducirá la suma de errores al cuadrado.* En consecuencia, se debe decidir si el incremento en la regresión es suficiente para garantizar su uso en el modelo. Como es de esperarse, el empleo de variables sin importancia reduciría la eficacia de la ecuación de predicción por el incremento de la variable de la respuesta estimada. Profundizaremos más en este punto al considerar la importancia de x_3 en el ejemplo 12.3. Inicialmente, se prueba

$$H_0\text{: } \beta_3 = 0,$$
$$H_1\text{: } \beta_3 \neq 0$$

usando la distribución t con 9 grados de libertad. Se tiene

$$t = \frac{b_3 - 0}{s\sqrt{c_{33}}} = \frac{-0.3433}{2.073\sqrt{0.0886}} = -0.556,$$

que indica que β_3 no difiere en forma significativa de cero y, por ello, bien se podría tener la justificación para eliminar x_3 del modelo. Suponga que se considera la regresión de Y sobre el conjunto (x_1, x_2), las ecuaciones normales de mínimos cuadrados ahora se reducen a

$$\begin{bmatrix} 13 & 59.43 & 81.82 \\ 59.43 & 394.7255 & 360.6621 \\ 81.82 & 360.6621 & 576.7264 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 377.50 \\ 1877.5670 \\ 2246.6610 \end{bmatrix}.$$

Los coeficientes de regresión estimados para este modelo reducido son

$$b_0 = 36.094, \quad b_1 = 1.031, \quad b_2 = -1.870,$$

y la suma de cuadrados de la regresión resultante con 2 grados de libertad es

$$R(\beta_1, \beta_2) = 398.12.$$

Aquí se utiliza la notación $R(\beta_1, \beta_2)$ para indicar la suma de cuadrados de la regresión del modelo restringido, y no debe confundirse con SSR, la suma de cuadrados de la regresión del modelo original con 3 grados de libertad. Entonces, la nueva suma de errores al cuadrado es

$$SST - R(\beta_1, \beta_2) = 438.13 - 398.12 = 40.01,$$

y el error cuadrático medio resultante con 10 grados de libertad es

$$s^2 = \frac{40.01}{10} = 4.001.$$

¿Una prueba T de variable única tiene una contraparte F?

La cantidad de variación en la respuesta, el porcentaje de supervivencia, que se atribuye a x_3, porcentaje de peso del tercer aditivo, en presencia de las variables x_1 y x_2, es

$$R(\beta_3|\beta_1, \beta_2) = SSR - R(\beta_1, \beta_2) = 399.45 - 398.12 = 1.33,$$

que representa una proporción pequeña de toda la variación de la regresión. Esta cantidad de regresión agregada es estadísticamente insignificante, como lo indica la prueba previa sobre β_3. Una prueba equivalente implica la formación de la razón

$$f = \frac{R(\beta_3|\beta_1, \beta_2)}{s^2} = \frac{1.33}{4.298} = 0.309,$$

que es un valor de la distribución F con 1 y 9 grados de libertad. Recuerde que la relación básica entre la distribución t con v grados de libertad y la distribución F con 1 y v grados de libertad es

$$t^2 = f(1, v),$$

y se observa que el valor f de 0.309 en efecto es el cuadrado del valor t de -0.56.

Para generalizar los conceptos anteriores, podemos evaluar el funcionamiento de una variable independiente x_i en el modelo general de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

con la observación de la cantidad de regresión atribuida a x_i **sobre y por arriba de aquella atribuida a las demás variables**, es decir, la regresión sobre x_i *ajustada para las demás variables*. Ésta se calcula restando de *SSR* la suma de cuadrados de la regresión para un modelo del que se descartó x_i. Por ejemplo, se dice que x_1 se evalúa calculando

$$R(\beta_1|\beta_2, \beta_3, \ldots, \beta_k) = SSR - R(\beta_2, \beta_3, \ldots, \beta_k),$$

donde $R(\beta_2, \beta_3, \ldots, \beta_k)$ es la suma de cuadrados de la regresión con $\beta_1 x_1$ retirados del modelo. Para probar la hipótesis

$$H_0\text{: } \beta_1 = 0,$$
$$H_1\text{: } \beta_1 \neq 0,$$

se calcula

$$f = \frac{R(\beta_1|\beta_2, \beta_3, \ldots, \beta_k)}{s^2},$$

y se compara con $f_\alpha(1, n - k - 1)$.

En forma similar, se puede probar para la significancia de un *conjunto* de las variables. Por ejemplo, para investigar simultáneamente la importancia de incluir x_1 y x_2 en el modelo, se prueba la hipótesis

$$H_0\text{: } \beta_1 = \beta_2 = 0,$$
$$H_1\text{: } \beta_1 \text{ y } \beta_2 \text{ no son cero las dos,}$$

calculando

$$f = \frac{[R(\beta_1, \beta_2|\beta_3, \beta_4, \ldots, \beta_k)]/2}{s^2} = \frac{[SSR - R(\beta_3, \beta_4, \ldots, \beta_k)]/2}{s^2}$$

y comparando con $f_\alpha(2, n - k - 1)$. El número de grados de libertad asociados con el numerador, en este caso 2, es igual al número de variables en el conjunto que se investiga.

12.7 Caso especial de ortogonalidad (opcional)

Antes de nuestro desarrollo original del problema general de regresión lineal, se hizo la suposición de que las variables independientes eran medidas sin error y con frecuencia estaban controladas por el experimentador. A menudo ocurren como resultado de un *experimento diseñado* con laboriosidad. De hecho, se puede incrementar la eficacia de la ecuación de predicción resultante utilizando un plan de experimentación adecuado.

Suponga el lector que otra vez consideramos la matriz $\mathbf{X}$ según se definió en la sección 12.3. Podemos rescribirla para que se lea como

$$\mathbf{X} = [\mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}],$$

donde $\mathbf{1}$ representa una columna de unos y $\mathbf{x_j}$ es un vector columna que representa los niveles de x_j. Si

$$\mathbf{x_p'x_q} = \mathbf{0}, \quad \text{para} \quad p \neq q,$$

se dice que las variables x_p y x_q son *ortogonales* entre sí. Hay ciertas ventajas evidentes en tener una situación por completo ortogonal, en la cual $\mathbf{x_p'x_q} = \mathbf{0}$ para toda posible p y q, $p \neq q$ y, además,

$$\sum_{i=1}^{n} x_{ji} = 0, \quad j = 1, 2, \ldots, k.$$

la $\mathbf{X'X}$ resultante es una matriz diagonal, y las ecuaciones normales de la sección 12.3 se reducen a

$$\begin{aligned} nb_0 &= \sum_{i=1}^{n} y_i, \\ b_1 \sum_{i=1}^{n} x_{1i}^2 &= \sum_{i=1}^{n} x_{1i} y_i, \\ &\vdots \qquad\qquad \vdots \\ b_k \sum_{i=1}^{n} x_{ki}^2 &= \sum_{i=1}^{n} x_{ki} y_i. \end{aligned}$$

Una ventaja importante es que es fácil hacer la partición de *SSR* en **componentes de un solo grado de libertad**, cada uno de los cuales corresponde a la cantidad de variación de Y debida a una variable controlada establecida. En la situación ortogonal, se escribe

$$\begin{aligned} SSR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} - b_0)^2 \\ &= b_1^2 \sum_{i=1}^{n} x_{1i}^2 + b_2^2 \sum_{i=1}^{n} x_{2i}^2 + \cdots + b_k^2 \sum_{i=1}^{n} x_{ki}^2 \\ &= R(\beta_1) + R(\beta_2) + \cdots + R(\beta_k). \end{aligned}$$

La cantidad $R(\beta_i)$ es la cantidad de la suma de cuadrados de la regresión asociada con un modelo que implica una sola variable independiente x_i.

Para probar simultáneamente la significancia de un conjunto de m variables en una situación ortogonal, la suma de cuadrados de la regresión se convierte en

$$R(\beta_1, \beta_2, \ldots, \beta_m | \beta_{m+1}, \beta_{m+2}, \ldots, \beta_k) = R(\beta_1) + R(\beta_2) + \cdots + R(\beta_m),$$

y así se tiene la simplificación mayor

$$R(\beta_1|\beta_2, \beta_3, \ldots, \beta_k) = R(\beta_1)$$

cuando se evalúa una sola variable independiente. Por lo tanto, la contribución de una variable dada o conjunto de variables se encuentra en esencia al *ignorar* las demás variables del modelo. Las evaluaciones independientes del beneficio de las variables individuales se llevan a cabo usando las técnicas de análisis de varianza que se dan en la tabla 12.4. La variación total en la respuesta está en forma de la partición de componentes de un solo grado de libertad más el término del error con $n - k - 1$ grados de libertad. Cada valor f calculado se utiliza para probar una de las hipótesis

$$\left.\begin{array}{ll} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{array}\right\} \qquad i = 1, 2, \ldots, k,$$

al compararlo con el punto crítico $f_\alpha(1, n - k - 1)$ o con la sola interpretación del valor P calculado a partir de la distribución f.

Tabla 12.4: Análisis de varianza para variables ortogonales

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
β_1	$R(\beta_1) = b_1^2 \sum_{i=1}^{n} x_{1i}^2$	1	$R(\beta_1)$	$\frac{R(\beta_1)}{s^2}$
β_2	$R(\beta_2) = b_2^2 \sum_{i=1}^{n} x_{2i}^2$	1	$R(\beta_2)$	$\frac{R(\beta_2)}{s^2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_k	$R(\beta_k) = b_k^2 \sum_{i=1}^{n} x_{ki}^2$	1	$R(\beta_k)$	$\frac{R(\beta_k)}{s^2}$
Error	SSE	$n - k - 1$	$s^2 = \frac{SSE}{n-k-1}$	
Total	$SST = S_{yy}$	$n - 1$		

Ejemplo 12.8: Suponga que un científico recaba datos experimentales del radio de un grano propulsor Y como función de la temperatura del polvo x_1, la tasa de extrusión x_2 y la temperatura del molde x_3. Ajuste un modelo de regresión lineal para predecir el radio del grano y determine la eficacia de cada variable que interviene en el modelo. Los datos se presentan en la tabla 12.5.

Solución: Observe que cada variable está controlada en dos niveles, y que el experimento representa cada una de las ocho combinaciones posibles. Por conveniencia, los datos sobre las variables independientes están codificados mediante las siguientes fórmulas:

$$x_1 = \frac{\text{temperatura del polvo} - 170}{20},$$
$$x_2 = \frac{\text{tasa de extrusión} - 18}{6},$$
$$x_3 = \frac{\text{temperatura del molde} - 235}{15}.$$

Tabla 12.5: Datos para el ejemplo 12.8

Radio del grano	**Temperatura del polvo**	**Tasa de extrusión**	**Temperatura del molde**
82	150 (−1)	12 (−1)	220 (−1)
93	190 (+1)	12 (−1)	220 (−1)
114	150 (−1)	24 (+1)	220 (−1)
124	150 (−1)	12 (−1)	250 (+1)
111	190 (+1)	24 (+1)	220 (−1)
129	190 (+1)	12 (−1)	250 (+1)
157	150 (−1)	24 (+1)	250 (+1)
164	190 (+1)	24 (+1)	250 (+1)

Los niveles resultantes de x_1, x_2 y x_3 toman los valores -1 y $+l$, según se indica en la tabla de los datos. Este diseño experimental permite la ortogonalidad que se ilustra aquí. En el capítulo 15 se analiza un tratamiento más completo de este tipo de diseño experimental. La matriz $\mathbf{X}$ es

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

y las condiciones de ortogonalidad se verifican con facilidad. Ahora, pueden calcularse los coeficientes

$$b_0 = \frac{1}{8}\sum_{i=1}^{8} y_i = 121.75, \qquad b_1 = \frac{1}{8}\sum_{i=1}^{8} x_{1i}y_i = \frac{20}{8} = 2.5,$$

$$b_2 = \frac{\sum_{i=1}^{8} x_{2i}y_i}{8} = \frac{118}{8} = 14.75, \qquad b_3 = \frac{\sum_{i=1}^{8} x_{3i}y_i}{8} = \frac{174}{8} = 21.75,$$

de manera que, en términos de las variables codificadas, la ecuación de predicción es

$$\hat{y} = 121.75 + 2.5\,x_1 + 14.75\,x_2 + 21.75\,x_3.$$

La tabla 12.6, del análisis de varianza, presenta las contribuciones independientes de cada variable para *SSR*. Los resultados, al compararse con $f_{0.05}(1.4)$ para el punto crítico de 7.71, indican que x_1 no contribuye de manera significativa con el nivel de 0.05; mientras que las variables x_2 y x_3 sí son significativas. En este ejemplo, el estimador para σ^2 es 23.1250. Igual que en el caso para una sola variable independiente, se desprende que dicho estimador no únicamente contiene variación por el error experimental, a menos que el modelo postulado sea correcto. De otro modo, el estimador estará "contaminado" por la falta de ajuste, además del error puro y la

Tabla 12.6: Análisis de varianza para los datos del radio de los granos

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**	**Valor *P***
β_1	$(2.5)^2(8) = 50$	1	50	2.16	0.2156
β_2	$(14.75)^2(8) = 1740.50$	1	1740.50	75.26	0.0010
β_3	$(21.75)^2(8) = 3784.50$	1	3784.50	163.65	0.0002
Error	92.5	4	23.1250		
Total	5667.50	7			

falta de ajuste sólo puede separarse si se obtienen observaciones experimentales múltiples en las distintas combinaciones (x_1, x_2, x_3).

Como x_1 no es significativa, puede eliminarse sin más del modelo y no se alterarán los efectos de las otras variables. Observe que tanto x_2 como x_3 tienen un efecto sobre el radio del grano de manera positiva, con x_3 como el factor más importante con base en lo pequeño de su valor *P*.

Ejercicios

12.31 Calcule e interprete el coeficiente de determinación múltiple para las variables del ejercicio 12.3 de la página 453.

12.32 Pruebe si la regresión explicada por el modelo del ejercicio 12.3 de la página 453 es significativa con el nivel de significancia de 0.01.

12.33 Pruebe si la regresión explicada por el modelo del ejercicio 12.9 de la página 454 es significativa con el nivel de significancia de 0.01.

12.34 Para el modelo del ejercicio 12.9 de la página 454, pruebe la hipótesis

H_0: $\beta_1 = \beta_2 = 0$,
H_1: β_1 y β_2 no son cero las dos.

12.35 Repita el ejercicio 12.17 de la página 464 usando el estadístico *F*.

12.36 Se condujo un pequeño experimento para ajustar una ecuación de regresión múltiple que relaciona el producto y con la temperatura x_1, el tiempo de reacción x_2 y la concentración de uno de los reactivos x_3. Se eligieron dos niveles de cada variable y se registraron mediciones correspondientes a las variables independientes codificadas, como sigue:

y	x_1	x_2	x_3
7.6	−1	−1	−1
8.4	1	−1	−1
9.2	−1	1	−1
10.3	−1	−1	1
9.8	1	1	−1
11.1	1	−1	1
10.2	−1	1	1
12.6	1	1	1

a) Con las variables codificadas, estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

b) Partición *SSR*, suma de cuadrados de la regresión, en tres componentes de un solo grado de libertad atribuibles a x_1, x_2 y x_3, respectivamente. Construya una tabla de análisis de varianza, con pruebas de significancia sobre cada variable.

12.37 Considere los datos de energía eléctrica del ejercicio 12.9 de la página 454. Pruebe H_0: $\beta_1 = \beta_2 = 0$, utilizando $R(\beta_1, \beta_2|\beta_3, \beta_4)$. Proporcione un valor *P* y saque conclusiones.

12.38 Considere los datos para el ejercicio 12.36. Calcule lo siguiente:

$$R(\beta_1|\beta_0), \qquad R(\beta_1|\beta_0, \beta_2, \beta_3),$$
$$R(\beta_2|\beta_0, \beta_1), \qquad R(\beta_2|\beta_0, \beta_1, \beta_3),$$
$$R(\beta_3|\beta_0, \beta_1, \beta_2).$$

Haga comentarios.

12.39 Considere los datos del ejercicio 11.63 de la página 439. Ajuste un modelo de regresión utilizando la razón de manejo y el peso como variables explicativas. Compare ese modelo con el de *RLS* (regresión lineal simple) únicamente con el empleo del peso. Utilice R^2, R^2_{aju} y cualesquiera estadísticos t (o F) que necesite para comparar la *RLS* con el modelo de regresión múltiple.

12.40 Considere el ejemplo 12.3. La figura 12.1 en la página 462 muestra una salida de *SAS* para un análisis del modelo que contiene las variables x_1, x_2 y x_3. Céntrese en el intervalo de confianza de la respuesta media μ_Y en las ubicaciones (x_1, x_2, x_3) que representan los 13 puntos de los datos. Considere el concepto que aparece en la salida indicado por C.V. Ése es el **coeficiente de variación**, que se define como

$$\text{C.V.} = \frac{s}{\bar{y}} \cdot 100,$$

donde $s = \sqrt{s^2}$ es la **raíz del error cuadrático medio**. El coeficiente de variación se utiliza con frecuencia como otro criterio para comparar modelos en competencia. Se trata de una cantidad sin escalas que expresa al estimador de σ, es decir s, como un porcentaje de la respuesta promedio $\bar{y}$. En competencia por el "mejor" modelo de un grupo de ellos en competencia, se prefiere aquel con un valor "pequeño" de C.V. Haga un análisis de regresión del conjunto de datos que se muestra en el ejemplo 12.3, pero elimine x_3. Compare el modelo completo (x_1, x_2, x_3) con el restringido (x_1, x_2) y céntrese en dos criterios: **i.** C.V.; **ii.** los anchos de los intervalos de confianza sobre μ_Y. Para el segundo criterio usted quizá desearía usar el ancho promedio. Haga comentarios.

12.41 Considere el ejemplo 12.4 de la página 451. Compare los dos modelos en competencia

$$\begin{aligned}
&\text{Primer orden:} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i,\\
&\text{Segundo orden:} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\\
&\qquad\qquad + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i.
\end{aligned}$$

En su comparación emplee R^2_{aju}, además de probar $H_0 : \beta_{11} = \beta_{22} = \beta_{12} = 0$. También utilice C.V. según lo hizo en el ejercicio 12.40.

12.42 En el ejemplo 12.8 se trata el caso de eliminar del modelo x_1, la temperatura del polvo, ya que el valor P basado en la prueba F es 0.2156, en tanto que los valores P para x_1 y x_2 son casi cero.

a) Reduzca el modelo con la eliminación de x_1, y genere un modelo completo y restringido (o reducido), y compárelos sobre la base de R^2_{aju}.

b) Compare los modelos completo y restringido usando intervalos de predicción de 95% de ancho sobre una nueva observación. El "mejor" de ambos modelos será aquel con intervalos de predicción más "estrechos". Utilice el promedio del ancho de los intervalos de predicción.

12.43 Considere los datos del ejercicio 12.15 de la página 455. ¿Puede explicarse la respuesta, el uso, en forma adecuada mediante una sola variable (sea la viscosidad o la carga) con una *RLS* en vez de con la regresión completa con dos variables? Justifique su respuesta con pruebas de hipótesis, así como con la comparación de los tres modelos en competencia.

12.44 Para el conjunto de datos que se da en el ejercicio 12.16 de la página 455, ¿es posible explicar la respuesta en forma adecuada usando dos variables regresoras cualesquiera? Analice.

12.8 Variables categóricas o indicadoras

Un caso especial de aplicación muy importante de la regresión lineal múltiple ocurre cuando una o más de las variables regresoras son **categóricas** o **indicadoras**. En un proceso químico, el ingeniero quizá desee modelar el producto contra regresores tales como la temperatura del proceso y el tiempo de reacción. Sin embargo, sería de interés el uso de dos catalizadores diferentes y, por ello, incluir de algún modo el "catalizador" en el modelo. El efecto del catalizador no puede medirse sobre un continuo, por lo que no es una variable categórica. Un analista podría desear modelar el precio de casas contra regresores que incluyeran los pies cuadrados de superficie habitable, x_1, la superficie del terreno, x_2, y la edad de la vivienda, x_3. Estos regresores son de naturaleza claramente continua. Sin embrago, es evidente que el costo de las casas podría variar en forma sustancial de una zona del país a otra. Así, podrían recabarse datos sobre las casas en el este, el medioeste, el sur y el oeste. Como resultado, se tiene una variable indicadora con **cuatro categorías**. En el ejemplo del proceso químico, si se usaran dos catalizadores se tendría una variable indicadora

con dos categorías. En un ejemplo biomédico, un fármaco se compara con un placebo y se realizan diversas mediciones continuas sobre todos los sujetos, tales como su edad y su presión sanguínea, entre otras, al igual que el género, que por supuesto es una variable categórica con dos categorías. Entonces, incluidas junto con las variables continuas existen dos variables indicadoras, el tratamiento con dos categorías (fármaco activo y placebo) y el género con dos categorías (masculino y femenino).

Modelo con variables categóricas

Para ilustrar la forma en que las variables indicadoras entran en el modelo, utilizaremos el ejemplo del procesamiento químico. Suponga que $y =$ producto, $x_1 =$ temperatura y $x_2 =$ tiempo de reacción. Ahora denotaremos con z la variable indicadora. Sea $z = 0$ para el catalizador 1 y $z = 1$ para el catalizador 2. La asignación del indicador (0, 1) al catalizador es arbitraria. Como resultado, el modelo se convierte en

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 z_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Tres categorías

Lo que sigue es la estimación de los coeficientes con el método de mínimos cuadrados. En el caso de tres niveles o categorías de una sola variable indicadora, el modelo incluirá **dos** regresores, por ejemplo z_1 y z_2, donde la asignación (0, 1) es como sigue:

$$\begin{array}{cc} z_1 & z_2 \\ \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \cdots & \cdots \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \\ \cdots & \cdots \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{array}\right] \end{array}$$

En otras palabras, si hay ℓ categorías, el modelo incluye $\ell - 1$ términos reales.

Puede ser instructivo observar la apariencia gráfica del modelo con 3 categorías. En aras de la simplicidad, se considerará una sola variable continua x.
Como resultado, el modelo está dado por

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \epsilon_i.$$

Así, la figura 12.2 refleja la naturaleza del modelo. Las siguientes son expresiones del modelo para las tres categorías.

$$\begin{aligned} E(Y) &= (\beta_0 + \beta_2) + \beta_1 x, && \text{categoría 1,} \\ E(Y) &= (\beta_0 + \beta_3) + \beta_1 x, && \text{categoría 2,} \\ E(Y) &= \beta_0 + \beta_1 x, && \text{categoría 3.} \end{aligned}$$

Como resultado, el modelo que incluye variables categóricas en esencia implica un **cambio en la intersección** conforme se pasa de una categoría a otra. Desde luego, aquí se supone que los **coeficientes de las variables continuas son las mismas a través de las categorías**.

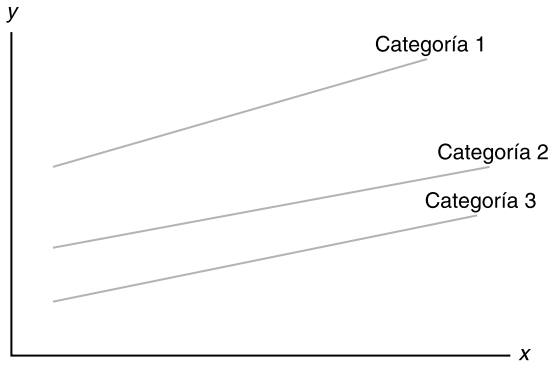


Figura 12.2: Caso de tres categorías.

Ejemplo 12.9: Considere los datos de la tabla 12.7. La respuesta y es la cantidad de sólidos en suspensión en un sistema de limpieza de carbón. La variable x es el pH del sistema. En éste se utilizan tres polímeros diferentes. Así, "polímero" es categórico con tres categorías, por lo que produce dos términos en el modelo, el cual está dado por

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \epsilon_i, \qquad i = 1, 2, \ldots, 18.$$

Aquí se tiene que

$$z_1 = \begin{cases} 1, & \text{para el polímero 1,} \\ 0, & \text{en cualquier otro caso,} \end{cases} \quad \text{y} \quad z_2 = \begin{cases} 1, & \text{para el polímero 2,} \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Serán de provecho algunos comentarios acerca de las conclusiones que se obtengan del análisis de la figura 12.3. El coeficiente b_1 para el pH es el estimador de la **pendiente común** que se acepta en el análisis de regresión. Todos los términos del modelo son estadísticamente significativos. Así, el pH y la naturaleza del polímero tienen un efecto sobre la cantidad de limpieza. Los signos y las magnitudes de los coeficientes de z_1 y z_2 indican que el polímero 1 es más eficaz (produce más sólidos en suspensión) en cuanto a la limpieza, seguido del polímero 2. El polímero 3 es el menos eficaz.

Tabla 12.7: Datos para el ejemplo 12.9

x (pH)	y (cantidad de sólidos en suspensión)	Polímero
6.5	292	1
6.9	329	1
7.8	352	1
8.4	378	1
8.8	392	1
9.2	410	1
6.7	198	2
6.9	227	2
7.5	277	2
7.9	297	2
8.7	364	2
9.2	375	2
6.5	167	3
7.0	225	3
7.2	247	3
7.6	268	3
8.7	288	3
9.2	342	3

```
                          Sum of
          Source  DF       Squares  Mean Square    F Value    Pr > F
           Model   3   80181.73127  26727.24376      73.68    <.0001
           Error  14    5078.71318    362.76523
Corrected Total  17   85260.44444

 R-Square     Coeff Var      Root MSE        y Mean
 0.940433      6.316049      19.04640      301.5556
                                            Standard
Parameter     Estimate        Error  t Value  Pr > |t|
Intercept -161.8973333  37.43315576    -4.32    0.0007
        x   54.2940260   4.75541126    11.42    <.0001
       z1   89.9980606  11.05228237     8.14    <.0001
       z2   27.1656970  11.01042883     2.47    0.0271
```

Figura 12.3: Salida del *SAS* para el ejemplo 12.9.

La pendiente puede variar con las categorías indicadoras

En el análisis efectuado hasta el momento, se ha supuesto que los términos de las variables indicadoras entran al modelo en forma aditiva, lo cual sugiere que las pendientes, como las que se aprecia en la figura 12.2, son constantes a través de las categorías. Es evidente que éste no siempre será el caso. Existe la posibilidad de que las pendientes varíen y, con ello, las pruebas para esta condición de **paralelismo** por la inclusión de términos de producto o **interacción** entre los términos indicadores y

las variables continuas. Por ejemplo, suponga que se elige un modelo con un regresor continuo y una variable indicadora con dos niveles. Se tiene el modelo

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon.$$

Éste sugiere que para la categoría 1 ($z = 1$),

$$E(y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x,$$

mientras que para la categoría 2 ($z = 0$),

$$E(y) = \beta_0 + \beta_1 x.$$

Así, se permite que varíen la intersección y las pendientes para las dos categorías. La figura 12.4 muestra las rectas de regresión con pendientes variables para las dos categorías.

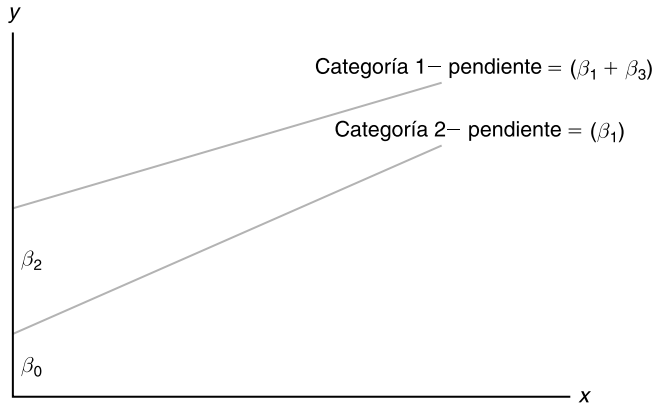


Figura 12.4: Falta de paralelismo en las variables categóricas.

En este caso, β_0, β_1 y β_2 son positivas; mientras que β_3 es negativa con $|\beta_3| < \beta_1$. Es evidente que si el coeficiente de interacción β_3 es insignificante, se está de vuelta en el modelo común de la pendiente.

Ejercicios

12.45 Se realizó un estudio para evaluar la eficacia en cuanto al costo de manejar un automóvil sedán de cuatro puertas en vez de una van o una SUV (vehículo deportivo utilitario). Las variables continuas son la lectura del odómetro y el octanaje de la gasolina empleada. La variable de respuesta está en millas por galón. Los datos se presentan a continuación.

a) Ajuste un modelo de regresión lineal que incluya dos variables indicadoras. Utilice 0, 0 para denotar al sedán de cuatro puertas.

b) ¿Qué tipo de vehículo parece desempeñarse mejor en cuanto a la distancia recorrida por unidad de gasolina?

c) Analice la diferencia entre una van y una SUV en términos del rendimiento de la gasolina en cuanto a la distancia.

MPG	Tipo de carro	Odómetro	Octanaje
34.5	sedán	75000	87.5
33.3	sedán	60000	87.5
30.4	sedán	88000	78.0
32.8	sedán	15000	78.0
35.0	sedán	25000	90.0
29.0	sedán	35000	78.0
32.5	sedán	102000	90.0
29.6	sedán	98000	87.5
16.8	van	56000	87.5
19.2	van	72000	90.0
22.6	van	14500	87.5
24.4	van	22000	90.0
20.7	van	66500	78.0
25.1	van	35000	90.0
18.8	van	97500	87.5
15.8	van	65500	78.0
17.4	van	42000	78.0
15.6	SUV	65000	78.0
17.3	SUV	55500	87.5
20.8	SUV	26500	87.5
22.2	SUV	11500	90.0
16.5	SUV	38000	78.0
21.3	SUV	77500	90.0
20.7	SUV	19500	78.0
24.1	SUV	87000	90.0

12.46 Se efectuó un estudio para determinar si el género del titular de la tarjeta de crédito era un factor importante con respecto a la generación de utilidades para cierta compañía de tarjetas de crédito. Las variables consideradas fueron el ingreso, el número de miembros de la familia y el género del titular de la tarjeta. Los datos son los siguientes:

Utilidad	Ingreso	Género	Miembros de la familia
157	45000	M	1
−181	55000	M	2
−253	45800	M	4
158	38000	M	3
75	75000	M	4
202	99750	M	4
−451	28000	M	1
146	39000	M	2
89	54350	M	1
−357	32500	M	1
522	36750	F	1
78	42500	F	3
5	34250	F	2
−177	36750	F	3
123	24500	F	2
251	27500	F	1
−56	18000	F	1
453	24500	F	1
288	88750	F	1
−104	19750	F	2

a) Ajuste un modelo de regresión lineal usando las variables disponibles. Con base en el modelo ajustado, ¿la compañía preferiría clientes del género masculino o del femenino?

b) ¿Diría usted que el ingreso fue un factor importante para explicar la variabilidad de la utilidad?

12.9 Métodos secuenciales para la selección del modelo

A veces, las pruebas de significancia estudiadas en la sección 12.6 son muy adecuadas para determinar cuáles variables deben usarse en el modelo final de la regresión. Dichas pruebas sin duda son eficaces si el experimento puede planearse y las variables son ortogonales entre sí. Incluso cuando las variables no sean ortogonales, las pruebas t individuales pueden usarse en muchos problemas donde es pequeño el número de variables que se investiga. No obstante, hay problemas en que es necesario utilizar técnicas más elaboradas para estudiar a las variables, en particular si el experimento muestra una desviación sustancial de la ortogonalidad. Los coeficientes de correlación de la muestra, $r_{x_i x_j}$, proporcionan mediciones útiles de **multicolinealidad** (dependencia lineal) entre las variables independientes. Como sólo nos ocupa la dependencia lineal entre variables independientes, no hay confusión, se eliminan las x de la notación y únicamente se escribe $r_{x_i x_j} = r_{ij}$, donde

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}.$$

Observe que, en sentido estricto, las r_{ij} no dan estimadores verdaderos de los coeficientes de correlación de la población, ya que las x en realidad no son variables

aleatorias en el contexto que se estudia aquí. Así, el término *correlación*, aunque estándar, quizá sea inadecuado.

Cuando uno o más de dichos coeficientes de correlación muestral se desvía de manera sustancial de cero, sería muy difícil encontrar el subconjunto de variables más eficaz para su inclusión en la ecuación de pronóstico. De hecho, en ciertos problemas la multicolinealidad será tan extrema que no podría hallarse un predictor adecuado, a menos que se investiguen todos los subconjuntos posibles de variables. En la bibliografía se mencionan los análisis informativos de Hocking para la selección de modelos de regresión. En el libro de Myers (1990), también citado, se estudian procedimientos para detectar la multicolinealidad.

El usuario de la regresión lineal múltiple busca lograr uno de tres objetivos:

1. Obtener estimadores de coeficientes individuales en un modelo completo.

2. Estudiar variables para determinar cuáles tienen un efecto significativo sobre la respuesta.

3. Llegar a la ecuación de pronóstico más eficaz.

En el punto **1.**, se sabe de antemano que todas las variables deben incluirse en el modelo. En el **2.**, la predicción es secundaria; mientras que en el **3.**, los coeficientes de regresión individuales no son tan importantes como la calidad de la respuesta estimada $\hat{y}$. Para cada una de las situaciones anteriores, la multicolinealidad en el experimento llega a tener un efecto profundo sobre el éxito de la regresión.

En esta sección se estudian algunos procedimientos secuenciales estándar para seleccionar variables, los cuales se basan en el concepto de que una sola variable o una colección de ellas no debería aparecer en la ecuación de estimación, a menos que origine un incremento significativo en la suma de cuadrados de la regresión o, en forma equivalente, un incremento significativo de R^2, el coeficiente de determinación múltiple.

Ilustración del estudio de las variables en presencia de colinealidad

Ejemplo 12.10: Considere los datos de la tabla 12.8, que muestra mediciones de 9 bebés. El propósito del experimento era llegar a una ecuación de estimación apropiada que relacionara la longitud del bebé con todas las variables independientes o un subconjunto de ellas. Los coeficientes de correlación muestral, que indican la dependencia lineal entre las variables independientes, se muestran en la matriz simétrica

$$\begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} 1.0000 & 0.9523 & 0.5340 & 0.3900 \\ 0.9523 & 1.0000 & 0.2626 & 0.1549 \\ 0.5340 & 0.2626 & 1.0000 & 0.7847 \\ 0.3900 & 0.1549 & 0.7847 & 1.0000 \end{bmatrix} \end{array}$$

Observe que parece haber una cantidad apreciable de multicolinealidad. Con la técnica de mínimos cuadrados que se describió en la sección 12.2, se ajustó la ecuación de regresión estimada usando el modelo completo:

$$\hat{y} = 7.1475 + 0.1000x_1 + 0.7264x_2 + 3.0758x_3 - 0.0300x_4.$$

El valor de s^2 con 4 grados de libertad es 0.7414, y el valor para el coeficiente de determinación para este modelo resulta 0.9908. En la tabla 12.9 se dan la suma

Tabla 12.8: Datos relacionados con la longitud de un bebé*

Longitud del bebé, y (cm)	**Edad, x_1 (días)**	**Longitud al nacer, x_2 (cm)**	**Peso al nacer x_3 (kg)**	**Tamaño del pecho al nacer, x_4 (cm)**
57.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67.0	88	49.0	5.52	36.5
53.5	67	43.0	3.21	27.2
62.7	80	48.0	4.32	27.7
56.2	74	48.0	2.31	28.3
68.5	94	53.0	4.30	30.3
69.2	102	58.0	3.71	28.7

*Datos analizados por el Statistical Consulting Center, Instituto Politécnico y Universidad Estatal de Virginia, Blacksburg, Virginia.

Tabla 12.9: Valores t para los datos de regresión de la tabla 12.8

Variable x_1	**Variable x_2**	**Variable x_3**	**Variable x_4**
$R(\beta_1\|\beta_2,\beta_3,\beta_4)$	$R(\beta_2\|\beta_1,\beta_3,\beta_4)$	$R(\beta_3\|\beta_1,\beta_2,\beta_4)$	$R(\beta_4\|j_1,\beta_2,\beta_3)$
$= 0.0644$	$= 0.6334$	$= 6.2523$	$= 0.0241$
$t = 0.2947$	$t = 0.9243$	$t = 2.9040$	$t = -0.1805$

de cuadrados de la regresión que mide la variación atribuida a cada variable individual en presencia de las demás, y los valores t correspondientes.

Una región crítica de doble cola con 4 grados de libertad con un nivel de significancia de 0.05 ocurre en $|t| > 2.776$. De los cuatro valores t calculados, **sólo la variable x_3 parece ser significativa**. Sin embargo, hay que recordar que aunque el estadístico t descrito en la sección 12.6 mide el beneficio que aporta una variable ajustada a todas las demás, no detecta la importancia potencial de una variable en combinación con un subconjunto de ellas. Por ejemplo, considere el modelo con solo las variables x_2 y x_3 en la ecuación. El análisis de los datos de la función de regresión

$$\hat{y} = 2.1833 + 0.9576x_2 + 3.3253x_3,$$

con $R^2 = 0.9905$, que por cierto no es una reducción sustancial de $R^2 = 0.9907$ para el modelo completo. Sin embargo, a menos que las características de desempeño de esta combinación particular hayan sido observadas, no se estaría al tanto de su potencial predictivo. Esto, por supuesto, apoya una metodología que observe *todas las regresiones posibles*, o un procedimiento secuencial sistemático diseñado para probar subconjuntos diferentes.

Regresión progresiva

Un procedimiento estándar para buscar el "subconjunto óptimo" de variables en ausencia de ortogonalidad es una técnica denominada **regresión progresiva**. Se basa en el procedimiento de introducir en forma secuencial las variables al modelo,

una por una. La descripción de la rutina progresiva se entenderá mejor si en primer lugar se describen los métodos de **selección hacia delante** y **selección hacia atrás**.

La **selección hacia delante** se basa en el concepto de que las variables deben insertarse una a la vez, hasta que se encuentre una ecuación de regresión satisfactoria. El procedimiento es como sigue:

PASO 1. Elija la variable que dé la mayor suma de cuadrados de la regresión, cuando se ejecute la regresión lineal simple con y o, en forma equivalente, aquella que dé el mayor valor de R^2. Esta variable inicial se llamará x_1.

PASO 2. Seleccione la variable que cuando entra al modelo da el incremento mayor de R^2, en presencia de x_1, sobre la R^2 hallada en el paso 1. Ésta, por supuesto, es la variable x_J, para la que

$$R(\beta_j|\beta_1) = R(\beta_1, \beta_j) - R(\beta_1)$$

es más grande. Dicha variable se llamará x_2. El modelo de regresión con el que x_1 y x_2 entonces es ajustado y R^2 observado.

PASO 3. Elija la variable x_j que da el valor más grande de

$$R(\beta_j|\beta_1, \beta_2) = R(\beta_1, \beta_2, \beta_j) - R(\beta_1, \beta_2),$$

otra vez resulta en el incremento mayor de R^2 sobre aquel obtenido en el paso 2. Esta variable se llamará x_3, y ahora se tiene un modelo de regresión que incluye x_1, x_2 y x_3.

Este proceso continúa hasta que la variable más reciente que ingresó ya no induce un incremento significativo en la regresión explicada. Tal incremento puede determinarse en cada paso con el uso adecuado de una prueba F o una t. Por ejemplo, en el paso 2, el valor

$$f = \frac{R(\beta_2|\beta_1)}{s^2}$$

se determina para probar la pertinencia de x_2 en el modelo. Aquí, el valor de s^2 es el error cuadrático medio para el modelo que contiene las variables x_1 y x_2. De manera similar, en el paso 3, la razón

$$f = \frac{R(\beta_3|\beta_1, \beta_2)}{s^2}$$

prueba la pertinencia de x_3 en el modelo. Sin embargo, ahora el valor de s^2 es el error cuadrático medio para el modelo que contiene las tres variables x_1, x_2 y x_3. Si en el paso 2, $f < f_\alpha\ (1, n-3)$ para un nivel de significancia preseleccionado, x_2 no está incluida y el proceso finaliza, lo que da como resultado una ecuación lineal simple que relaciona y y x_1. Sin embargo, si $f > f_\alpha(1, n-3)$, se pasa al paso 3. De nuevo si, en el paso 3, $f < f_\alpha(1, n-4)$ x_3 no queda incluida y el proceso concluye con la ecuación de regresión apropiada que contiene las variables x_1 y x_2.

La **eliminación hacia atrás** implica los mismos conceptos que la selección hacia delante, excepto que se comienza todas las variables en el modelo. Por ejemplo, suponga que hay cinco variables en consideración. Los pasos son:

PASO 1. Ajuste una ecuación de regresión con las cinco variables incluidas en el modelo. Elija la variable que dé el valor más pequeño de la suma de cuadrados de la regresión **ajustada para las demás**. Suponga que dicha variable es x_2. Elimine x_2 del modelo si

$$f = \frac{R(\beta_2|\beta_1, \beta_3, \beta_4, \beta_5)}{s^2}$$

es insignificante.

PASO 2. Ajuste una ecuación de regresión que use las variables restantes x_1, x_3, x_4 y x_5, y repita el paso 1. Suponga que esta vez se elige la variable x_5. Otra vez, si

$$f = \frac{R(\beta_5|\beta_1, \beta_3, \beta_4)}{s^2}$$

es insignificante, se retira del modelo la variable x_5. En cada paso, la s^2 que se usa en la prueba F es el error cuadrático medio para el modelo de regresión en esa etapa.

Este proceso se repite hasta que en algún paso la variable con la suma de cuadrados de la regresión ajustada da como resultado un valor f significativo para algún nivel de significancia predeterminado.

La **regresión progresiva** se lleva a cabo con una modificación ligera pero importante del procedimiento de selección hacia delante. La modificación requiere efectuar más pruebas en cada etapa, para garantizar la eficacia continuada de las variables que se hubieran incluido en el modelo durante alguna etapa anterior. Esto representa una mejoría sobre la selección hacia delante, ya que es muy posible que una variable que haya entrado a la ecuación de regresión en una etapa temprana podría carecer de importancia, o ser redundante, debido a las relaciones que existen entre ella y las demás variables de las etapas posteriores. Por lo tanto, en una etapa en que una variable nueva haya ingresado a la ecuación de regresión con un incremento significativo de R^2 según lo determina la prueba F, todas las variables que ya estén en el modelo quedan sujetas a pruebas F (o, en forma equivalente, a pruebas t) a la luz de esta variable nueva, y si no muestran un valor f significativo, se eliminan. El procedimiento continúa hasta que se alcance una etapa donde no puedan insertarse ni eliminarse variables adicionales. Este procedimiento hacia delante se ilustra con el siguiente ejemplo.

Ejemplo 12.11: Usando técnicas de regresión progresiva, encuentre un modelo de regresión lineal adecuado para predecir la longitud de los bebés cuyos datos se presentan en la tabla 12.8.

Solución: PASO 1. Se considera cada variable por separado y se ajustan cuatro ecuaciones individuales de regresión lineal simple. Se calculan las siguientes sumas de cuadrados de la regresión que son pertinentes:

$$R(\beta_1) = 288.1468, \qquad R(\beta_2) = 215.3013,$$
$$R(\beta_3) = 186.1065, \qquad R(\beta_4) = 100.8594.$$

Es claro que la variable x_1 da la suma de cuadrados de la regresión más elevada. El error cuadrático medio para la ecuación que implica sólo x_1 es $s^2 = 4.7276$ y como

$$f = \frac{R(\beta_1)}{s^2} = \frac{288.1468}{4.7276} = 60.9500,$$

que excede $f_{0.05}(1.7) = 5.59$, se introduce al modelo la variable x_1.

PASO 2. En esta etapa se ajustan tres ecuaciones de regresión, todas las cuales contienen x_1. Los resultados importantes para las combinaciones (x_1, x_2), (x_1, x_3) y (x_1, x_4) son

$$R(\beta_2|\beta_1) = 23.8703, \quad R(\beta_3|\beta_1) = 29.3086, \quad R(\beta_4|\beta_1) = 13.8178.$$

La variable x_3 muestra la mayor suma de cuadrados de la regresión en presencia de x_1. La regresión que implica x_1 y x_3 da un valor nuevo de $s^2 = 0.6307$ y como

$$f = \frac{R(\beta_3|\beta_1)}{s^2} = \frac{29.3086}{0.6307} = 46.47,$$

que excede $f_{0.05}(1, 6) = 5.99$, la variable x_3 se incluye en el modelo junto con x_1. Ahora, debemos someter a x_1 a una prueba de significancia en la presencia de x_3. Encontramos que $R(\beta_1|\beta_3) = 131.349$, por lo que

$$f = \frac{R(\beta_1|\beta_3)}{s^2} = \frac{131.349}{0.6307} = 208.26,$$

que es muy significativa. Por lo tanto, se mantiene x_1 junto con x_3.

PASO 3. Con x_1 y x_3 ya en el modelo, ahora se requiere $R(\beta_2|\beta_1\beta_3)$ y $R(\beta_4|\beta_1\beta_3)$, con la finalidad de determinar cuál, si alguna, de las dos variables restantes debe entrar en esta etapa. Del análisis de regresión, usando x_2 junto con x_1 y x_3, se encuentra que $R(\beta_2|\beta_1\beta_3) = 0.7948$, y cuando x_4 se utiliza con x_1 y x_3 se obtiene $R(\beta_4|\beta_1\beta_3) = 0.1855$. El valor de s^2 es 0.5979 para la combinación (x_1, x_2, x_3), y de 0.7198 para la combinación (x_1, x_2, x_4). Como ningún valor f es significativo con el nivel $\alpha = 0.05$, el modelo final de regresión sólo incluye las variables x_1 y x_3. Se encuentra que la ecuación de estimación es

$$\hat{y} = 20.1084 + 0.4136x_1 + 2.0253x_3,$$

y el coeficiente de determinación para este modelo es $R^2 = 0.9882$.

Aunque (x_1, x_3) es la combinación elegida por la regresión progresiva, no necesariamente es la combinación de dos variables que da el valor más grande de R^2. De hecho, ya se observó que la combinación (x_2, x_3) da un valor de $R^2 = 0.9905$. Desde luego, el procedimiento progresivo nunca observó en realidad dicha combinación. Podría plantearse un argumento racional de que en realidad hay una diferencia despreciable en el desempeño entre esas dos ecuaciones de estimación, al menos en términos del porcentaje de variación explicado. Sin embargo, es interesante observar que la eliminación hacia atrás da la combinación (x_2, x_3) en la ecuación final (véase el ejercicio 12.49 en la página 496).

Resumen

La función principal de cada uno de los procedimientos explicados en esta sección consiste en exponer las variables a una metodología sistemática, diseñada para garantizar la inclusión final de las combinaciones mejores de ellas. Es evidente que no

es seguro que esto pase en todos los problemas, y, por supuesto, es posible que la multicolinealidad sea tan extensa que no haya más alternativa que apoyarse en procedimientos de estimación diferentes de los mínimos cuadrados. Tales procedimientos de estimación se estudian en Myers (1990), citado en la bibliografía.

Los procedimientos secuenciales que se estudian aquí representan tres de muchos métodos parecidos que aparecen en la bibliografía y para los cuales se dispone de varios paquetes de regresión por computadora. Estos métodos fueron diseñados para ser eficientes en cuanto a computación, aunque, desde luego, no dan resultados para todos los subconjuntos posibles de las variables. Como resultado, los procedimientos son más eficaces en conjuntos de datos que incluyen un **número grande de variables**. En problemas de regresión que implican un número relativamente pequeño de variables, los paquetes modernos de cómputo para la regresión permiten el cálculo y resumen la información cuantitativa de todos los modelos para cada subconjunto posible de variables. En la sección 12.11 se dan ilustraciones de ello.

12.10 Estudio de los residuos y trasgresión de las suposiciones (verificación del modelo)

En un punto anterior de este capítulo se sugirió que los residuos, o errores en el ajuste de regresión, con frecuencia dan información muy valiosa para el analista de los datos. Los $e_i = y_i - \hat{y}_i,\ i = 1, 2, \ldots, n$, que son la contraparte numérica de los ϵ_i, los errores del modelo, arrojan luz sobre la posible trasgresión de las suposiciones o la presencia de datos de puntos "sospechosos". Suponga que el vector $\mathbf{x_i}$ denota los valores de las variables regresoras que corresponden al i-ésimo dato de los puntos, que incluye un 1 en la posición inicial. Es decir,

$$\mathbf{x}_i' = [1, x_{1i}, x_{2i}, \ldots, x_{ki}].$$

Considere la cantidad

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i, \qquad i = 1, 2, \ldots, n.$$

El lector debería notar que h_{ii} se utilizó en la sección 12.5 para calcular los intervalos de confianza sobre la respuesta media. Además de σ^2, h_{ii} representa la varianza del valor ajustado $\hat{y}_i$. Los valores h_{ii} son los elementos de la diagonal de la **matriz TESTADA**

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

que desempeña un papel importante en cualquier estudio de los residuos y en otros aspectos modernos del análisis de regresión (véase la referencia a Myers, 1990, en la bibliografía). El término *matriz TESTADA* se deriva del hecho de que $\mathbf{H}$ genera las "y testadas", o valores ajustados cuando se multiplica por el vector $\mathbf{y}$ de respuestas observadas. Es decir, $\hat{\mathbf{y}} = \mathbf{Xb}$, por lo que

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Hy},$$

donde $\hat{\mathbf{y}}$ es el vector cuyo i-ésimo elemento es $\hat{y}_i$.

Si se hacen las suposiciones usuales de que los ϵ_i son independientes y están distribuidos normalmente con media cero y varianza σ^2, las propiedades estadísticas de los residuos quedan caracterizadas con facilidad. Entonces,

$$E(e_i) = E(y_i - \hat{y}_i) = 0, \quad \text{y} \quad \sigma^2_{\epsilon_i} = (1 - h_{ii})\sigma^2,$$

para $i = 1, 2, \ldots, n$. (Para mayores detalles, véase Myers, 1990.) Puede demostrarse que los valores de la diagonal de la matriz TESTADA están acotados de acuerdo con la desigualdad

$$\frac{1}{n} \leq h_{ii} \leq 1.$$

Además, $\sum_{i=1}^{n} h_{ii} = k + 1$, el número de parámetros de la regresión. Como resultado, cualquier dato de punto cuyo elemento diagonal TESTADA sea grande, es decir, muy por encima del valor promedio de $(k + 1)/n$, está en una posición dentro del conjunto de datos donde la varianza de $\hat{y}_i$ es relativamente grande, y la varianza de un residuo es relativamente pequeña. Como resultado, el analista de datos puede obtener alguna perspectiva de qué tan grande puede ser un residuo antes de que su desviación de cero se atribuya a algo distinto de la mera aleatoriedad. Muchos de los paquetes comerciales para computadora sobre la regresión producen el conjunto de **residuos studentizados***.

Residuo studentizado

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \ldots, n$$

Aquí, cada residuo se dividió en una **estimación de su desviación estándar**, con la creación de un estadístico *tipo t* diseñado para dar al analista una cantidad libre de escala, que proporcione información sobre el *tamaño* del residuo. Además, es frecuente que los paquetes de cómputo estándar proporcionen valores de otro conjunto de residuos tipo studentizados, denominados **valores *R* de Student**.

Residuo *R* de Student

$$t_i = \frac{e_i}{s_{-i}\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \ldots, n$$

donde s_{-i} es un estimador de la desviación estándar del error, calculado con el ***i*-ésimo dato de los puntos eliminado**.

Hay tres tipos de trasgresiones de las suposiciones que se detectan con facilidad utilizando los residuos o las *gráficas de residuos*. En tanto que las gráficas de residuos crudos, los e_i, son de ayuda, con frecuencia es más informativo graficar los residuos studentizados. Las tres trasgresiones son como sigue:

1. Presencia de valores extremos.
2. Varianza del error heterogénea.
3. Mala especificación del modelo.

En el caso 1, elegimos definir un **valor extremo** como dato de punto que tiene una desviación de la suposición usual de que $E(\epsilon_i) = 0$ para un valor específico de i. Si hay una razón para creer que un dato de punto específico es un valor extremo y ejerce una influencia grande sobre el modelo ajustado, r_i o t_i pueden dar información. Es de esperarse que los valores R de Student sean más sensibles a los valores extremos que los valores r_i.

En realidad, en condiciones en que $E(\epsilon_i) = 0$, t_i es un valor de una variable aleatoria que sigue una distribución t con $n - 1 - (k + 1) = n - k - 2$ grados de

*Por *Student*, seudónimo del autor de la distribución t. NT.

libertad. Así, es posible utilizar una prueba t de dos colas para obtener información para detectar si el punto i-ésimo es un valor extremo o no.

Aunque el estadístico R de Student t_i produce una prueba t exacta para detectar un valor extremo en una ubicación específica, la distribución t no se aplicaría para probar simultáneamente varios de ellos en todas las ubicaciones. Como resultado, los residuos studentizados o valores R de Student deberían usarse estrictamente como herramientas de diagnóstico *sin* pruebas de hipótesis formales como mecanismo. La implicación es que dichos estadísticos resaltan datos de puntos en los cuales el error del ajuste es mayor de lo esperado por la sola aleatoriedad. Valores R de Student de magnitud grande sugieren la necesidad de "verificar" los datos con todos los recursos disponibles. La práctica de eliminar observaciones de conjuntos de datos de la regresión no debería llevarse a cabo en forma indiscriminada. (Para más información sobre el uso de los diagnósticos sobre valores extremos, véase la referencia a Myers, 1990, en la bibliografía.)

Ilustración de la detección de valores extremos

Ejemplo 12.12: En un experimento biológico efectuado en el Departamento de Entomología del Instituto Politécnico y Universidad Estatal de Virginia, se hicieron n corridas experimentales con dos métodos diferentes para capturar saltamontes. Los métodos son captura en red de caída y captura en red de barrido. Para cada método, se registró el número promedio de saltamontes atrapados en un conjunto de cuadrantes del campo en una fecha dada. También se registró una variable regresora adicional: la altura promedio de las plantas en los cuadrantes. Los datos experimentales aparecen en la tabla 12.10.

Tabla 12.10: Conjunto de datos para el ejemplo 12.12

Observación	Captura con red de caída, y	Captura con red de barrido, x_1	Altura de las plantas, x_2 (cm)
1	18.0000	4.15476	52.705
2	8.8750	2.02381	42.069
3	2.0000	0.15909	34.766
4	20.0000	2.32812	27.622
5	2.3750	0.25521	45.879
6	2.7500	0.57292	97.472
7	3.3333	0.70139	102.062
8	1.0000	0.13542	97.790
9	1.3333	0.12121	88.265
10	1.7500	0.10937	58.737
11	4.1250	0.56250	42.386
12	12.8750	2.45312	31.274
13	5.3750	0.45312	31.750
14	28.0000	6.68750	35.401
15	4.7500	0.86979	64.516
16	1.7500	0.14583	25.241
17	0.1333	0.01562	36.354

El objetivo es poder estimar la captura de saltamontes empleando únicamente el método de la red de barrido, que es menos costoso. Hay cierta preocupación acerca de la validez del cuarto dato de los puntos. La captura observada que se reportó con el uso del método de la red de caída parece inusualmente alta, dadas las demás condiciones y, asimismo, había la sensación de que la cifra podía ser errónea. Ajuste un modelo del tipo

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

a los 17 datos de los puntos y estudie los residuos para determinar si el dato del punto 4 es un valor extremo.

Solución: Un paquete de cómputo generó el modelo de regresión ajustado

$$\hat{y} = 3.6870 + 4.1050x_1 - 0.0367x_2$$

junto con los estadísticos $R^2 = 0.9244$ y $s^2 = 5.580$. También se obtuvieron los residuos y otra información de diagnóstico, y se registraron en la tabla 12.11.

Tabla 12.11: Información sobre los residuos para el conjunto de datos del ejemplo 12.12

Obs.	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	h_{ii}	$s\sqrt{1-h_{ii}}$	r_i	t_i
1	18.000	18.809	−0.809	0.2291	2.074	−0.390	−0.3780
2	8.875	10.452	−1.577	0.0766	2.270	−0.695	−0.6812
3	2.000	3.065	−1.065	0.1364	2.195	−0.485	−0.4715
4	20.000	12.231	7.769	0.1256	2.209	3.517	9.9315
5	2.375	3.052	−0.677	0.0931	2.250	−0.301	−0.2909
6	2.750	2.464	0.286	0.2276	2.076	0.138	0.1329
7	3.333	2.823	0.510	0.2669	2.023	0.252	0.2437
8	1.000	0.656	0.344	0.2318	2.071	0.166	0.1601
9	1.333	0.947	0.386	0.1691	2.153	0.179	0.1729
10	1.750	1.982	−0.232	0.0852	2.260	−0.103	−0.0989
11	4.125	4.442	−0.317	0.0884	2.255	−0.140	−0.1353
12	12.875	12.610	0.265	0.1152	2.222	0.119	0.1149
13	5.375	4.383	0.992	0.1339	2.199	0.451	0.4382
14	28.000	29.841	−1.841	0.6233	1.450	−1.270	−1.3005
15	4.750	4.891	−0.141	0.0699	2.278	−0.062	−0.0598
16	1.750	3.360	−1.610	0.1891	2.127	−0.757	−0.7447
17	0.133	2.418	−2.285	0.1386	2.193	−1.042	−1.0454

Como se esperaba, el residuo en la cuarta ubicación parece inusualmente grande, 7.769. La cuestión fundamental aquí es si este residuo es más grande o no que el que se esperaría debido al azar. El error estándar para el punto 4 es 2.209. El valor R de Student, t_4, resulta de 9.9315. Al ver éste como el valor de una variable aleatoria que tiene una distribución t con 13 grados de libertad, se concluiría sin duda que el residuo de la cuarta observación es algo mayor que 0, y que la medición del presunto error está apoyada por el estudio de los residuos. Observe que ningún otro valor de los residuos en un valor R de Student es motivo de alarma.

Gráfica de los residuos

En el capítulo 11 estudiamos con cierto detalle la utilidad de graficar los residuos en el análisis de regresión. Es frecuente que con base en dichas gráficas se detecte la trasgresión de las suposiciones del modelo. En la regresión múltiple, la probabilidad normal de la gráfica de los residuos o de los residuos contra $\hat{y}$ es de utilidad. Sin embargo, con frecuencia es preferible graficar los residuos studentizados.

Hay que recordar que la preferencia de los residuos studentizados sobre los residuos ordinarios para propósitos de graficación surge del hecho de que como la varianza del i-ésimo residuo depende del i-ésimo elemento en la diagonal de la matriz TESTADA, las varianzas de los residuos diferirán si hay dispersión en las diagonales TESTADAS. Así, la apariencia de una gráfica de residuos puede ser heterogénea debido a que éstos no se comportan, en general, en forma ideal. El propósito de utilizar residuos studentizados es proporcionar una *estandarización*. Es claro que si se conociera σ, entonces en condiciones ideales (es decir, un modelo correcto y una varianza homogénea), se tendría

$$E\left(\frac{e_i}{\sigma\sqrt{1-h_{ii}}}\right) = 0, \quad \text{y} \quad Var\left(\frac{e_i}{\sigma\sqrt{1-h_{ii}}}\right) = 1.$$

Por lo que los residuos studentizados producen un conjunto de estadísticos que se comportan en forma estándar bajo condiciones ideales. La figura 12.5 muestra una gráfica de valores ***R* de Student** para los datos de los saltamontes del ejemplo 12.12. Observe que el valor para la observación 4 se aparta de los demás. La gráfica R de Student se generó con el software SAS. La gráfica presenta los residuos contra los valores $\hat{y}$.

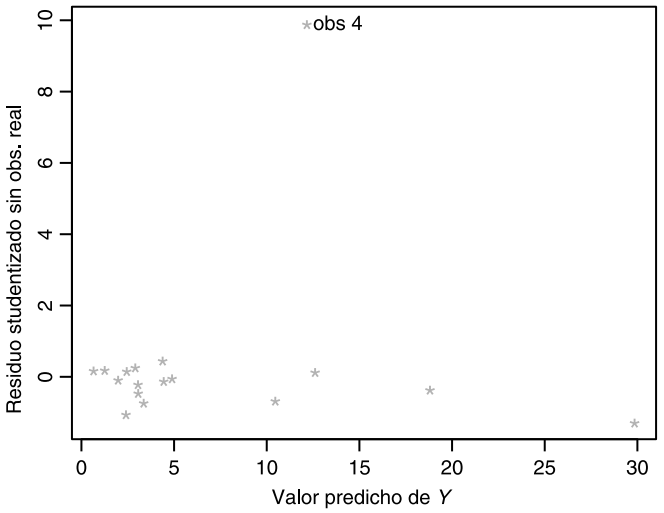


Figura 12.5: Valores R de Student graficados contra los valores predichos para los datos de los saltamontes del ejemplo 12.12.

Verificación de la normalidad

El lector debe recordar la importancia que tiene verificar la normalidad utilizando la gráfica de la probabilidad normal, según se estudió en el capítulo 11. La misma recomendación es válida para el caso de la regresión lineal múltiple. Las gráficas de probabilidad normal pueden generarse con el empleo de software estándar para regresión. Sin embargo, otra vez, pueden ser más eficaces si no se usan residuos ordinarios, sino studentizados o valores R de Student.

12.11 Validación cruzada, x_1, y otros criterios para la selección del modelo

Para muchos problemas de regresión, el experimentador debe elegir entre distintos modelos alternativos o formas de modelo que se desarrollen a partir del mismo conjunto de datos. En efecto, con mucha frecuencia se requiere el modelo que predice o estima mejor la respuesta media. El experimentador debe tomar en cuenta los tamaños relativos de los valores de s^2 para los modelos candidatos y, sin duda, la naturaleza general de los intervalos de confianza sobre la respuesta media. También debe considerarse lo bien que prediga el modelo valores de la respuesta que **no se hayan utilizado para construir los modelos candidato**. Los modelos deben estar sujetos a **validación cruzada**. Entonces, lo que se requiere son los errores de la validación cruzada en vez de los errores del ajuste. Tales errores en la predicción son los **residuos PRESS**.

$$\delta_i = y_i - \hat{y}_{i,-i}, \quad i = 1, 2, \ldots, n,$$

donde $\hat{y}_i - i$, es la predicción del i-ésimo dato de punto por medio de un modelo que no utiliza el i-ésimo punto en el cálculo de los coeficientes. Estos residuos PRESS se calculan a partir de la fórmula

$$\delta_i = \frac{e_i}{1 - h_{ii}}, \quad i = 1, 2, \ldots, n,$$

(La obtención de ésta se encuentra en el libro de texto sobre regresión de Myers, 1990.)

Uso del estadístico PRESS

La motivación para utilizar PRESS y la herramienta de los residuos PRESS es muy fácil de entender. El propósito de extraer o *separar* datos de puntos uno a la vez es permitir el empleo de metodologías separadas para ajustar y evaluar un modelo específico. Para evaluar un modelo, la "$-i$" indica que el residuo PRESS comete un error de predicción donde la observación que se predice es *independiente del ajuste del modelo.*

Los criterios que usan los residuos PRESS están dados por

$$\sum_{i=1}^{n} |\delta_i| \quad \text{y} \quad \text{PRESS} = \sum_{i=1}^{n} \delta_i^2.$$

El término PRESS es un acrónimo de **suma de cuadrados de la predicción** (prediction sum of squares). Se sugiere emplear ambos criterios. Es posible que PRESS

sea dominado por uno o algunos residuos PRESS grandes. Es claro que el criterio $\sum_{i=1}^{n} |\delta_i|$ es menos sensible a un número pequeño de valores grandes.

Además del estadístico PRESS en sí, el analista puede tan sólo calcular otro "semejante al R" que refleje la bondad de la predicción. Es frecuente que dicho estadístico se denomine R^2_{pred} y está dado como sigue:

R^2 de la predicción

Dado un modelo ajustado con valor específico para PRESS, R^2_{pred} está dado por

$$R^2_{\text{pred}} = 1 - \frac{\text{PRESS}}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Observe que R^2_{pred} es tan sólo el estadístico ordinario R^2 con la *SSE* reemplazada por el estadístico PRESS.

En el ejemplo siguiente se ilustra un "estudio de caso" donde se ajustan muchos modelos candidato a un conjunto de datos y se elige el mejor de ellos. No se emplean los procedimientos secuenciales descritos en la sección 12.9. En cambio, se ilustra el papel de los residuos PRESS y otros valores de estadísticos para seleccionar la mejor ecuación de regresión.

Ejemplo 12.13: **Estudio de caso** La fortaleza en las piernas es un requisito necesario para un pateador exitoso en el fútbol americano. Una medida de la calidad de un buen pateador es el "tiempo de vuelo", que es el tiempo que el balón se mantiene en el aire antes de ser atrapado por el regresador de patadas. Para determinar cuáles factores de la fortaleza en las piernas influyen en el tiempo de vuelo y desarrollar un modelo empírico para predecir esta respuesta, el Departamento de Salud, Educación Física y Recreación, del Instituto Politécnico y Universidad Estatal de Virginia, llevó a cabo un estudio sobre *La relación entre variables seleccionadas de desempeño físico y la capacidad de despejes en el fútbol*. Se eligieron 13 pateadores para el experimento, y cada uno pateó 10 veces el balón. El tiempo de vuelo promedio, junto con las mediciones de fortaleza usada en el análisis, están registrados en la tabla 12.12.

Cada variable regresora se define como sigue:

1. **RLS**, fortaleza en la pierna derecha (libras).
2. **LLS**, fortaleza en la pierna izquierda (libras).
3. **RHF**, flexibilidad del músculo del tendón derecho (grados).
4. **LHF**, flexibilidad del músculo del tendón izquierdo (grados).
5. **Potencia**, fortaleza conjunta de las piernas (pies-libra).

Determine el modelo más adecuado para predecir el tiempo de vuelo.

Solución: En la búsqueda del "mejor" de los modelos candidato para predecir el tiempo de vuelo, se obtuvo la información de la tabla 12.13 a partir de un paquete de cómputo para regresión. Los modelos están clasificados en orden ascendente con respecto a los valores del estadístico PRESS. Esta presentación brinda información suficiente acerca de todos los modelos posibles, con la finalidad de permitir que el usuario elimine algunos de éstos. El modelo que contiene a x_2 y x_5 (*LLS* y *Potencia*), denotado con x_2x_5, parece superior para predecir el tiempo del vuelo para los pateadores. Asimismo,

Tabla 12.12: Datos para el ejemplo 12.13

Pateador	Tiempo de, vuelo y(seg)	RLS, x_1	LLS, x_2	RHF, x_3	LHF, x_4	Potencia, x_5
1	4.75	170	170	106	106	240.57
2	4.07	140	130	92	93	195.49
3	4.04	180	170	93	78	152.99
4	4.18	160	160	103	93	197.09
5	4.35	170	150	104	93	266.56
6	4.16	150	150	101	87	260.56
7	4.43	170	180	108	106	219.25
8	3.20	110	110	86	92	132.68
9	3.02	120	110	90	86	130.24
10	3.64	130	120	85	80	205.88
11	3.68	120	140	89	83	153.92
12	3.60	140	130	92	94	154.64
13	3.85	160	150	95	95	240.57

observe que todos los modelos con PRESS bajo, s^2 baja, bajo $\sum_{i=1}^{n} |\delta_i|$, y valores altos de R^2, contienen esas dos variables.

Con el propósito de obtener alguna perspectiva de los residuos de la regresión ajustada

$$\hat{y}_i = b_0 + b_2 x_{2i} + b_5 x_{5i},$$

se generaron los residuos y los residuos PRESS. El modelo de predicción real (véase el ejercicio 12.47 en la página 496) está dado por

$$\hat{y} = 1.10765 + 0.01370 x_2 + 0.00429 x_5.$$

En la tabla 12.14 se listan los residuos, los valores de la diagonal testada y los valores PRESS.

Observe el ajuste a los datos relativamente bueno de los modelos de regresión con dos variables. Los residuos PRESS reflejan la capacidad de la ecuación de regresión para predecir el tiempo de vuelo si se hicieran predicciones independientes. Por ejemplo, para el pateador número 4, el tiempo de vuelo de 4.180 tendría un error de predicción de 0.039 si se construyera el modelo usando los 12 lanzadores restantes. Para este modelo, el error promedio de la predicción, o error de validación cruzada, es

$$\frac{1}{13} \sum_{i=1}^{n} |\delta_i| = 0.1489 \text{ segundos},$$

que es pequeño comparado con el tiempo de vuelo promedio para los 13 lanzadores.

En la sección 12.9 se indica que cuando se busca el modelo mejor, con frecuencia es aconsejable utilizar todos los subconjuntos posibles de regresión. Los paquetes de software para estadística más conocidos contiene una rutina de *todas las regresiones posibles*. Tales algoritmos calculan diversos criterios para todos los subconjuntos de términos del modelo. Es evidente que criterios como R^2, s^2 y PRESS son razonables para elegir entre subconjuntos de candidatos. Otro estadístico muy popular y útil, en particular para áreas de las ciencias físicas e ingeniería, es el estadístico C_p, que se describe a continuación.

Tabla 12.13: Comparación de diferentes modelos de regresión

Modelo	s^2	$\sum \lvert\delta_i\rvert$	PRESS	R^2
x_2x_5	0.036907	1.93583	0.54683	0.871300
$x_1x_2x_5$	0.041001	2.06489	0.58998	0.871321
$x_2x_4x_5$	0.037708	2.18797	0.59915	0.881658
$x_2x_3x_5$	0.039636	2.09553	0.66182	0.875606
$x_1x_2x_4x_5$	0.042265	2.42194	0.67840	0.882093
$x_1x_2x_3x_5$	0.044578	2.26283	0.70958	0.875642
$x_2x_3x_4x_5$	0.042421	2.55789	0.86236	0.881658
$x_1x_3x_5$	0.053664	2.65276	0.87325	0.831580
$x_1x_4x_5$	0.056279	2.75390	0.89551	0.823375
x_1x_5	0.059621	2.99434	0.97483	0.792094
x_2x_3	0.056153	2.95310	0.98815	0.804187
x_1x_3	0.059400	3.01436	0.99697	0.792864
$x_1x_2x_3x_4x_5$	0.048302	2.87302	1.00920	0.882096
x_2	0.066894	3.22319	1.04564	0.743404
x_3x_5	0.065678	3.09474	1.05708	0.770971
x_1x_2	0.068402	3.09047	1.09726	0.761474
x_3	0.074518	3.06754	1.13555	0.714161
$x_1x_3x_4$	0.065414	3.36304	1.15043	0.794705
$x_2x_3x_4$	0.062082	3.32392	1.17491	0.805163
x_2x_4	0.063744	3.59101	1.18531	0.777716
$x_1x_2x_3$	0.059670	3.41287	1.26558	0.812730
x_3x_4	0.080605	3.28004	1.28314	0.718921
x_1x_4	0.069965	3.64415	1.30194	0.756023
x_1	0.080208	3.31562	1.30275	0.692334
$x_1x_3x_4x_5$	0.059169	3.37362	1.36867	0.834936
$x_1x_2x_4$	0.064143	3.89402	1.39834	0.798692
$x_3x_4x_5$	0.072505	3.49695	1.42036	0.772450
$x_1x_2x_3x_4$	0.066088	3.95854	1.52344	0.815633
x_5	0.111779	4.17839	1.72511	0.571234
x_4x_5	0.105648	4.12729	1.87734	0.631593
x_4	0.186708	4.88870	2.82207	0.283819

El estadístico C_p

Es muy frecuente que la selección del modelo más adecuado implique muchas consideraciones. Evidentemente, es importante el número de términos del modelo; el tema de la parsimonia es una consideración que no debe ignorarse. Por otro lado, el analista no quedaría satisfecho con un modelo que sea demasiado simple, hasta el punto en que hubiera una simplificación excesiva. En este sentido, un estadístico único que representa un compromiso aceptable es C_p. (Véase la referencia a Mallows, en la bibliografía.)

El estadístico C_p parece agradable al sentido común y se desarrolla a partir de consideraciones del compromiso apropiado entre el sesgo excesivo, en que se incurre cuando se subajusta (se eligen muy pocos términos para el modelo), y la varianza excesiva de la predicción que se genera cuando se sobreajusta (hay redundancias en

Tabla 12.14: Residuos PRESS

Pateador	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	h_{ii}	δ_i
1	4.750	4.470	0.280	0.198	0.349
2	4.070	3.728	0.342	0.118	0.388
3	4.040	4.094	−0.054	0.444	−0.097
4	4.180	4.146	0.034	0.132	0.039
5	4.350	4.307	0.043	0.286	0.060
6	4.160	4.281	−0.121	0.250	−0.161
7	4.430	4.515	−0.085	0.298	−0.121
8	3.200	3.184	0.016	0.294	0.023
9	3.020	3.174	−0.154	0.301	−0.220
10	3.640	3.636	0.004	0.231	0.005
11	3.680	3.687	−0.007	0.152	−0.008
12	3.600	3.553	0.047	0.142	0.055
13	3.850	4.196	−0.346	0.154	−0.409

el modelo). El estadístico C_p es una función sencilla del número total de parámetros en el modelo candidato y el error cuadrático medio s^2.

Aquí no se presentará el desarrollo completo del estadístico C_p. (Para mayores detalles, se recomienda que el lector consulte el libro de texto de Myers que se cita en la bibliografía.) El C_p para un subconjunto particular de modelos es *una estimación de lo siguiente*:

$$\Gamma_{(p)} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Var}(\hat{y}_i) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(\text{Sesgo } \hat{y}_i)^2.$$

Se descubre que, con las suposiciones estándar de los mínimos cuadrados que se indicaron con anterioridad en este capítulo, y si se supone que el modelo "verdadero" es aquel que contiene todas las variables candidatas,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Var}(\hat{y}_i) = p \qquad \text{(número de parámetros en el modelo candidato)}$$

(véase el ejercicio de repaso 12.61) y un estimador insesgado de

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\text{Sesgo } \hat{y}_i)^2 \text{ está dado por } \frac{1}{\sigma^2}\sum_{i=1}^{n}(\widehat{\text{Sesgo}}\ \hat{y}_i)^2 = \frac{(s^2-\sigma^2)(n-p)}{\sigma^2}.$$

En éstas, s^2 es el error cuadrático medio para el modelo candidato, y σ^2 es la varianza del error de la población. Así, si se supone que se dispone de algún estimador para $\hat{\sigma}^2$, entonces C_p está dado por

Estadístico C_p

$$C_p = p + \frac{(s^2-\hat{\sigma}^2)(n-p)}{\hat{\sigma}^2},$$

donde p es el número de parámetros en el modelo, s^2 es el error cuadrático medio para el modelo candidato, y $\hat{\sigma}^2$ es un estimador de σ^2.

Tabla 12.15: Datos para el ejemplo 12.14

Distrito	Cuentas promocionales, x_1	Cuentas activas, x_2	Marcas en competencia, x_3	Potencial, x_4	Ventas, y (miles)
1	5.5	31	10	8	$ 79.3
2	2.5	55	8	6	200.1
3	8.0	67	12	9	163.2
4	3.0	50	7	16	200.1
5	3.0	38	8	15	146.0
6	2.9	71	12	17	177.7
7	8.0	30	12	8	30.9
8	9.0	56	5	10	291.9
9	4.0	42	8	4	160.0
10	6.5	73	5	16	339.4
11	5.5	60	11	7	159.6
12	5.0	44	12	12	86.3
13	6.0	50	6	6	237.5
14	5.0	39	10	4	107.2
15	3.5	55	10	4	155.0

Es evidente que el científico debería adoptar modelos con valores pequeños de C_p. El lector tiene que observar que, a diferencia del estadístico PRESS, C_p está libre de escala. Además, se puede obtener alguna perspectiva respecto de lo adecuado de un modelo candidato observando el valor de su C_p. Por ejemplo, $C_p > p$ indica un modelo sesgado debido a que está subajustado; mientras que $C_p \approx p$ indica un modelo razonable.

Con frecuencia hay confusión acerca de donde proviene $\hat{\sigma}^2$ en la fórmula para C_p. Es notorio que el científico o ingeniero no tienen acceso a la cantidad σ^2 de la población. En aplicaciones donde se dispone de corridas repetidas, digamos en situaciones de diseño experimental, se dispone de un estimador de σ^2 independiente del modelo (véanse los capítulos 11 y 15). Sin embargo, la mayoría de los paquetes de software utilizan $\hat{\sigma}^2$ como el *error cuadrático medio del modelo más completo*. Evidentemente si éste no es un estimador bueno, la porción de sesgo del estadístico C_p puede ser negativa. Así, C_p llega a ser menor que p.

Ejemplo 12.14: Considere el conjunto de datos de la tabla 12.15, en los cuales un fabricante de grava asfáltica se interesa en la relación entre las ventas durante un año específico y los factores que influyen en ellas. (Los datos están tomados de Neter, Wassermann y Kutner; véase la bibliografía.)

De los modelos de subconjuntos posibles, hay tres que revisten interés especial. Estos tres son los de x_2x_3, $x_1x_2x_3$ y $x_1x_2x_3x_4$. A continuación se presenta la información pertinente para comparar los tres modelos. Con el objetivo de ayudar a la toma de decisiones se incluyen los estadísticos PRESS de los tres modelos.

Modelo	R^2	R^2_{pred}	s^2	PRESS	C_p
x_2x_3	0.9940	0.9913	44.5552	782.1896	11.4013
$x_1x_2x_3$	0.9970	0.9928	24.7956	643.3578	3.4075
$x_1x_2x_3x_4$	0.9971	0.9917	26.2073	741.7557	5.0

Dependent Variable: sales

Number in Model	C(p)	R-Square	Adjusted R-Square	MSE	Variables in Model
3	3.4075	0.9970	0.9961	24.79560	x1 x2 x3
4	5.0000	0.9971	0.9959	26.20728	x1 x2 x3 x4
2	11.4013	0.9940	0.9930	44.55518	x2 x3
3	13.3770	0.9940	0.9924	48.54787	x2 x3 x4
3	1053.643	0.6896	0.6049	2526.96144	x1 x3 x4
2	1082.670	0.6805	0.6273	2384.14286	x3 x4
2	1215.316	0.6417	0.5820	2673.83349	x1 x3
1	1228.460	0.6373	0.6094	2498.68333	x3
3	1653.770	0.5140	0.3814	3956.75275	x1 x2 x4
2	1668.699	0.5090	0.4272	3663.99357	x1 x2
2	1685.024	0.5042	0.4216	3699.64814	x2 x4
1	1693.971	0.5010	0.4626	3437.12846	x2
2	3014.641	0.1151	-.0324	6603.45109	x1 x4
1	3088.650	0.0928	0.0231	6248.72283	x4
1	3364.884	0.0120	-.0640	6805.59568	x1

Figura 12.6: Salida del *SAS* de todos los subconjuntos posibles sobre los datos de las ventas para el ejemplo 12.14.

De la información de la tabla, parece claro que el modelo $x_1x_2x_3$ es preferible sobre los otros dos. Observe que para el modelo completo, $C_p = 5.0$. Esto ocurre porque la *porción de sesgo* es igual a cero, y $\hat{\sigma}^2 = 26.2073$ es el error cuadrático medio del modelo completo.

La figura 12.6 es una salida anotada del *SAS* PROC REG que muestra información sobre todas las regresiones posibles. De ahí es posible hacer comparaciones de otros modelos con (x_1, x_2, x_3). Observe que (x_1, x_2, x_3) parece muy bueno en comparación con todos los modelos.

Como verificación final del modelo (x_1, x_2, x_3), la figura 12.7 presenta una gráfica de probabilidad normal de los residuos del modelo.

Ejercicios

12.47 Considere el "tiempo de vuelo" para los datos del despeje que se dan en el ejemplo 12.13, utilizando sólo las variables x_2 y x_3.

a) Verifique la ecuación de regresión que se presenta en la página 492.

b) Prediga el tiempo de vuelo para el pateador con LLS = 180 libras y potencia = 260 pies-libras.

c) Construya un intervalo de confianza de 95% para el tiempo de vuelo medio de un pateador con LLS = 180 libras y potencia = 260 pies-libras.

12.48 Para los datos del ejercicio 12.11 de la página 454, utilice las técnicas de

a) *selección hacia delante* con un nivel de significancia de 0.05 para elegir un modelo de regresión lineal;

b) *eliminación hacia atrás* con un nivel de significancia de 0.05 para seleccionar un modelo de regresión lineal;

c) *regresión progresiva* con un nivel de significancia de 0.05, para escoger un modelo de regresión lineal.

12.49 Emplee las técnicas de *eliminación hacia atrás* con $\alpha = 0.05$ para elegir una ecuación de predicción para los datos de la tabla 12.8.

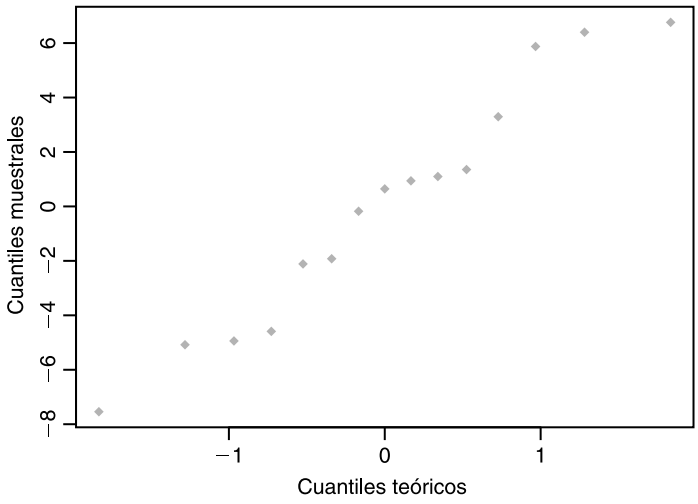


Figura 12.7: Gráfica de la probabilidad normal de los residuos, utilizando el modelo $x_1x_2x_3$ para el ejemplo 12.14.

12.50 Para los datos del pateador del ejemplo 12.13, también se registró una respuesta adicional, la "distancia de pateo". Los siguientes son los valores de distancia promedio para cada uno de los 13 pateadores:

Pateador	**Distancia, y (pies)**
1	162.50
2	144.00
3	147.50
4	163.50
5	192.00
6	171.75
7	162.00
8	104.93
9	105.67
10	117.59
11	140.25
12	150.17
13	165.16

a) Con los datos de distancia en vez de los de tiempo de vuelo, estime un modelo de regresión lineal múltiple del tipo

$$\mu_{Y|x_1,x_2,x_3,x_4,x_5} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5$$

para predecir la distancia del despeje.

b) Utilice regresión progresiva con un nivel de significancia de 0.10 para seleccionar una combinación de variables.

c) Genere valores para s^2, R^2, PRESS y $\sum_{i=1}^{13} |\delta_i|$ para todo el conjunto de 31 modelos. Utilice esta información para determinar la mejor combinación de variables para predecir la distancia del despeje.

d) Para el modelo final que seleccione, grafique los residuos estandarizados contra Y y elabore una gráfica de probabilidad normal de los residuos ordinarios. Haga comentarios.

12.51 El siguiente es un conjunto de datos para y, la cantidad de dinero (miles de dólares) aportados a la asociación de alumnos del Virginia Tech, por la Clase de 1960; y para x, el número de años posteriores a la graduación:

y	x	y	x
812.52	1	2755.00	11
822.50	2	4390.50	12
1211.50	3	5581.50	13
1348.00	4	5548.00	14
1301.00	8	6086.00	15
2567.50	9	5764.00	16
2526.50	10	8903.00	17

a) Ajuste un modelo de regresión del tipo

$$\mu_{Y|x} = \beta_0 + \beta_1x.$$

b) Ajuste un modelo cuadrático del tipo

$$\mu_{Y|x} = \beta_0 + \beta_1x + \beta_{11}x^2.$$

c) Determine cuál de los modelos de los incisos *a*) o *b*) es preferible. Utilice s^2, R^2 y los residuos PRESS para apoyar su decisión.

12.52 Para el modelo del ejercicio 12.50*a*), pruebe la hipótesis

$$H_0\text{: } \beta_4 = 0$$
$$H_1\text{: } \beta_4 \neq 0$$

Utilice un valor P para su conclusión.

12.53 Para el modelo cuadrático del ejercicio 12.51*b*), proporcione estimadores de las varianzas y las covarianzas de los estimadores de β_1 y β_{11}.

12.54 En un esfuerzo para modelar las remuneraciones de los ejecutivos en el año de 1979, se seleccionaron 33 empresas y se recabaron datos acerca de remuneraciones, ventas, ganancias y empleo. Considere el modelo

$$y_i = \beta_0 + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \beta_3 \ln x_{3i} + \epsilon_i, \qquad i = 1, 2, \ldots, 33.$$

a) Ajuste la regresión con el modelo anterior.

b) ¿Es un modelo con un subconjunto de las variables preferible al modelo completo?

Empresa	Compensación, y (miles)	Ventas, x_1, (millones)	Ganancias, x_2, (millones)	Empleo, x_3
1	\$450	\$4,600.6	\$128.1	48,000
2	387	9,255.4	783.9	55,900
3	368	1,526.2	136.0	13,783
4	277	1,683.2	179.0	27,765
5	676	2,752.8	231.5	34,000
6	454	2,205.8	329.5	26,500
7	507	2,384.6	381.8	30,800
8	496	2,746.0	237.9	41,000
9	487	1,434.0	222.3	25,900
10	383	470.6	63.7	8,600
11	311	1,508.0	149.5	21,075
12	271	464.4	30.0	6,874
13	524	9,329.3	577.3	39,000
14	498	2,377.5	250.7	34,300
15	343	1,174.3	82.6	19,405
16	354	409.3	61.5	3,586
17	324	724.7	90.8	3,905
18	225	578.9	63.3	4,139
19	254	966.8	42.8	6,255
20	208	591.0	48.5	10,605
21	518	4,933.1	310.6	65,392
22	406	7,613.2	491.6	89,400
23	332	3,457.4	228.0	55,200
24	340	545.3	54.6	7,800
25	698	22,862.8	3011.3	337,119
26	306	2,361.0	203.0	52,000
27	613	2,614.1	201.0	50,500
28	302	1,013.2	121.3	18,625
29	540	4,560.3	194.6	97,937
30	293	855.7	63.4	12,300
31	528	4,211.6	352.1	71,800
32	456	5,440.4	655.2	87,700
33	417	1,229.9	97.5	14,600

12.55 La blancura del rayón es un factor importante para los científicos que estudian la calidad de las telas. La blancura se ve afectada por la calidad de la pulpa y otras variables de procesamiento. Algunas de estas incluyen la temperatura del baño con ácido, °C (x_1); concentración del ácido en cascada, % (x_2); temperatura del agua, °C (x_3); concentración del sulfuro, % (x_4); cantidad del blanqueador de cloro, lb/min (x_5); la temperatura de terminación de la tela, °C (x_6). A continuación se da un conjunto de datos de especímenes de rayón. La respuesta, y, es la medición de la blancura.

a) Utilice los criterios MSE, c_p y PRESS para dar el mejor modelo del subconjunto de todos los modelos.

b) Grafique los residuos estandarizados contra Y y haga una gráfica de probabilidad normal de los residuos para el "mejor" modelo. Comente los resultados.

y	x_1	x_2	x_3	x_4	x_5	x_6
88.7	43	0.211	85	0.243	0.606	48
89.3	42	0.604	89	0.237	0.600	55
75.5	47	0.450	87	0.198	0.527	61
92.1	46	0.641	90	0.194	0.500	65
83.4	52	0.370	93	0.198	0.485	54
44.8	50	0.526	85	0.221	0.533	60
50.9	43	0.486	83	0.203	0.510	57
78.0	49	0.504	93	0.279	0.489	49
86.8	51	0.609	90	0.220	0.462	64
47.3	51	0.702	86	0.198	0.478	63
53.7	48	0.397	92	0.231	0.411	61
92.0	46	0.488	88	0.211	0.387	88
87.9	43	0.525	85	0.199	0.437	63
90.3	45	0.486	84	0.189	0.499	58
94.2	53	0.527	87	0.245	0.530	65
89.5	47	0.601	95	0.208	0.500	67

12.56 Un cliente del Departamento de Ingeniería Mecánica se acercó al Centro de Consulta del Instituto Politécnico y Universidad Estatal de Virginia, para que lo ayudaran a analizar un experimento sobre motores de turbina de gas. Se midieron varias salidas del voltaje de los motores con distintas combinaciones de velocidad de las aspas y del voltaje que mide la extensión de los sensores. Los datos son los siguientes:

y (voltios)	Velocidad, x_1 (pulg/seg)	Extensión, x_2 (pulg)
1.95	6336	0.000
2.50	7099	0.000
2.93	8026	0.000
1.69	6230	0.000
1.23	5369	0.000
3.13	8343	0.000
1.55	6522	0.006
1.94	7310	0.006
2.18	7974	0.006
2.70	8501	0.006
1.32	6646	0.012
1.60	7384	0.012
1.89	8000	0.012

y (voltios)	Velocidad, x_1 (pulg/seg)	Extensión, x_2 (pulg)
2.15	8545	0.012
1.09	6755	0.018
1.26	7362	0.018
1.57	7934	0.018
1.92	8554	0.018

a) Haga un ajuste de regresión lineal múltiple a los datos.

b) Calcule las pruebas t sobre los coeficientes. Proporcione valores F.

c) Diga sus comentarios sobre la calidad del modelo ajustado.

12.57 La resistencia a la tracción de una unión de alambre es una característica importante. La siguiente tabla brinda información sobre la resistencia a la tracción y, la altura del molde x_1, la altura del perno x_2, altura del lazo x_3, longitud del alambre x_4, ancho de la unión sobre el molde x_5 y ancho del molde sobre el perno x_6. [Datos tomados de Myers y Montgomery (2002).]

y	x_1	x_2	x_3	x_4	x_5	x_6
8.0	5.2	19.6	29.6	94.9	2.1	2.3
8.3	5.2	19.8	32.4	89.7	2.1	1.8
8.5	5.8	19.6	31.0	96.2	2.0	2.0
8.8	6.4	19.4	32.4	95.6	2.2	2.1
9.0	5.8	18.6	28.6	86.5	2.0	1.8
9.3	5.2	18.8	30.6	84.5	2.1	2.1
9.3	5.6	20.4	32.4	88.8	2.2	1.9
9.5	6.0	19.0	32.6	85.7	2.1	1.9
9.8	5.2	20.8	32.2	93.6	2.3	2.1
10.0	5.8	19.9	31.8	86.0	2.1	1.8
10.3	6.4	18.0	32.6	87.1	2.0	1.6
10.5	6.0	20.6	33.4	93.1	2.1	2.1
10.8	6.2	20.2	31.8	83.4	2.2	2.1
11.0	6.2	20.2	32.4	94.5	2.1	1.9
11.3	6.2	19.2	31.4	83.4	1.9	1.8
11.5	5.6	17.0	33.2	85.2	2.1	2.1
11.8	6.0	19.8	35.4	84.1	2.0	1.8
12.3	5.8	18.8	34.0	86.9	2.1	1.8
12.5	5.6	18.6	34.2	83.0	1.9	2.0

a) Ajuste un modelo de regresión usando todas las variables independientes.

b) Use regresión progresiva con un nivel de significancia de entrada de 0.25 y un nivel de significancia de 0.05 para la remoción. Proporcione el modelo final.

c) Utilice todos los modelos de regresión posibles y calcule R^2, C_p, s^2 y R^2 ajustada, para todos los modelos.

d) Dé el modelo final.

e) Para el modelo del inciso *d*), grafique los residuos studentizados (o la R de Student) y haga comentarios al respecto.

12.58 Para el ejercicio 12.57, pruebe H_0: $\beta_1 = \beta_6 = 0$. Proporcione valores P y haga sus comentarios.

12.59 En el ejercicio 12.28 de la página 464, se tienen los datos siguientes sobre el uso de un rodamiento:

y (uso)	x_1 (viscosidad del aceite)	x_2 (carga)
193	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

a) Puede considerar el siguiente modelo para describir los datos:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

para $i = 1, 2, \ldots, 6$. El término x_1x_2 es de "interacción". Ajuste este modelo y estime los parámetros.

b) Utilice los modelos (x_1), (x_1, x_2), (x_2), (x_1, x_2, x_1x_2) y calcule PRESS, C_p, y s^2 para determinar el "mejor" modelo.

12.12 Modelos especiales no lineales para condiciones no ideales

En gran parte del material anterior de este capítulo y del 11 hemos tenido muchos beneficios por la suposición de que los errores del modelo, los ϵ_i, tienen distribución normal con media igual a cero y varianza constante σ^2. Sin embargo, hay muchas situaciones de la vida real en las cuales es evidente que la respuesta no es normal. Por ejemplo, existen aplicaciones donde la **respuesta es binaria** (0 o 1) y, por ello, su naturaleza es de Bernoulli. En las ciencias sociales el problema sería desarrollar un modelo que prediga si un individuo representa riesgos para un crédito o no (0 o 1), como función de ciertos regresores socioeconómicos como ingreso, edad, género y nivel académico. En una prueba biomédica para un fármaco, es frecuente que el paciente responda favorablemente o no a cierto medicamento, en tanto que los regresores incluyen la dosis y factores biológicos como edad, peso y presión sanguínea. Otra

vez, la respuesta es de naturaleza binaria. Las aplicaciones también son abundantes en las áreas de manufactura en que ciertos factores controlables influyen para decidir si cierto artículo fabricado está **defectuoso o no**.

Un segundo tipo de aplicación que no es normal y del que haremos una mención breve tiene que ver con el **control de datos**. Aquí, con frecuencia es conveniente suponer una respuesta de Poisson. En aplicaciones biomédicas, el número de colonias de células cancerosas es la respuesta que se modela contra las dosis de medicamentos. En la industria textil, el número de imperfecciones por yarda de tela es una respuesta razonable que se modela contra ciertas variables de los procesos.

Varianza no homogénea

El lector debería notar la comparación de la situación ideal (es decir, la respuesta normal) con aquella de la respuesta de Bernoulli (o binomial) o la de Poisson. Estamos acostumbrados al hecho de que el caso normal es muy especial en que la varianza es **independiente de la media**. Resulta claro que éste no es el caso para las respuestas de Bernoulli ni de Poisson. Por ejemplo, si la respuesta es 0 o 1, lo cual sugiere una respuesta de Bernoulli, entonces el modelo es de la forma

$$p = f(\boldsymbol{x}, \beta),$$

donde p es la **probabilidad de éxito** (por ejemplo, la respuesta = 1). El parámetro p juega el papel de $\mu_{Y|x}$ en el caso normal. Sin embargo, la varianza de Bernoulli es $p(1-p)$ que, desde luego, también es función del regresor $\boldsymbol{x}$. Como resultado, la varianza no es constante. Estas reglas utilizan los mínimos cuadrados estándar que se han utilizado en nuestro trabajo de regresión lineal hasta este momento. Lo mismo se aplica para el caso de Poisson, ya que el modelo es de la forma

$$\lambda = f(\boldsymbol{x}, \beta),$$

con $\text{Var}(y) = \mu_y = \lambda$, que varía con $\boldsymbol{x}$.

Respuesta binaria (regresión logística)

El enfoque más popular para modelar respuestas binarias es la técnica llamada **regresión logística**. Se emplea mucho en las ciencias biológicas, en la investigación biomédica y en la ingeniería. Pero incluso en las ciencias sociales se encuentra que las respuestas binarias son de abundantes. La distribución básica para la respuesta es la de Bernoulli o la binomial. La primera se encuentra en estudios observacionales donde no hay corridas repetidas en cada nivel de regresor; mientras que la segunda será el caso cuando se utilice un diseño experimental. Por ejemplo, en un ensayo clínico en el cual se evalúe un fármaco nuevo, el objetivo sería determinar la dosis del medicamento que es eficaz. Así, en el experimento se utilizarán ciertas dosis y para cada una de ellas se emplearán a varios sujetos. Este caso se denomina **caso agrupado**.

¿Cuál es el modelo para la regresión logística?

En el caso de respuestas binarias, la respuesta media es una probabilidad. En la ilustración clínica anterior, puede decirse que se desea estimar la probabilidad de que el

paciente responda en forma positiva al fármaco (P(éxito)). Entonces, el modelo se escribe en términos de una probabilidad. Dados los regresores $\boldsymbol{x}$, la función logística está dada por

$$p = \frac{1}{1 + e^{-\mathbf{x}'\beta}}.$$

La porción $\boldsymbol{x}'\beta$ se llama **predictor lineal** y, en el caso de un solo regresor x, puede escribirse $\boldsymbol{x}'\beta = \beta_0 + \beta_1 x$. Por supuesto, en el llamado predictor lineal no se trabaja con regresores múltiples y términos polinomiales. En el caso agrupado, el modelo implica el modelado de la media de una binomial en vez de una de Bernoulli, por lo que se tiene la media dada por

$$np = \frac{n}{1 + e^{-\mathbf{x}'\beta}}.$$

Características de la función logística

Una gráfica de la función logística revela mucho sobre sus características y del porqué se utiliza para este tipo de problema. En primer lugar, la función es no lineal. Además, la gráfica de la figura 12.8 revela la forma de S con la función que tiende a la asíntota en $p = 1.0$. En este caso, $\beta_1 > 0$. Así, nunca se experimentaría una probabilidad estimada mayor que 1.0.

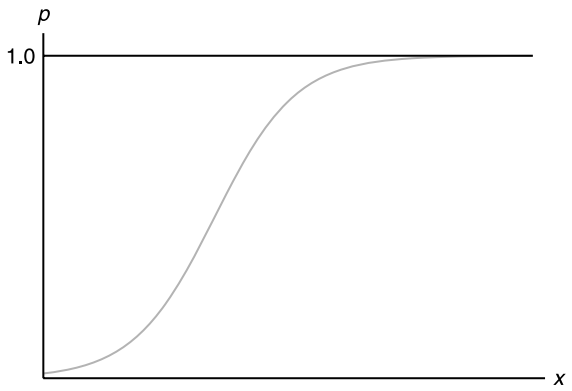


Figura 12.8: La función logística.

Los coeficientes de regresión en el predictor lineal se estiman con el método de máxima verosimilitud según se describe en el capítulo 9. La solución de las ecuaciones de verosimilitud requiere una metodología iterativa que no veremos aquí. Sin embargo, presentaremos un ejemplo y analizaremos la salida por computadora y las conclusiones.

Ejemplo 12.15: El conjunto de datos de la tabla 12.16 es un ejemplo del uso de la regresión logística para analizar un ensayo biológico cuantal de agente único en un experimento de toxicidad. Los resultados muestran el efecto de dosis diferentes de la nicotina sobre la mosca común de la fruta.

El propósito del experimento fue utilizar la regresión logística para llegar a un modelo adecuado que relacionara la probabilidad de "muerte" con la concentración.

Tabla 12.16. Conjunto de datos para el ejemplo 12.15

x Concentración (gramos/100 cc)	$n_{i\star}$ Número de insectos	y Número de muertes	$\star$ Porcentaje de muertes
0.10	47	8	17.0
0.15	53	14	26.4
0.20	55	24	43.6
0.30	52	32	61.5
0.50	46	38	82.6
0.70	54	50	92.6
0.95	52	50	96.2

Además, el analista buscaba la denominada **dosis eficaz** (DE), es decir, la concentración de nicotina que da como resultado cierta probabilidad. La DE_{50} tiene interés particular, ya que es la concentración que produce una probabilidad de 0.5 de que el "insecto muera".

Este ejemplo se agrupa, por lo que el modelo está dado por

$$E(Y_i) = n_i p_i = \frac{n_i}{1 + e^{-(\beta_0 + \beta_1 x_i)}}.$$

Los estimadores de β_0 y β_1 y sus errores estándar se encuentran usando el método de máxima verosimilitud. Las pruebas sobre los coeficientes individuales se encuentran con el estadísticos χ^2 en vez de t, puesto que no hay una varianza común σ^2. El estadístico χ^2 se obtiene a partir de $\left(\frac{\text{coef}}{\text{error estándar}}\right)^2$.

Así, se llega a la siguiente salida de *SAS* PROC LOGIST.

	Análisis de los estimadores de los parámetros				
	df	**Estimador**	**Error estándar**	**Chi-cuadrada**	**Valor *P***
β_0	1	−1.7361	0.2420	51.4482	< 0.0001
β_1	1	6.2954	0.7422	71.9399	< 0.0001

Ambos coeficientes son significativamente distintos de cero. Por lo que el modelo ajustado que se emplea para predecir la probabilidad de "muerte" está dado por

$$\hat{p} = \frac{1}{1 + e^{-(-1.7361 + 6.2954x)}}.$$

Estimación de la dosis eficaz

El estimador de la DE_{50} se encuentra de forma muy sencilla a partir del estimador b_0 para β_0 y b_1 para β_1. Con la función logística se observa que

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x.$$

Como resultado, para $p = 0.5$, se halla un estimador de x a partir de

$$b_0 + b_1 x = 0.$$

Así, la DE_{50} está dada por

$$x = -\left(\frac{b_0}{b_1}\right) = 0.276 \text{ gramos/100 cc.}$$

Concepto de razón de probabilidad

Otra forma de inferencia que se lleva a cabo de manera conveniente con la regresión logística se obtiene del uso de la razón de probabilidad, la cual está diseñada para determinar cómo se incrementa la **razón de probabilidad** $= \frac{p}{1-p}$, conforme ocurren cambios en los valores del regresor. Por ejemplo, en el caso del ejemplo 12.15, quizá se deseara saber cómo se incrementan las probabilidades si se incrementara la dosis en, digamos, 0.2 gramos/100 cc.

Definición 12.1: En la regresión logística, una **razón de probabilidad** es la razón de la probabilidad de éxito en la condición 2 a la de la condición 1 en los regresores, es decir,

$$\frac{[p/(1-p)]_2}{[p/(1-p)]_1}.$$

Esto permite que el analista tenga un sentido de la utilidad al cambiar el regresor en cierto número de unidades. Ahora, como $\left(\frac{p}{1-p}\right) = e^{\beta_0+\beta_1 x}$, entonces para el ejemplo 12.15, la razón que refleja el incremento de las probabilidades de éxito cuando aumenta la dosis de nicotina en 0.2 gramos/100 cc, está dada por

$$e^{0.2b_1} = e^{(0.2)(6.2954)} = 3.522.$$

La implicación de que una razón de probabilidades sea de 3.522 es que la probabilidad de éxito mejora en un factor de 3.522 cuando la dosis de nicotina aumenta en 0.2 gramos/100 cc.

Ejercicios de repaso

12.60 En el Departamento de Pesca y Vida Silvestre en el Instituto Politécnico y Universidad Estatal de Virginia, se realizó un experimento para estudiar el efecto de las características de la corriente sobre la biomasa de los peces. Las variables regresoras son las siguientes: profundidad promedio (de 50 celdas) (x_1); área de la cubierta en la corriente (es decir, riberas socavadas, troncos, cantos rodados, etcétera) (x_2); cubierta porcentual de material translúcido (promedio de 12) (x_3); área $\geq$ 25 centímetros en profundidad (x_4). La respuesta es y, la biomasa de los peces. Los datos son los siguientes:

Obs.	y	x_1	x_2	x_3	x_4
1	100	14.3	15.0	12.2	48.0
2	388	19.1	29.4	26.0	152.2
3	755	54.6	58.0	24.2	469.7
4	1288	28.8	42.6	26.1	485.9
5	230	16.1	15.9	31.6	87.6
6	0	10.0	56.4	23.3	6.9
7	551	28.5	95.1	13.0	192.9
8	345	13.8	60.6	7.5	105.8
9	0	10.7	35.2	40.3	0.0
10	348	25.9	52.0	40.3	116.6

a) Ajuste una regresión lineal múltiple que incluya las cuatro variables regresoras.

b) Utilice C_p, R^2 y s^2 para determinar el mejor subconjunto de variables. Calcule dichos estadísticos para todos los subconjuntos posibles.

c) Compare lo adecuado de los modelos de los incisos *a*) y *b*), para efectos de predecir la biomasa de los peces.

12.61 Demuestre que, en un conjunto de datos de regresión lineal múltiple,

$$\sum_{i=1}^{n} h_{ii} = p.$$

12.62 Se efectuó un experimento sencillo para ajustar una ecuación de regresión múltiple que relaciona al producto y con la temperatura x_1, el tiempo de reacción x_2 y la concentración de uno de los reactivos x_3. Se eligieron dos niveles de cada variable y se hicieron mediciones correspondientes a las variables independientes definidas, como sigue:

y	x_1	x_2	x_3
7.6	−1	−1	−1
5.5	1	−1	−1
9.2	−1	1	−1
10.3	−1	−1	1
11.6	1	1	−1
11.1	1	−1	1
10.2	−1	1	1
14.0	1	1	1

a) Con las variables definidas, estime la ecuación de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

b) Haga la partición de SSR, la suma cuadrática de la regresión, en tres componentes de un grado de libertad atribuibles a x_1, x_2 y x_3, respectivamente. Construya una tabla de análisis de varianza donde se indiquen pruebas de significancia sobre cada variable. Comente los resultados.

12.63 En un experimento de ingeniería química que tiene que ver con la transferencia de calor en una capa de fluido superficial, se recabaron datos sobre las cuatro variables regresoras siguientes: tasa de flujo del gas fluido, lb/hr (x_1); tasa de flujo del gas supercaliente, lb/hr (x_2); abertura de la boquilla de entrada del gas supercaliente, milímetros (x_3); temperatura de entrada del gas SUPERNATANT, °F (x_4). Las respuestas medidas son la eficacia de la transferencia de calor (y_1); la eficacia térmica (y_2). Los datos son los siguientes:

Obs.	y_1	y_2	x_1	x_2	x_3	x_4
1	41.852	38.75	69.69	170.83	45	219.74
2	155.329	51.87	113.46	230.06	25	181.22
3	99.628	53.79	113.54	228.19	65	179.06
4	49.409	53.84	118.75	117.73	65	281.30
5	72.958	49.17	119.72	117.69	25	282.20
6	107.702	47.61	168.38	173.46	45	216.14
7	97.239	64.19	169.85	169.85	45	223.88
8	105.856	52.73	169.85	170.86	45	222.80
9	99.348	51.00	170.89	173.92	80	218.84
10	111.907	47.37	171.31	173.34	25	218.12
11	100.008	43.18	171.43	171.43	45	219.20
12	175.380	71.23	171.59	263.49	45	168.62
13	117.800	49.30	171.63	171.63	45	217.58
14	217.409	50.87	171.93	170.91	10	219.92
15	41.725	54.44	173.92	71.73	45	296.60
16	151.139	47.93	221.44	217.39	65	189.14
17	220.630	42.91	222.74	221.73	25	186.08
18	131.666	66.60	228.90	114.40	25	285.80
19	80.537	64.94	231.19	113.52	65	286.34
20	152.966	43.18	236.84	167.77	45	221.72

Considere el modelo para predecir la respuesta del coeficiente de transferencia de calor

$$y_{1i} = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ji} + \sum_{i=1}^{4} \beta_{jj} x_{ji}^2 + \sum\sum_{j \neq l} \beta_{jl} x_{ji} x_{li} + \epsilon_i, \quad i = 1, 2, \ldots, 20.$$

a) Calcule PRESS y $\sum_{i=1}^{n} |y_i - \hat{y}_{i,-i}|$ para ajustar con regresión por mínimos cuadrados al modelo anterior.

b) Ajuste un modelo de segundo orden con x_4 eliminada por completo (es decir, elimine todos los términos que impliquen x_4). Calcule los criterios de predicción para el modelo reducido. Comente sobre lo adecuado de x_4 para la predicción del coeficiente de transferencia de calor.

c) Repita los incisos *a*) y *b*) para la eficacia térmica.

12.64 En la fisiología del deporte, una medición objetiva de la condición física es el consumo de oxígeno en volumen por unidad de peso corporal por unidad de tiempo. Se estudiaron 31 individuos en un experimento con la finalidad de modelar el consumo de oxígeno contra: edad en años (x_1); peso en kilogramos (x_2); tiempo en que se corre 1.5 millas (x_3); tasa del pulso en reposo (x_4); tasa del pulso al final de la carrera (x_5); tasa máxima del pulso durante la carrera (x_6).

ID	y	x_1	x_2	x_3	x_4	x_5	x_6
1	44.609	44	89.47	11.37	62	178	182
2	45.313	40	75.07	10.07	62	185	185
3	54.297	44	85.84	8.65	45	156	168
4	59.571	42	68.15	8.17	40	166	172
5	49.874	38	89.02	9.22	55	178	180
6	44.811	47	77.45	11.63	58	176	176
7	45.681	40	75.98	11.95	70	176	180
8	49.091	43	81.19	10.85	64	162	170
9	39.442	44	81.42	13.08	63	174	176
10	60.055	38	81.87	8.63	48	170	186
11	50.541	44	73.03	10.13	45	168	168
12	37.388	45	87.66	14.03	56	186	192
13	44.754	45	66.45	11.12	51	176	176
14	47.273	47	79.15	10.60	47	162	164
15	51.855	54	83.12	10.33	50	166	170
16	49.156	49	81.42	8.95	44	180	185
17	40.836	51	69.63	10.95	57	168	172
18	46.672	51	77.91	10.00	48	162	168
19	46.774	48	91.63	10.25	48	162	164
20	50.388	49	73.37	10.08	76	168	168
21	39.407	57	73.37	12.63	58	174	176
22	46.080	54	79.38	11.17	62	156	165
23	45.441	52	76.32	9.63	48	164	166
24	54.625	50	70.87	8.92	48	146	155
25	45.118	51	67.25	11.08	48	172	172
26	39.203	54	91.63	12.88	44	168	172
27	45.790	51	73.71	10.47	59	186	188
28	50.545	57	59.08	9.93	49	148	155

ID	y	x_1	x_2	x_3	x_4	x_5	x_6
29	48.673	49	76.32	9.40	56	186	188
30	47.920	48	61.24	11.50	52	170	176
31	47.467	52	82.78	10.50	53	170	172

a) Realice una regresión progresiva con un nivel de significancia de 0.25 en la entrada. Proporcione el modelo final.

b) Estudie todos los subconjuntos posibles usando s^2, C_p, R^2 y R^2_{aju}. Tome una decisión y determine el modelo final.

12.65 Considere los datos del ejercicio de repaso 12.62. Suponga que es de interés agregar algunos términos de "interacción". En específico, considere el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_{12} x_{1i} x_{2i} + \beta_{13} x_{1i} x_{3i} + \beta_{23} x_{2i} x_{3i} + \beta_{123} x_{1i} x_{2i} x_{3i} + \epsilon_i.$$

a) ¿Aún se tiene ortogonalidad? Comente.

b) Considerando el modelo ajustado en el inciso *a*), ¿puede usted encontrar intervalos de predicción y de confianza sobre la respuesta media? ¿Por qué?

c) Utilice el modelo con $\beta_{123} x_1 x_2 x_3$ eliminada. Para determinar si son necesarias las interacciones (como un todo), pruebe

$$H_0: \ \beta_{12} = \beta_{13} = \beta_{23} = 0.$$

Dé valores P y saque conclusiones.

12.66 Para extraer petróleo crudo, se utiliza una técnica de inyección de dióxido de carbono (CO_2). El flujo de CO_2 envuelve el petróleo y lo desplaza. En el experimento, se introducen tubos de flujo en bolsones de muestras de petróleo que contienen una cantidad conocida de éste. Los bolsones de petróleo se inyectan con CO_2, y se registra el porcentaje de petróleo desplazado, usando tres valores diferentes de presión del flujo y tres valores diferentes de ángulos de introducción. Considere el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i.$$

Ajuste el modelo anterior a los datos y sugiera cualquier modificación al modelo que considere necesaria.

Presión lb/pulg², x_1	Ángulo de inyección, x_2	Porcentaje de recuperación de petróleo, y
1000	0	60.58
1000	15	72.72
1000	30	79.99
1500	0	66.83
1500	15	80.78
1500	30	89.78
2000	0	69.18
2000	15	80.31
2000	30	91.99

Fuente: Wang, G. C. "Microscopic Investigations of CO_2 Flooding Process", *Journal of Petroleum Technology*, vol. 34, núm. 8, agosto de 1982.

12.67 Un artículo del *Journal of Pharmaceutical Sciences* (vol. 80, 1991) presenta datos de la solubilidad de una fracción molar de un soluto a temperatura constante. También se midió la dispersión, x_1, y los parámetros de solubilidad del enlace bipolar y de hidrógeno x_2, y x_3. En la tabla siguiente se presenta una parte de los datos. En el modelo, y es el logaritmo negativo de la fracción molar. Ajuste el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i,$$

para $i = 1, 2, \ldots, 20$.

a) Pruebe H_0: $\beta_1 = \beta_2 = \beta_3 = 0$.

b) Grafique los residuos studentizados contra x_1, x_2 y x_3 (tres gráficas). Haga comentarios.

c) Considere dos modelos adicionales que son competidores del modelo anterior:

Modelo 2: Agregue x_1^2, x_2^2, x_3^2

Modelo 3: Agregue $x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3$.

Con estos tres modelos utilice PRESS y C_p para llegar al mejor de los tres.

Obs.	y	x_1	x_2	x_3
1	0.2220	7.3	0.0	0.0
2	0.3950	8.7	0.0	0.3
3	0.4220	8.8	0.7	1.0
4	0.4370	8.1	4.0	0.2
5	0.4280	9.0	0.5	1.0
6	0.4670	8.7	1.5	2.8
7	0.4440	9.3	2.1	1.0
8	0.3780	7.6	5.1	3.4
9	0.4940	10.0	0.0	0.3
10	0.4560	8.4	3.7	4.1
11	0.4520	9.3	3.6	2.0
12	0.1120	7.7	2.8	7.1
13	0.4320	9.8	4.2	2.0
14	0.1010	7.3	2.5	6.8
15	0.2320	8.5	2.0	6.6
16	0.3060	9.5	2.5	5.0
17	0.0923	7.4	2.8	7.8
18	0.1160	7.8	2.8	7.7
19	0.0764	7.7	3.0	8.0
20	0.4390	10.3	1.7	4.2

12.68 Se realizó un estudio para determinar si cambios en el estilo de vida podrían sustituir la medicación para reducir la presión sanguínea de los individuos hipertensos. Los factores considerados fueron una dieta saludable con un programa de ejercicios, la dosis común de medicamentos contra la hipertensión y la no intervención. También se calculó el índice de masa corporal (IMC) anterior al tratamiento, debido a que se sabe que afecta la presión sanguínea. La respuesta considerada en este estudio

Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	490177488	98035498	237.79	<.0001
Error	11	4535052	412277		
Corrected Total	16	494712540			

Root MSE	642.08838	R-Square	0.9908
Dependent Mean	4978.48000	Adj R-Sq	0.9867
Coeff Var	12.89728		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1962.94816	1071.36170	1.83	0.0941
x1	Average Daily Patient Load	1	-15.85167	97.65299	-0.16	0.8740
x2	Monthly X-Ray Exposure	1	0.05593	0.02126	2.63	0.0234
x3	Monthly Occupied Bed Days	1	1.58962	3.09208	0.51	0.6174
x4	Eligible Population in the Area/100	1	-4.21867	7.17656	-0.59	0.5685
x5	Average Length of Patients Stay in Days	1	-394.31412	209.63954	-1.88	0.0867

Figura 12.9: Salida del *SAS* para el ejercicio de repaso 12.69; parte I.

cambió con la presión sanguínea. El grupo de variables tiene los siguientes niveles.

1 = Dieta saludable y programa de ejercicios.

2 = Medicación.

3 = No intervención.

Cambio en la presión sanguínea	Grupo	IMC
−32	1	27.3
−21	1	22.1
−26	1	26.1
−16	1	27.8
−11	2	19.2
−19	2	26.1
−23	2	28.6
−5	2	23.0
−6	3	28.1
5	3	25.3
−11	3	26.7
14	3	22.3

a) Ajuste un modelo adecuado utilizando los datos anteriores. ¿Pareciera que el ejercicio y la dieta podrían utilizarse en forma eficaz para disminuir la presión sanguínea? Explique su respuesta a partir de los resultados.

b) ¿El ejercicio y la dieta serían una alternativa eficaz a la medicación?

(*Sugerencia:* Para responder a estas preguntas, quizás usted desee construir el modelo en más de una forma.)

12.69 Estudio de caso: Considere el conjunto de datos para el ejercicio 12.12 de la página 454 (datos de un hospital). El conjunto de datos se repite en seguida.

a) Las salidas de *SAS* PROC REG presentadas en las figuras 12.9 y 12.10 suministran una cantidad considerable de información. El propósito es detectar los valores extremos y, a final de cuentas, determinar cuáles términos del modelo deben utilizarse en la versión final de éste.

b) Haga comentarios sobre cuáles son otros análisis que deberían hacerse.

c) Elabore análisis apropiados y escriba sus conclusiones con respecto al modelo final.

Sitio	x_1	x_2	x_3	x_4	x_5	y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1003.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20106	3655.08	180.5	6.15	3503.93
11	96.00	13313	2912.00	60.9	5.88	3571.59
12	131.42	10771	3921.00	103.7	4.88	3741.40
13	127.21	15543	3865.67	126.8	5.50	4026.52
14	252.90	36194	7684.10	157.7	7.00	10343.81
15	409.20	34703	12446.33	169.4	10.75	11732.17
16	463.70	39204	14098.40	331.4	7.05	15414.94
17	510.22	86533	15524.00	371.6	6.35	18854.45

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	566.5200	775.0251	241.2323	244.0765	1306	-734.6494	2285
2	696.8200	740.6702	331.1402	11.8355	1470	-849.4275	2331
3	1033	1104	278.5116	490.9234	1717	-436.5244	2644
4	1604	1240	268.1298	650.3459	1831	-291.0028	2772
5	1611	1564	211.2372	1099	2029	76.6816	3052
6	1613	2151	279.9293	1535	2767	609.5796	3693
7	1854	1690	218.9976	1208	2172	196.5345	3183
8	2161	1736	468.9903	703.9948	2768	-13.8306	3486
9	2306	2737	290.4749	2098	3376	1186	4288
10	3504	3682	585.2517	2394	4970	1770	5594
11	3572	3239	189.0989	2823	3655	1766	4713
12	3741	4353	328.8507	3630	5077	2766	5941
13	4027	4257	314.0481	3566	4948	2684	5830
14	10344	8768	252.2617	8213	9323	7249	10286
15	11732	12237	573.9168	10974	13500	10342	14133
16	15415	15038	585.7046	13749	16328	13126	16951
17	18854	19321	599.9780	18000	20641	17387	21255

Obs	Residual	Std Error Residual	Student Residual	-2-1 0 1 2
1	-208.5051	595.0	-0.350	| | |
2	-43.8502	550.1	-0.0797	| | |
3	-70.7734	578.5	-0.122	| | |
4	363.1244	583.4	0.622	| |* |
5	46.9483	606.3	0.0774	| | |
6	-538.0017	577.9	-0.931	| *| |
7	164.4696	603.6	0.272	| | |
8	424.3145	438.5	0.968	| |* |
9	-431.4090	572.6	-0.753	| *| |
10	-177.9234	264.1	-0.674	| *| |
11	332.6011	613.6	0.542	| |* |
12	-611.9330	551.5	-1.110	| **| |
13	-230.5684	560.0	-0.412	| | |
14	1576	590.5	2.669	| |***** |
15	-504.8574	287.9	-1.753	| ***| |
16	376.5491	263.1	1.431	| |** |
17	-466.2470	228.7	-2.039	| ****| |

Figura 12.10: Salida de *SAS* para el ejercicio de repaso 12.69; parte II.

12.70 Demuestre que al elegir el llamado mejor modelo del subconjunto de entre una serie de modelos candidato, si se selecciona el modelo con la menor s^2, ello equivale a escoger el modelo con el R^2_{aj} más pequeño.

12.71 A partir de un conjunto de datos de respuesta a la dosis de estreptomicina, un investigador desea desarrollar una relación entre la proporción de linfoblastos muestreados que contienen aberraciones y la dosis del medicamento. Se aplicaron cinco niveles de dosis a los conejos que se emplearon para el experimento. Los datos son los siguientes:

Dosis (mg/kg)	Número de linfoblastos	Número de aberraciones
0	600	15
30	500	96
60	600	187
75	300	100
90	300	145

En la bibliografía, véase Myers, 1990.

a) Haga una regresión logística para el conjunto de datos, y así estime β_0 y β_1 en el modelo,

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}},$$

donde n es el número de linfoblastos, x es la dosis y p la probabilidad de una aberración.

b) Muestre los resultados de pruebas χ^2 que revelen la significancia de los coeficientes de regresión β_0 y β_1.

c) Estime la DE$_{50}$ e interprétela.

12.72 En un experimento para estudiar el efecto de la carga, x, en lb/pulgadas^2, sobre la probabilidad de falla de especímenes de cierto tipo de tela, varios especímenes se expusieron a cargas de entre 5 lb/pulg^2 a 90 lb/pulg^2. Se observaron los números de "fallas". Los datos son los siguientes:

Carga	Número de especímenes	Número de Fallas
5	600	13
35	500	95
70	600	189
80	300	95
90	300	130

a) Utilice regresión logística para ajustar el modelo

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}},$$

donde p es la probabilidad de falla y x es la carga.

b) Emplee el concepto de razón de probabilidad para determinar el incremento de la probabilidad de falla que resulta de aumentar la carga en 20 lb/pulg^2.

12.13 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

En este capítulo se estudiaron varios procedimientos para usarlos en el "intento" de encontrar el mejor modelo. Sin embargo, uno de los errores más importantes en el trabajo de los científicos e ingenieros novatos es que existe un **modelo lineal verdadero**, y que es posible encontrarlo. En la mayoría de fenómenos de la ciencia, las relaciones entre las variables científicas son de naturaleza no lineal y se desconoce el modelo verdadero. Los modelos estadísticos lineales son **aproximaciones empíricas**.

A veces, la selección del modelo por adoptar depende de cuál es la información que necesita obtenerse del mismo. ¿Va a usarse para realizar predicciones? ¿Para explicar el papel de cada regresor? Esta "selección" puede ser difícil ante la presencia de colinealidad. Es un hecho que para muchos problemas de regresión hay modelos múltiples muy similares en cuanto a su desempeño. Para mayores detalles, véase la referencia de Myers (1990).

Uno de los equívocos más nocivos del material de este capítulo consiste en dar demasiada importancia a R^2 en la selección del llamado *mejor modelo*. Es importante recordar que para cualquier conjunto de datos, se puede obtener una R^2 tan grande como se desee, dentro de la restricción de que $0 \leq R^2 \leq 1$. **Prestar mucha atención a R^2 con frecuencia lleva al sobreajuste**.

En este capítulo se dio mucha atención a la detección de los valores extremos. Un error clásico y serio de los estadísticos estriba en la decisión acerca de la detección de los valores extremos. Los autores esperan que quede claro que el analista no debería por ningún motivo detectar los valores extremos, eliminarlos del conjunto de datos, ajustar un modelo nuevo, informar sobre los valores extremos, y así sucesivamente. Éste es un procedimiento tentador y desastroso para llegar a un modelo que se ajuste bien a los datos, lo cual resulta en un ejemplo de **cómo mentir con estadísticos**.

Si se detecta un valor extremo, debe revisarse la historia de los datos en busca de posibles errores de captura o de procedimiento antes de eliminarlos del conjunto de datos. Debe recordarse que, por definición, un valor extremo es aquel para el cual el modelo no se ajusta bien. El problema podría no estar en los datos sino en la selección del modelo. Cambiar el modelo quizás haría que el punto no se detecte como un valor extremo.

Capítulo 13

Experimentos con un solo factor: General

13.1 Técnica del análisis de varianza

En el material sobre estimación y prueba de hipótesis que se cubrió en los capítulos 9 y 10, en cada caso nos restringimos a no considerar más de dos parámetros de la población. Ése fue el caso, por ejemplo, en la prueba de la igualdad de dos medias poblacionales, usando muestras independientes de poblaciones normales con varianza común pero desconocida, donde fue necesario obtener una estimación de unión de σ^2.

Dicho material, que trata con inferencias de dos muestras, representa un caso especial de lo que se denomina el *problema de un solo factor*. Por ejemplo, en el ejercicio 35, sección 10.8, el tiempo de supervivencia está medido para dos muestras de ratones, de los que una muestra recibió un tratamiento de suero contra la leucemia y la otra no lo recibió. En este caso, decimos que hay *un factor*, llamado *tratamiento*, y el factor se halla en *dos niveles*. Si en el proceso de muestreo se utilizaran varios tratamientos en competencia, serían necesarias más muestras de ratones. En ese caso, el problema implicaría un factor con más de dos niveles y, por ello, con más de dos muestras.

En el problema de $k > 2$ muestras, se supone que hay k muestras provenientes de k poblaciones. Un procedimiento muy común que se utiliza cuando se prueban medias poblacionales se denomina ***análisis de varianza***, o ANOVA.

El análisis de varianza no es, por supuesto, una técnica nueva, si el lector ha estudiado el material acerca de la teoría de la regresión. Se usa el enfoque del análisis de varianza para hacer una partición de la suma total de cuadrados en una parte que se deba a la regresión, y otra que se deba al error.

Suponga que en un experimento industrial a un ingeniero le interesa la forma en que la absorción media de humedad del concreto varía para 5 agregados de concreto diferentes. Las muestras se exponen a la humedad durante 48 horas. Se decidió que para cada agregado deben probarse 6 muestras, lo que hace que se requiera probar un total de 30 muestras. En la tabla 13.1 se muestran los datos registrados.

El modelo que se considera para esta situación es el siguiente. Se tomaron 6 observaciones de cada una de las 5 poblaciones, con medias μ_1, $\mu_2, \ldots,$ μ_5, respectivamente. Se desea probar

H_0: $\mu_1 = \mu_2 = \cdots = \mu_5$,

H_1: Al menos dos de las medias no son iguales.

Tabla 13.1: Absorción de humedad en agregados para concreto

Agregado:	**1**	**2**	**3**	**4**	**5**	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Media	553.33	569.33	610.50	465.17	610.67	561.80

Además, estamos interesados en realizar comparaciones individuales entre estas 5 medias poblacionales.

Dos fuentes de variabilidad en los datos

En el procedimiento del análisis de varianza, se supone que cualquier variación que exista entre los promedios de los agregados se atribuye a **1.** la variación en la absorción entre observaciones dentro de los tipos de agregados, y **2.** la variación debida a los tipos de agregados, es decir, a las diferencias en la composición química de los agregados. Por supuesto, la **variación dentro de los agregados** se debe a varias causas. Quizá las condiciones de temperatura y humedad no se mantuvieron constantes durante el experimento. Es posible que haya habido cierta cantidad de heterogeneidad en los lotes de materias primas que se usaron. En todo caso, debe considerarse la variación dentro de la muestra como una **variación aleatoria o al azar**, y parte del objetivo del análisis de varianza es determinar si las diferencias entre las 5 medias muestrales son lo que se esperaría debido a la sola variación aleatoria.

En esta etapa surgen muchas preguntas acerca del problema anterior. Por ejemplo, ¿cuántas muestras deben probarse para cada agregado? Ésta es una pregunta que desafía continuamente al analista. Además, ¿qué pasa si la variación al interior de la muestra es tan grande que sería difícil para un procedimiento estadístico detectar las diferencias sistemáticas? ¿Es posible controlar de manera sistemática fuentes externas de variación y así eliminarlas de la parte que llamamos variación aleatoria? En las secciones siguientes intentaremos responder estas y otras preguntas.

13.2 La estrategia del diseño de experimentos

En los capítulos 9 y 10 se estudiaron el concepto de la estimación y la prueba de hipótesis para el caso de dos muestras, con la salvedad importante de la manera en que se realiza el experimento. Esto forma parte de la categoría amplia del diseño experimental. Por ejemplo, para la **prueba *t* combinada** que se estudió en el capítulo 10, se supone que los niveles de los factores (los tratamientos, en el ejercicio de los ratones) se asignan al azar a las unidades experimentales (los ratones). En los capítulos 9 y 10 se analizó el concepto de unidades experimentales, y se ilustró con varios ejemplos. En pocas palabras, las unidades experimentales son las unidades

(ratones, pacientes, especímenes de concreto, tiempo) que **proporcionan la heterogeneidad que lleva al error experimental** en una investigación científica. La asignación al azar elimina el sesgo que podría originarse en una asignación sistemática. El objetivo consiste en distribuir en forma uniforme entre los niveles de los factores los riesgos que introduce la heterogeneidad de las unidades experimentales. Una asignación al azar simula mejor las condiciones presentes en el modelo. En la sección 13.8 se estudia el **bloqueo** en los experimentos. En los capítulos 9 y 10 se presentó el concepto de bloqueo, cuando se efectuaron comparaciones entre las medias usando el **pareo**, es decir, la división de las unidades experimentales en pares homogéneos denominados **bloques**. Entonces, los niveles de los factores o tratamientos se asignan al azar dentro de los bloques. El propósito del bloque es reducir el error experimental eficaz. En este capítulo se extiende de manera natural el pareo a bloques de tamaño mayor, con el análisis de varianza como la herramienta analítica principal.

13.3 Análisis de varianza de un solo factor: Diseño completamente al azar (ANOVA de un solo factor)

De k poblaciones se seleccionan muestras aleatorias de tamaño n. Las k poblaciones diferentes se clasifican con base en un criterio único, como tratamientos o grupos diferentes. En la actualidad, el término **tratamiento** se utiliza, por lo general, para designar las diversas clasificaciones, ya sean diferentes agregados, analistas, fertilizadores o regiones del país.

Suposiciones e hipótesis del ANOVA de un solo factor

Se supone que las k poblaciones son independientes y están distribuidas en forma normal con medias $\mu_1, \mu_2, \ldots, \mu_k$, y varianza común σ^2. Como se indicó en la sección 13.2, estas suposiciones son más aceptables mediante la aleatoriedad. Se desean obtener métodos adecuados para probar las hipótesis

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$.

H_1: Al menos dos de las medias no son iguales.

Sea que y_{ij} denote la j-ésima observación del i-ésimo tratamiento, y el acomodo de los datos es el que se observa en la tabla 13.2. Aquí, $Y_{i.}$ es el total de todas las observaciones de la muestra, del i-ésimo tratamiento, $\bar{y}_{i.}$, es la media de todas las observaciones en la muestra del i-ésimo tratamiento, $Y_{..}$ es el total de todas las nk observaciones, y $\bar{y}_{..}$ es la media de todas las nk observaciones.

Modelo de ANOVA para un solo factor

Cada observación puede escribirse en la forma

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

donde ϵ_{ij} mide la desviación que tiene la observación j-ésima de la i-ésima muestra, con respecto de la media del tratamiento correspondiente. El término ϵ_{ij} representa el error aleatorio y juega el mismo papel que los términos del error en los modelos

Tabla 13.2: k muestras aleatorias

Tratamiento:	**1**	**2**	**...**	***i***	**...**	***k***	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	$\cdots$	$Y_{i.}$	$\cdots$	$Y_{k.}$	$Y_{..}$
Media	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\cdots$	$\bar{y}_{i.}$	$\cdots$	$\bar{y}_{k.}$	$\bar{y}_{..}$

de regresión. Una forma alternativa y preferible de esta ecuación se obtiene al sustituir $\mu_{\text{i}} = \mu + \alpha_i$, sujeta a la restricción $\sum_{i=1}^{k} \alpha_i = 0$. Por lo tanto, se escribe

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

donde μ tan sólo es la **media global** de todas las μ_i, es decir,

$$\mu = \frac{1}{k} \sum_{i=1}^{k} \mu_i,$$

y α_i se denomina el **efecto** del i-ésimo tratamiento.

La prueba de la hipótesis nula de que k medias poblacionales son iguales, contra la alternativa de que al menos dos de las medias son distintas, ahora puede reemplazarse por las hipótesis equivalentes.

H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$,

H_0: Al menos una de las α_i no es igual a cero.

Resolución de la variabilidad total en componentes

Nuestra prueba se basará en una comparación de dos estimadores independientes de la varianza poblacional común σ^2. Dichos estimadores se obtendrán haciendo la partición de la variabilidad total de nuestros datos, denotados mediante la sumatoria doble

$$\sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2,$$

en dos componentes.

Teorema 13.1:

Identidad de la suma de cuadrados

$$\sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2.$$

En lo que sigue, será conveniente identificar los términos de la identidad de la suma de cuadrados con la notación que se presenta en seguida:

Tres medidas importantes de variabilidad

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \text{suma total de cuadrados,}$$

$$SSA = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = \text{suma de los cuadrados del tratamiento,}$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 = \text{suma de los cuadrados de los errores.}$$

Ahora, la identidad de la suma de los cuadrados puede representarse simbólicamente con la ecuación

$$SST = SSA + SSE.$$

La identidad anterior expresa cómo las variaciones entre tratamientos y dentro de éstos se suman para formar la suma total de cuadrados. Sin embargo, puede ampliarse mucho la perspectiva si se investiga el **valor esperado tanto de *SSA*** como de ***SSE***. Finalmente, se desarrollarán estimadores de la varianza que formulen la razón que se va a usar para probar la igualdad de las medias poblacionales.

Teorema 13.2:

$$E(SSA) = (k-1)\sigma^2 + n\sum_{i=1}^{k}\alpha_i^2.$$

La prueba del teorema se deja como ejercicio para el lector (véase el ejercicio 13.2 de la página 521).

Si H_0 es verdadera, un estimador de σ^2 con base en $k-1$ grados de libertad, está dado por la expresión

Media cuadrática del tratamiento

$$s_1^2 = \frac{SSA}{k-1}.$$

Si H_0 es verdadera y por ello cada α_i en el teorema 13.2 es igual a cero, se observa que

$$E\left(\frac{SSA}{k-1}\right) = \sigma^2,$$

y s_1^2 es un estimador insesgado de σ^2. Sin embargo, si H_1 es verdadera, se tiene que

$$E\left(\frac{SSA}{k-1}\right) = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k}\alpha_i^2,$$

y s_1^2 estima a σ^2 más un término adicional, que mide la variación debida a los efectos sistemáticos.

Otro estimador independiente de σ^2, con base en $k(n-1)$ grados de libertad, es la fórmula familiar

Error cuadrático medio

$$s^2 = \frac{SSE}{k(n-1)}.$$

Resulta instructivo puntualizar la importancia de los valores esperados de las medias cuadráticas recién expresados. En la sección siguiente se estudia el empleo de la **razón *F*** con la media cuadrática del tratamiento en el numerador. Se observa que cuando H_1 es verdadera, la presencia de la condición $E(s_1^2) > E(s^2)$ sugiere que la razón F se utiliza en el contexto de una **prueba unilateral de cola superior**. Es decir, cuando H_1 es verdadera se esperaría que el numerador s_1^2 fuera mayor que el denominador.

Uso de la prueba *F* en el ANOVA

El estimador s^2 es insesgado sin que importe la verdad o falsedad de la hipótesis nula (véase el ejercicio 13.1 de la página 521). Es importante notar que la identidad de la suma de cuadrados ha hecho la partición no sólo de la variabilidad total de los datos, sino también del número total de grados de libertad. Es decir,

$$nk - 1 = k - 1 + k(n - 1).$$

Razón *F* para probar la igualdad de las medias

Cuando H_0 es verdadera, la razón $f = s_1^2/s^2$ es un valor de la variable aleatoria F, que tiene distribución F con $k - 1$ y $k(n - 1)$ grados de libertad. Como s_1^2 sobrestima σ^2 cuando H_0 es falsa, se tiene una prueba de una cola con la región crítica contenida por entero en la cola derecha de la distribución.

Con un nivel de significancia de α, se rechaza la hipótesis nula H_0 cuando

$$f > f_\alpha[k - 1, k(n - 1)].$$

Otro enfoque, el del valor P, sugiere que la evidencia a favor o en contra de H_0 es

$$P = P[f[k - 1, k(n - 1)] > f].$$

Los cálculos para un problema de análisis de varianza, por lo general, se resumen en forma tabular, como se presenta en la tabla 13.3.

Tabla 13.3: Análisis de varianza del ANOVA para un solo factor

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
Tratamientos	SSA	$k - 1$	$s_1^2 = \frac{SSA}{k-1}$	$\frac{s_1^2}{s^2}$
Error	SSE	$k(n - 1)$	$s^2 = \frac{SSE}{k(n-1)}$	
Total	SST	$kn - 1$		

Ejemplo 13.1: Pruebe la hipótesis de que $\mu_1 = \mu_2 = \cdots = \mu_5$ con un nivel de significancia de 0.05, para los datos de la tabla 13.1 sobre la absorción de humedad por varios tipos de agregados para cemento.

Solución: H_0: μ_1,$= \mu_2 = \cdots = \mu_5$,

H_1: Al menos dos de las medias no son iguales.

$\alpha = 0.05$.

Región crítica: $f > 2.76$ con $v_1 = 4$ y $v_2 = 25$ grados de libertad. Los cálculos de la suma de cuadrados dan

$$SST = 209,377,$$

$$SSA = 85,356,$$

$$SSE = 209,377 - 85,356 = 124,021.$$

En la figura 13.1 se muestran estos resultados y el resto de los cálculos del procedimiento de SAS para el ANOVA.

```
                        The GLM Procedure
Dependent Variable: moisture

                                          Sum of
Source             DF         Squares     Mean Square    F Value    Pr > F
Model               4      85356.4667      21339.1167       4.30    0.0088
Error              25     124020.3333       4960.8133
Corrected Total    29     209376.8000

  R-Square     Coeff Var      Root MSE    moisture Mean
  0.407669      12.53703      70.43304         561.8000

Source             DF       Type I SS     Mean Square    F Value    Pr > F
aggregate           4     85356.46667     21339.11667       4.30    0.0088
```

Figura 13.1: Salida de *SAS* para el procedimiento de análisis de varianza.

Decisión: Rechace H_0 y concluya que los agregados no tienen la misma media de absorción. El valor P para $f = 4.30$ es más pequeño que 0.01.

Durante el trabajo experimental es frecuente que se pierdan algunas de las observaciones deseadas. Los animales del experimento mueren, el material experimental se daña o los seres humanos abandonan el estudio. El análisis anterior para un tamaño igual de muestra todavía debe validarse con la modificación leve de las fórmulas de la suma de cuadrados. Ahora se supondrá que las k muestras aleatorias son de tamaño $n_1, n_2, \ldots, n_k$, respectivamente.

Suma de cuadrados; tamaños desiguales de las muestras

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2, \quad SSA = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2, \quad SSE = SST - SSA$$

Después, se hace la partición de los grados de libertad, como antes: $N - 1$ para SST, $k - 1$ para SSA, y $N - 1 - (k - 1) = N - k$ para SSE, donde $N = \sum_{i=1}^{k} n_i$.

Ejemplo 13.2: Parte de un estudio dirigido por el Instituto Politécnico y Universidad Estatal de Virginia se diseñó para medir los niveles de actividad del suero fosfatado alcalino (en unidades de Bessey-Lowry) en niños con crisis epilépticas que recibieron terapia anticonvulsiva al cuidado de un médico privado. Para el estudio se reclutaron 45 sujetos y se clasificaron en cuatro grupos, según el medicamento:

G-1: Control (no recibieron anticonvulsivos ni tenían historia de crisis epilépticas).

G-2: Fenobarbital.

G-3: Carbamazepina.

G-4: Otros anticonvulsivos.

De las muestras de sangre tomadas de cada sujeto, se determinó el nivel de actividad del suero fosfatado alcalino y se registró según se observa en la tabla 13.4. Pruebe la hipótesis de que con un nivel de significancia de 0.05, el promedio del nivel de actividad del suero fosfatado alcalino es el mismo para los cuatro grupos del medicamento.

Tabla 13.4: Nivel de actividad del suero fosfatado alcalino

G-1		G-2	G-3	G-4
49.20	97.50	97.07	62.10	110.60
44.54	105.00	73.40	94.95	57.10
45.80	58.05	68.50	142.50	117.60
95.84	86.60	91.85	53.00	77.71
30.10	58.35	106.60	175.00	150.00
36.50	72.80	0.57	79.50	82.90
82.30	116.70	0.79	29.50	111.50
87.85	45.15	0.77	78.40	
105.00	70.35	0.81	127.50	
95.22	77.40			

Solución:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$,

H_1: Al menos dos de las medias no son iguales.

$\alpha = 0.05$.

Región crítica: $f > 2.836$, con interpolación de los valores de la tabla A.6.
Cálculos: $Y_{1.} = 1460.25$, $Y_{2.} = 440.36$, $Y_{3.} = 842.45$, $Y_{4.} = 707.41$ y $Y_{..} = 3450.47$. En la salida de MINITAB que se presenta en la figura 13.2 se incluye el análisis de varianza.
Decisión: Rechace H_0 y concluya que los niveles de actividad del suero fosfatado alcalino para los cuatro grupos de medicamentos no siempre son los mismos. El valor P es de 0.02.

Como conclusión del análisis de la varianza para la clasificación de un solo factor, mencionaremos las ventajas que tiene que elegir muestras del mismo tamaño en vez de otras de tamaños distintos. La primera ventaja es que la razón f no es sensible a fallos pequeños de la suposición de varianzas iguales para las k poblaciones cuando las muestras son del mismo tamaño. La segunda consiste en que las muestras de tamaño igual minimizan la probabilidad de cometer un error del tipo II.

13.4 Pruebas para la igualdad de diversas varianzas

Aunque la razón f que se obtiene con el procedimiento del análisis de varianza no es sensible a fallos de la suposición de varianzas iguales para las k poblaciones normales

```
One-way ANOVA: G-1, G-2, G-3, G-4

Source  DF     SS    MS     F      P
Factor   3  13939  4646  3.57  0.022
Error   41  53376  1302
Total   44  67315

S = 36.08   R-Sq = 20.71%   R-Sq(adj) = 14.90%

                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level   N    Mean  StDev  --+---------+---------+---------+-------
G-1    20   73.01  25.75             (----*-----)
G-2     9   48.93  47.11  (-------*-------)
G-3     9   93.61  46.57                  (-------*-------)
G-4     7  101.06  30.76                    (--------*--------)
                         --+---------+---------+---------+-------
                         30        60        90       120

Pooled StDev = 36.08
```

Figura 13.2: Análisis de *MINITAB* de la tabla 13.4.

si las muestras son de igual tamaño, debe tenerse precaución y efectuar una prueba preliminar sobre la homogeneidad de las varianzas. En el caso de muestras de tamaños distintos, es claramente aconsejable realizar una prueba como ésa, si existe duda razonable acerca de la homogeneidad de las varianzas de la población. Por lo tanto, suponga que se desea probar la hipótesis nula

$$H_0\colon \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

contra la alternativa

$$H_1\colon \text{No todas las varianzas son iguales.}$$

La prueba que usaremos, denominada **prueba de Bartlett**, se basa en un estadístico cuya distribución muestral proporciona valores críticos exactos cuando los tamaños de muestra son iguales. Dichos valores críticos para tamaños iguales de muestras también pueden utilizarse para obtener aproximaciones muy exactas de los valores críticos para tamaños distintos de muestras.

En primer lugar, calculamos las varianzas de las k muestras $s_1^2,\ s_2^2, \ldots,\ s_k^2$ de tamaños $n_1,\ n_2, \ldots,\ n_k$ con $\sum_{i=1}^{k} n_i = N$. En segundo lugar, se combinan las varianzas muestrales para dar la estimación de unión

$$s_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)s_i^2.$$

Ahora,

$$b = \frac{[(s_1^2)^{n_1-1}(s_2^2)^{n_2-1}\cdots(s_k^2)^{n_k-1}]^{1/(N-k)}}{s_p^2}$$

es un valor de una variable aleatoria B que tiene la **distribución de Bartlett**. Para el caso especial en que $n_1 = n_2 = \cdots = n_k = n$, se rechaza H_0 con un nivel de significancia α si

$$b < b_k(\alpha; n),$$

donde $b_k(\alpha; n)$ es el valor crítico que deja un área de tamaño α en el extremo izquierdo de la distribución de Bartlett. La tabla A.10 da los valores críticos, $b_k(\alpha; n)$, para $\alpha = 0.01$ y 0.05; $k = 2, 3, \ldots, 10$; y valores seleccionados de n, desde 3 hasta 100.

Cuando los tamaños de las muestras son distintos, se rechaza la hipótesis nula con el nivel de significancia α si

$$b < b_k(\alpha; n_1, n_2, \ldots, n_k),$$

donde

$$b_k(\alpha; n_1, n_2, \ldots, n_k) \approx \frac{n_1 b_k(\alpha; n_1) + n_2 b_k(\alpha; n_2) + \cdots + n_k b_k(\alpha; n_k)}{N}.$$

Igual que antes, todas las $b_k(\alpha; n_i)$ para tamaños de muestra $n_1, n_2, \ldots, n_k$, se obtienen de la tabla A.10.

Ejemplo 13.3: Utilice la prueba de Bartlett para probar la hipótesis de que, con un nivel de significancia de 0.01, son iguales las varianzas poblacionales de los cuatro grupos de medicamentos del ejemplo 13.2.

Solución:

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$,
H_1: Las varianzas no son iguales.
$\alpha = 0.01$.

Región crítica: En relación con el ejemplo 13.2, tenemos que $n_1 = 20$, $n_2 = 9$, $n_3 = 9$, $n_4 = 7$, $N = 45$ y $k = 4$. Por lo tanto, se rechaza cuando

$$\begin{aligned} b &< b_4(0.01; 20, 9, 9, 7) \\ &\approx \frac{(20)(0.8586) + (9)(0.6892) + (9)(0.6892) + (7)(0.6045)}{45} \\ &= 0.7513. \end{aligned}$$

Cálculos: El primero es

$$s_1^2 = 662.862,\ s_2^2 = 2219.781,\ s_3^2 = 2168.434,\ s_4^2 = 946.032,$$

y después

$$\begin{aligned} s_p^2 &= \frac{(19)(662.862) + (8)(2219.781) + (8)(2168.434) + (6)(946.032)}{41} \\ &= 1301.861. \end{aligned}$$

Ahora,

$$b = \frac{[(662.862)^{19}(2219.781)^{8}(2168.434)^{8}(946.032)^{6}]^{1/41}}{1301.861} = 0.8557.$$

Decisión: no rechace la hipótesis y concluya que las varianzas poblacionales de los cuatro grupos de medicamentos no son significativamente distintas.

Aunque la prueba de Bartlett es la que se utiliza con mayor frecuencia para probar la homogeneidad de varianzas, se dispone de otros métodos. Uno que se debe a Cochran brinda un procedimiento de cálculo sencillo; aunque está limitado a situaciones en que los tamaños de las muestras son iguales. La **prueba de Cochran** es útil en particular para detectar si alguna de las varianzas es mucho mayor que las demás. El estadístico que se emplea es:

$$G = \frac{\text{más grande } S_i^2}{\sum_{i=1}^{k} S_i^2},$$

y se rechaza la hipótesis de igualdad de varianzas si $g > g_\alpha$, donde el valor de g_α se obtiene de la tabla A.11.

Para ilustrar la prueba de Cochran nos remitiremos otra vez a los datos de la tabla 13.1, sobre la absorción de humedad de los agregados para concreto. ¿Se justificó aceptar varianzas iguales cuando se realizó el análisis de varianza en el ejemplo 13.1? Se encontró que

$$s_1^2 = 12,134,\ s_2^2 = 2303,\ s_3^2 = 3594,\ s_4^2 = 3319,\ s_5^2 = 3455.$$

Por lo tanto,

$$g = \frac{12,134}{24,805} = 0.4892,$$

la cual no excede el valor de la tabla $g_{0.05} = 0.5065$. Entonces, se concluye que es razonable la suposición de que las varianzas son iguales.

Ejercicios

13.1 Demuestre que el error cuadrático medio,

$$s^2 = \frac{SSE}{k(n-1)}$$

es un estimador insesgado de σ^2 para el análisis de varianza en una clasificación de un solo factor.

13.2 Demuestre el teorema 13.2.

13.3 Están en consideración seis máquinas diferentes para utilizarlas en la manufactura de juntas de caucho. Las máquinas se comparan con respecto de la resistencia a la tensión del producto. Se emplea una muestra aleatoria de 4 juntas procedentes de cada máquina, para determinar si la resistencia media a la tensión varía de una máquina a otra. Las siguientes son las mediciones de esa resistencia en kilogramos por centímetro cuadrado $\times 10^{-1}$:

Máquina					
1	**2**	**3**	**4**	**5**	**6**
17.5	16.4	20.3	14.6	17.5	18.3
16.9	19.2	15.7	16.7	19.2	16.2
15.8	17.7	17.8	20.8	16.5	17.5
18.6	15.4	18.9	18.9	20.5	20.1

Lleve a cabo el análisis de varianza con un nivel de significancia de 0.05, e indique si las resistencias medias a la tensión difieren o no en forma significativa para las 6 máquinas.

13.4 Los datos de la tabla siguiente representan el número de horas de alivio que proporcionaron 5 marcas diferentes de comprimidos para el dolor de cabeza que se

suministraron a 25 sujetos que tenían fiebre de 38 °C o más. Realice el análisis de varianza y pruebe la hipótesis de que, con un nivel de significancia de 0.05, el número medio de horas de alivio que dieron los comprimidos es el mismo para las 5 marcas. Analice los resultados.

Comprimido				
A	**B**	**C**	**D**	**E**
5.2	9.1	3.2	2.4	7.1
4.7	7.1	5.8	3.4	6.6
8.1	8.2	2.2	4.1	9.3
6.2	6.0	3.1	1.0	4.2
3.0	9.1	7.2	4.0	7.6

13.5 En el artículo *Shelf-Space Strategy in Retailing*, que se publicó en *Proceedings: Southern Marketing Association*, se investigó en los supermercados el efecto que tenía la altura de los anaqueles sobre las ventas de alimento enlatado para perro. Se llevó a cabo un experimento en un supermercado pequeño durante un periodo de 8 días, para las ventas de una marca de alimento para perro conocida como *Arf*, y que implicaba tres niveles de altura de anaquel: a las rodillas, a la cintura y a los ojos. Cada día se cambiaba al azar, en tres ocasiones distintas, la altura del anaquel en la que estaba dicho alimento. Las secciones restantes de la góndola que contenía la marca dada se llenaban con una mezcla de marcas de comida canina, las cuales resultaban tanto familiares como desconocidas para los consumidores de esa área geográfica específica. Las ventas diarias, expresadas en cientos de dólares, del alimento *Arf* para las tres alturas de anaquel, fueron las siguientes:

Altura de anaquel		
A las rodillas	**A la cintura**	**A los ojos**
77	88	85
82	94	85
86	93	87
78	90	81
81	91	80
86	94	79
77	90	87
81	87	93

¿Existe una diferencia significativa en el promedio de ventas diarias de dicho alimento, con base en la altura del anaquel? Utilice un nivel de significancia de 0.01.

13.6 La inmovilización de los venados silvestres de cola blanca usando tranquilizantes da a los investigadores la oportunidad de estudiarlos de cerca y obtener información psicológica valiosa. En el estudio denominado *Influence of Physical Restraint and Restraint Facilitating Drugs on Blood Measurements of White-Tailed Deer and Other Selected Mammals*, realizado por el Instituto Politécnico y Universidad Estatal de Virginia Polytechnic Institute and State University, los biólogos de la vida silvestre probaron el tiempo del *derribamiento* (el periodo transcurrido entre la inyección y la inmovilidad) de tres sustancias tranquilizantes distintas. En este caso, la inmovilidad se define como el punto en que el animal ya no tiene control muscular suficiente para permanecer de pie. Se asignaron al azar 30 venados machos de cola blanca a cada uno de tres tratamientos. El grupo A recibió 5 miligramos de cloruro de sucinilcolina líquida (SCC); al grupo B se le suministraron 8 miligramos de SCC en polvo; y al grupo C, 200 miligramos de hidrocloruro de fenciclidina. A continuación se presentan los tiempos de derribamiento, en minutos. Haga un análisis de varianza con un nivel de significancia de 0.01, y determine si el tiempo promedio de derribamiento es el mismo o no para las tres sustancias.

Grupo		
A	**B**	**C**
11	10	4
5	7	4
14	16	6
7	7	3
10	7	5
7	5	6
23	10	8
4	10	3
11	6	7
11	12	3

13.7 Se ha demostrado que el fertilizante a base de fosfato de amonio de magnesio, $MgNH_4PO_4$, es un proveedor eficaz de los nutrientes necesarios para el crecimiento de las plantas. Los compuestos que suministra son muy solubles en agua, lo cual permite su aplicación directa sobre la superficie del suelo o que se mezcle con el sustrato del crecimiento durante su colocación en una maceta. Se efectuó un estudio denominado *Effect of Magnesium Ammonium Phosphate on Height of Chrysanthemus*, en la Universidad George Mason, para determinar el nivel óptimo posible de la fertilización, con base en la mejoría de la respuesta del crisantemo en cuanto a su crecimiento vertical. Se dividieron 40 semillas de crisantemo en 4 grupos de diez plantas cada uno. Se sembró cada una en una maceta similar que contenía un medio uniforme de crecimiento. Se agregó a cada grupo de plantas una concentración cada vez mayor de $MgNH_4PO_4$, medido en gramos por bushel. Se cultivaron durante cuatro semanas los cuatro grupos de plantas en condiciones uniformes en un invernadero. En la tabla que sigue se presentan los tratamientos y los cambios respectivos de sus alturas, medidas en centímetros:

Tratamiento			
50 g/bu	**100 g/bu**	**200 g/bu**	**400 g/bu**
13.2	16.0	7.8	21.0
12.4	12.6	14.4	14.8
12.8	14.8	20.0	19.1
17.2	13.0	15.8	15.8
13.0	14.0	17.0	18.0
14.0	23.6	27.0	26.0
14.2	14.0	19.6	21.1
21.6	17.0	18.0	22.0
15.0	22.2	20.2	25.0
20.0	24.4	23.2	18.2

Con un nivel de significancia de 0.05, ¿podría concluirse que concentraciones diferentes de $MgNH_4PO_4$ afectan la estatura promedio que alcanzan los crisantemos? ¿Qué cantidad del fertilizante parece ser la mejor?

13.8 Un estudio mide la tasa de *sorción* (ya sea *absorción* o *adsorción*) de tres tipos diferentes de solventes químicos orgánicos. Estos solventes se utilizan para limpiar partes industriales metálicas, y son desechos potencialmente riesgosos. Se probaron muestras independientes de solventes de cada tipo y se registraron sus tasas de sorción como porcentaje molar. [Véase McClave, Dictrich y Sincich (1997).]

Aromáticos		**Cloroalcalinos**		**Ésteres**		
1.06	0.95	1.58	1.12	0.29	0.43	0.06
0.79	0.65	1.45	0.91	0.06	0.51	0.09
0.82	1.15	0.57	0.83	0.44	0.10	0.17
0.89	1.12	1.16	0.43	0.55	0.53	0.17
1.05				0.61	0.34	0.60

¿Existe diferencia significativa en la tasa media de sorción de los tres solventes? Para obtener sus conclusiones emplee un valor *P*. ¿Qué solvente usaría?

13.9 La enzima mitocondrial NAPH:NAD transhidrogenasa, de la tenia de la rata común (*Hymenolepiasis diminuta*) cataliza el hidrógeno en transferencia de NADYH a NAD, y produce NADH. Se sabe que esta enzima desempeña un papel vital en el metabolismo anaerobio de la tenia, y recientemente se planteó la hipótesis de que podría servir como una bomba de intercambio de protones, es decir, para transferir protones a través de la membrana mitocondrial. El estudio *Effect of Various Substrate Concentrations on the Conformational Variation of the NADPH:NAD Transhydrogenase of Hymenolepiasis diminuta* llevado a cabo por la Universidad Estatal Bowling Green, se diseñó para evaluar la capacidad de dicha enzima para sufrir cambios en su conformación o su forma. Los cambios en la actividad específica de la enzima ocasionados por las variaciones en la concentración de NADP podrían interpretarse como un apoyo de la teoría del cambio de conformación. La enzima en cuestión se localiza en la membrana interior de las mitocondrias de la tenia. Se homogeneizaron las tenias, y se aisló la enzima mediante una serie de centrifugaciones. Después, se agregaron diferentes concentraciones de NADP a la solución de enzima aislada y la mezcla se incubó durante tres minutos en un baño de agua a 56 °C. Luego, se analizó la enzima con un espectrómetro de rayo dual y se calcularon los resultados siguientes, en términos de la actividad específica de la enzima, en nanomoles por minuto por miligramo de proteína:

Concentración de NADP (nm)				
0	**80**	**160**	**360**	
11.01	11.38	11.02	6.04	10.31
12.09	10.67	10.67	8.65	8.30
10.55	12.33	11.50	7.76	9.48
11.26	10.08	10.31	10.13	8.89
			9.36	

Pruebe la hipótesis de que la actividad específica promedio es la misma para las cuatro concentraciones, con un nivel de significancia de 0.01.

13.10 Para los datos del ejercicio 13.7, use la prueba de Bartlett para probar si las varianzas son iguales.

13.5 Comparaciones con un grado de libertad

El análisis de varianza en la clasificación de un solo factor, o experimento de un solo factor, como se le denomina con frecuencia, tan sólo indica si puede rechazarse o no la hipótesis de tratamiento igual. Por lo general, el experimentador preferiría efectuar un análisis más profundo. Por ejemplo, en el ejemplo 13.1, el rechazo de la hipótesis nula permite concluir que las medias no son iguales, pero aún no sabemos si hay diferencias entre los agregados. El ingeniero quizás intuya *a priori* que los agregados 1 y 2 deberían poseer propiedades similares de absorción, al igual que los agregados 3 y 5. Sin embargo, resulta de interés estudiar las diferencias entre los dos grupos. Así, parece apropiado probar las hipótesis

$$H_0\colon \mu_1 + \mu_2 - \mu_3 - \mu_5 = 0,$$
$$H_1\colon \mu_1 + \mu_2 - \mu_3 - \mu_5 \neq 0.$$

Se observa que la hipótesis es una función lineal de las medias poblacionales, en las cuales los coeficientes suman cero.

Definición 13.1:

Cualquier función lineal de la forma

$$\omega = \sum_{i=1}^{k} c_i \mu_i,$$

donde $\sum_{i=1}^{k} c_i = 0$, se llama **comparación** o **contraste** de las medias de los tratamientos.

Es frecuente que el experimentador realice comparaciones múltiples al probar la significancia de los contrastes de las medias de los tratamientos, es decir, al probar una hipótesis del tipo

---

Hipótesis para un contraste

$$H_0: \sum_{i=1}^{k} c_i \mu_i = 0,$$

$$H_1: \sum_{i=1}^{k} c_i \mu_i \neq 0,$$

donde $\sum_{i=1}^{k} c_i = 0$.

La prueba se efectúa al calcular, primero, un contraste similar de las medias de los tratamientos.

$$w = \sum_{i=1}^{k} c_i \bar{y}_{i\cdot}.$$

Como $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{k\cdot}$ son variables aleatorias independientes que tienen distribuciones normales con medias $\mu_1, \mu_2, \ldots, \mu_k$ y varianzas $\sigma_1^2/n_1, \sigma_2^2/n_2, \ldots, \sigma_k^2/n_k$, respectivamente, el teorema 7.11 nos garantiza que w es un valor de la variable aleatoria normal W con media

$$\mu_W = \sum_{i=1}^{k} c_i \mu_i$$

y varianza

$$\sigma_W^2 = \sigma^2 \sum_{i=1}^{k} \frac{c_i^2}{n_i}.$$

Por lo tanto, cuando H_0 es verdadera, $\mu w = 0$ y, según el ejemplo 7.5, el estadístico

$$\frac{W^2}{\sigma_W^2} = \frac{\left(\sum_{i=1}^{k} c_i \bar{Y}_{i\cdot}\right)^2}{\sigma^2 \sum_{i=1}^{k} (c_i^2/n_i)}$$

es una variable aleatoria con distribución chi-cuadrada con 1 grado de libertad. Nuestra hipótesis se prueba con un nivel de significancia de α calculando

Estadístico para probar un contraste

$$f = \frac{\left(\sum_{i=1}^{k} c_i \bar{y}_{i.}\right)^2}{s^2 \sum_{i=1}^{k} (c_i^2/n_i)} = \frac{\left[\sum_{i=1}^{k} (c_i Y_{i.}/n_i)\right]^2}{s^2 \sum_{i=1}^{k} (c_i^2/n_i)} = \frac{SSw}{s^2}.$$

Aquí, f es un valor de la variable aleatoria F que tiene distribución F con 1 y $N-k$ grados de libertad.

Cuando los tamaños de las muestras son iguales a n,

$$SSw = \frac{\left(\sum_{i=1}^{k} c_i Y_{i.}\right)^2}{n \sum_{i=1}^{k} c_i^2}.$$

La cantidad SSw, que se denomina **suma de cuadrados de los contrastes**, indica la porción de SSA que se explica por el contraste en cuestión.

Esta suma de cuadrados se empleará para probar la hipótesis de que el contraste

$$\sum_{i=1}^{k} c_i \mu_i = 0.$$

Con frecuencia es de interés probar contrastes múltiples, en particular, contrastes que son linealmente independientes u ortogonales. Como resultado, se vuelve necesaria la siguiente definición:

Definición 13.2: Se dice que los dos contrastes

$$\omega_1 = \sum_{i=1}^{k} b_i \mu_i \qquad \text{y} \qquad \omega_2 = \sum_{i=1}^{k} c_i \mu_i$$

son **ortogonales**, si $\sum_{i=1}^{k} b_i c_i / n_i = 0$ o, cuando las n_i son iguales a n, si

$$\sum_{i=1}^{k} b_i c_i = 0.$$

Si ω_1 y ω_2 son ortogonales, entonces las cantidades SSw_1 y SSw_2 son componentes de SSA, cada una con un solo grado de libertad. Es posible hacer la partición de la suma de los cuadrados de los tratamientos con $k-1$ grados de libertad, en un máximo de $k-1$ sumas independientes de cuadrados, de los contrastes con un grado de libertad, que satisfacen la identidad

$$SSA = SSw_1 + SSw_2 + \cdots + SSw_{k-1},$$

si los contrastes son ortogonales entre sí.

Ejemplo 13.4: En relación con el ejemplo 13.1, encuentre la suma de cuadrados de los contrastes que corresponden a los contrastes ortogonales

$$\omega_1 = \mu_1 + \mu_2 - \mu_3 - \mu_5, \qquad \omega_2 = \mu_1 + \mu_2 + \mu_3 - 4\mu_4 + \mu_5,$$

y efectúe las pruebas de significancia adecuadas. En este caso, tiene interés *a priori* comparar los dos grupos (1, 2) y (3, 5). Un contraste importante e independiente es la comparación entre el conjunto de agregados (1, 2, 3, 5) y el agregado 4.

Solución: Es evidente que los dos contrastes son ortogonales, puesto que

$$(1)(1) + (1)(1) + (-1)(1) + (0)(-4) + (-1)(1) = 0.$$

El segundo contraste indica una comparación entre los agregados (1, 2, 3 y 5) y el agregado 4. Podemos escribir dos contrastes adicionales ortogonales a los dos primeros, es decir:

$$\omega_3 = \mu_1 - \mu_2 \qquad \text{(agregado 1 contra agregado 2)},$$
$$\omega_4 = \mu_3 - \mu_5 \qquad \text{(agregado 3 contra agregado 5)}.$$

De los datos de la tabla 13.1, se tiene que

$$SSw_1 = \frac{(3320 + 3416 - 3663 - 3664)^2}{6[(1)^2 + (1)^2 + (-1)^2 + (-1)^]} = 14,553,$$
$$SSw_2 = \frac{[3320 + 3416 + 3663 + 3664 - 4(2791)]^2}{6[(1)^2 + (1)^2 + (1)^2 + (1)^2 + (-4)^2]} = 70,035.$$

En la tabla 13.5 se presenta un análisis de varianza más amplio. Se observa que las dos sumas de cuadrados de los contrastes intervienen en casi todas las sumas de cuadrados de los agregados. Existe una diferencia significativa entre las propiedades de absorción de los agregados, y el contraste ω_1 es significativo marginalmente. Sin embargo, el valor f de 14.12 para ω_2 es más significativo, y se rechaza la hipótesis

$$H_0\text{: } \mu_1 + \mu_2 + \mu_3 + \mu_5 = 4\mu_4.$$

Tabla 13.5: Análisis de varianza usando contrastes ortogonales

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media de los cuadrados**	***f* calculada**
Agregados	85,356	4	21,339	4.30
$(1,2)$ vs. $(3,5)$	14,553	1	14,533	2.93
$(1,2,3,5)$ vs. 4	70,035	1	70,035	14.12
Error	124,021	25	4,961	
Total	209,377	29		

Los contrastes ortogonales permiten al profesional hacer la partición de la variación del tratamiento en componentes independientes. Hay varias elecciones posibles al seleccionar los contrastes ortogonales, excepto para el último. Es normal que el experimentador tenga interés en hacer ciertos contrastes. Ése fue el caso en nuestro ejemplo, donde había consideraciones que sugerían *a priori* que los agregados (1, 2) y (3, 5)

constituían grupos distintos con propiedades diferentes de absorción, postulado que no se sostenía mucho con la prueba de significancia. Sin embargo, la segunda comparación apoya la conclusión de que el agregado 4 parecía "destacar" de los demás. En este caso, no era necesaria la partición completa de *SSA*, ya que dos de las cuatro comparaciones independientes posibles intervenían en la mayoría de variaciones de los tratamientos.

En la figura 13.3 se presenta un procedimiento *SAS* GLM, que muestra el conjunto completo de contrastes ortogonales. Observe que la suma de cuadrados de los cuatro contrastes se agrega a la suma de cuadrados de los agregados. Asimismo, los últimos dos contrastes (1 contra 2, 3 contra 5) revelan comparaciones insignificantes.

```
                          The GLM Procedure
Dependent Variable: moisture
                              Sum of
 Source             DF      Squares   Mean Square  F Value  Pr > F
 Model               4   85356.4667    21339.1167     4.30  0.0088
 Error              25  124020.3333     4960.8133
 Corrected Total    29  209376.8000

        R-Square     Coeff Var      Root MSE    moisture Mean
        0.407669      12.53703      70.43304         561.8000

 Source            DF     Type I SS     Mean Square    F Value    Pr > F
 aggregate          4   85356.46667     21339.11667       4.30    0.0088

 Source            DF   Type III SS     Mean Square    F Value    Pr > F
 aggregate          4   85356.46667     21339.11667       4.30    0.0088

 Contrast          DF   Contrast SS     Mean Square    F Value    Pr > F
 (1,2,3,5) vs. 4    1   70035.00833     70035.00833      14.12    0.0009
 (1,2) vs. (3,5)    1   14553.37500     14553.37500       2.93    0.0991
 1 vs. 2            1     768.00000       768.00000       0.15    0.6973
 3 vs. 5            1       0.08333         0.08333       0.00    0.9968
```

Figura 13.3: Un conjunto de procedimientos ortogonales.

13.6 Comparaciones múltiples

El análisis de varianza es un procedimiento poderoso para probar la homogeneidad de un conjunto de medias. No obstante, si se rechazara la hipótesis nula y se aceptara la alternativa que se planteó, acerca de que no todas las medias son iguales, aún no se sabría cuáles de las medias poblacionales son iguales y cuáles diferentes.

En la sección 13.5 se describe el uso de contrastes ortogonales con la finalidad de realizar comparaciones entre conjuntos de niveles de factores o tratamientos. El concepto de ortogonalidad permite al analista hacer pruebas que implican contrastes

independientes. Así, puede hacerse la partición de la variación entre los tratamientos, SSA, en componentes con solo un grado de libertad, y así atribuir las proporciones de esta variación a contrastes específicos. Sin embargo, hay situaciones en que el empleo de contrastes no es un enfoque apropiado. Es frecuente que sea de interés efectuar varias (quizá todas las que sea posible) **comparaciones por pares** entre los tratamientos. En realidad, una comparación por pares puede verse como un contraste simple, es decir, una prueba de

$$H_0: \ \mu_i - \mu_j = 0,$$
$$H_1: \ \mu_i - \mu_j \neq 0,$$

para toda $i \neq j$. Todas las comparaciones posibles por pares entre las medias pueden ser muy benéficas cuando no se conocen *a priori* contrastes complejos particulares. Por ejemplo, suponga que se desea probar las hipótesis siguientes, con los datos de los agregados de la tabla 13.1:

$$H_0: \ \mu_1 - \mu_5 = 0,$$
$$H_1: \ \mu_1 - \mu_5 \neq 0,$$

Se desarrolla la prueba usando una F o una t, o con el enfoque de los intervalos de confianza. Con la t, se tiene que

$$t = \frac{\bar{y}_{1.} - \bar{y}_{5.}}{s\sqrt{2/n}},$$

donde s es la raíz cuadrada del error cuadrático medio, y $n = 6$ es el tamaño de la muestra por tratamiento. En este caso

$$t = \frac{553.33 - 610.67}{\sqrt{4961}\sqrt{1/3}} = -1.41.$$

El valor P para la prueba t con 25 grados de libertad es 0.17. Así, no hay evidencia suficiente para rechazar H_0.

Relación entre *t* y *F*

En lo sucesivo, analizaremos el empleo de una prueba t agrupada, junto con los lineamientos que se estudiaron en el capítulo 10. La estimación de unión proviene del error cuadrático medio, con la finalidad de aprovechar los grados de libertad agrupados a través de las cinco muestras. Además, probamos un contraste. El lector debería observar que si el valor t está elevado al cuadrado, el resultado tiene la forma exacta del valor de f para la prueba del contraste que se examinó en la sección precedente. En efecto,

$$f = \frac{(\bar{y}_{1.} - \bar{y}_{5.})^2}{s^2(1/6 + 1/6)} = \frac{(553.33 - 610.67)^2}{4961(1/3)} = 1.988,$$

que es, por supuesto, t^2.

Enfoque del intervalo de confianza para una comparación por pares

Es fácil resolver el mismo problema de una comparación por pares (o contraste) usando el enfoque del intervalo de confianza. Es claro que si se calcula un intervalo

de confianza de $100(1-\alpha)\%$ sobre $\mu_1 - \mu_5$, se tiene que

$$\bar{y}_{1.} - \bar{y}_{5.} \pm t_{\alpha/2} s \sqrt{\frac{2}{6}},$$

donde $t_{\alpha/2}$ es el punto superior de $100(1-\alpha/2)\%$ de una distribución t con 25 grados de libertad (grados de libertad que provienen de s^2). Esta conexión inmediata entre las pruebas de hipótesis y los intervalos de confianza debería ser notoria a partir de los análisis que se hicieron en los capítulos 9 y 10. La prueba de un contraste simple $\mu_1 - \mu_5$ implica algo no más allá de observar si el intervalo de confianza cubre o no al cero. Al sustituir los números, se tiene lo siguiente como intervalo de confianza de 95%:

$$(553.33 - 610.67) \pm 2.060\sqrt{4961}\sqrt{\frac{1}{3}} = -57.34 \pm 83.77.$$

Así, como el intervalo de confianza cubre al cero, el contraste no es significativo. En otras palabras, no hay diferencia significativa entre las medias de los agregados 1 y 5.

Tasa de error por experimento

Se ha demostrado que un contraste simple (es decir, una comparación de dos medias) se puede hacer utilizando una prueba F, como se vio en la sección 13.5, una prueba t, calculando un intervalo de confianza sobre la diferencia entre las dos medias. Sin embargo, hay muchas dificultades cuando el analista intenta hacer varias o todas las comparaciones por pares posibles. Para el caso de k medias, habrá, desde luego, $r = k(k-1)/2$ comparaciones por pares posibles. Si se suponen comparaciones independientes, la **tasa de error por experimento** (es decir, la probabilidad del rechazo falso de al menos una de las hipótesis) está dada por $1 - (1-\alpha)^r$, donde α es la probabilidad seleccionada del error tipo I para una comparación específica. Es claro que esta medida del error tipo I del experimento sabio podría ser bastante grande. Por ejemplo, aun si sólo hubiera 6 comparaciones, digamos, en el caso de 4 medias y $\alpha = 0.05$, la tasa de error por experimento sería

$$1 - (0.95)^6 \approx 0.26.$$

Junto con la tarea de probar muchas comparaciones por pares, por lo general, hay la necesidad de hacer el contraste eficaz sobre una sola comparación más conservadora. Es decir, usando el enfoque del intervalo de confianza, los intervalos serían mucho más anchos que $\pm t_{\alpha/2} s\sqrt{2/n}$, que se emplea para el caso en que se realiza una sola comparación.

Prueba de Tukey

Hay varios métodos estándar para realizar comparaciones por pares que den credibilidad a la tasa del error tipo I. Aquí se analizarán e ilustrarán dos de ellos. El primero, denominado **procedimiento de Tukey**, permite la formación de intervalos de confianza de $(1-\alpha)100\%$ para todas las comparaciones por pares. El método se basa en la distribución del rango *studentizado*. El punto apropiado del percentil es una función de α, k y v = grados de libertad para s^2. En la tabla A.12 se presenta una lista de puntos porcentuales superiores adecuados para $\alpha = 0.05$. El método de Tukey de comparaciones por pares implica encontrar diferencias significativas entre las medias i y j $(i \neq j)$ si $|\bar{y}_{i.} - \bar{y}_{j.}|$ excede $q[\alpha, k, v]\sqrt{\frac{s^2}{n}}$.

El procedimiento de Tukey se ilustra con facilidad. Considere un ejemplo hipotético en el que se tienen 6 tratamientos, en un diseño completamente aleatorio de un solo factor, con 5 observaciones tomadas por tratamiento. Suponga que el error cuadrático medio tomado de la tabla del análisis de varianza es $s^2 = 2.45$ (24 grados de libertad). Las medias muestrales están en orden ascendente,

$\bar{y}_{2.}$	$\bar{y}_{5.}$	$\bar{y}_{1.}$	$\bar{y}_{3.}$	$\bar{y}_{6.}$	$\bar{y}_{4.}$
14.50	16.75	19.84	21.12	22.90	23.20.

Con $\alpha = 0.05$, el valor de $q(0.05, 6, 24) = 4.37$. Así, todas las diferencias absolutas tienen que compararse con

$$4.37\sqrt{\frac{2.45}{5}} = 3.059.$$

Como resultado, las siguientes representan medias que son diferentes en forma significativa, según el procedimiento de Tukey:

4 y 1, 4 y 5, 4 y 2, 6 y 1, 6 y 5,
6 y 2, 3 y 5, 3 y 2, 1 y 5, 1 y 2.

¿De dónde proviene el nivel α en la prueba de Tukey?

Se mencionó brevemente el concepto de **intervalos de confianza simultáneos** que se emplean para el procedimiento de Tukey. El lector obtendrá una perspectiva útil del concepto de comparaciones múltiples, si comprende lo que se quiere decir con intervalos de confianza simultáneos.

En el capítulo 9 vimos que si se calcula un intervalo de confianza sobre, por ejemplo, una media μ, entonces la probabilidad de que el intervalo cubra la media verdadera μ es 0.95. Sin embargo, como lo estudiamos para el caso de comparaciones múltiples, la probabilidad efectiva de interés está ligada con la tasa de error por experimento, y debe hacerse énfasis en que los intervalos de confianza del tipo $\bar{y}_{i.} - \bar{y}_{j.} \pm q[\alpha, k, v]s\sqrt{1/n}$ no son independientes, ya que todos implican s y muchos utilizan los mismos promedios, las $\bar{y}_{i.}$. A pesar de tales dificultades, si se utiliza la $q(0.05, k, v)$, el nivel de confianza simultáneo está 95% controlado. Lo mismo es cierto para $q(0.01, k, v)$, es decir, el nivel de confianza está 99% controlado. En el caso de $\alpha = 0.05$, hay una probabilidad de 0.05 de que se encuentre equivocadamente que al menos un par de mediciones son diferentes (rechazo falso de al menos una hipótesis). En el caso en que $\alpha = 0.01$, la probabilidad correspondiente será 0.01.

Prueba de Duncan

El segundo procedimiento que se estudiará se llama **procedimiento de Duncan** o **prueba de Duncan de rango múltiple**. Este procedimiento también se basa en el concepto general del rango studentizado. El rango de cualquier subconjunto de p medias muestrales debe superar cierto valor antes de que se encuentre que cualquiera de las p medias es diferente. Este valor recibe el nombre de **rango mínimo significativo** para las p medias, y se denota como R_p, donde

$$R_p = r_p\sqrt{\frac{s^2}{n}}.$$

Los valores de la cantidad r_p, llamados **rango mínimo significativo studentizado**, dependen del nivel de significancia deseado y del número de grados de libertad del error cuadrático medio. Estos valores se obtienen de la tabla A.13, para $p = 2, 3, \ldots, 10$ medias.

Para ilustrar el procedimiento de rango múltiple, consideremos el ejemplo hipotético en el cual se comparan 6 tratamientos con 5 observaciones por tratamiento. Éste es el mismo ejemplo que se empleó para ilustrar la prueba de Tukey. Se obtiene R_p multiplicando cada r_p por 0.70. Los resultados de estos cálculos se resumen como sigue:

p	2	3	4	5	6
r_p	2.919	3.066	3.160	3.226	3.276
R_p	2.043	2.146	2.212	2.258	2.293

Al comparar estos rangos menos significativos con las diferencias en medias ordenadas, se llega a las conclusiones siguientes:

1. Como $\bar{y}_{4.} - \bar{y}_{2.} = 8.70 > R_6 = 2.293$, se concluye que μ_4 y μ_2 son significativamente distintas.
2. Al comparar $\bar{y}_{4.} - \bar{y}_{5.}$ y $\bar{y}_{6.} - \bar{y}_{2.}$ con R_5, se concluye que μ_4 es significativamente mayor que μ_5, y μ_6 es significativamente mayor que μ_2.
3. Al comparar $\bar{y}_{4.} - \bar{y}_{1.}$, $\bar{y}_{6.} - \bar{y}_{5.}$ y $\bar{y}_{3.} - \bar{y}_{2.}$ con R_4, se concluye que cada diferencia es significativa.
4. Al comparar $\bar{y}_{4.} - \bar{y}_{3.}$, $\bar{y}_{6.} - \bar{y}_{1.}$, $\bar{y}_{3.} - \bar{y}_{5}$ y $\bar{y}_{1.} - \bar{y}_{2.}$ con R_3, se encuentra que todas las diferencias son significativas excepto $\mu_4 - \mu_3$. Por lo tanto, μ_3, μ_4 y μ_6 constituyen un subconjunto de medias homogéneas.
5. Al comparar $\bar{y}_{3.} - \bar{y}_{1.}$, $\bar{y}_{1.} - \bar{y}_{5.}$ y $\bar{y}_{5.} - \bar{y}_{2.}$ con R_2, se concluye que sólo μ_3 y μ_1 no son significativamente distintas.

Se acostumbra a resumir las conclusiones anteriores con el dibujo de una línea debajo de cualquier subconjunto de medias adyacentes que no sean significativamente distintas. Así, tenemos

$\bar{y}_{2.}$	$\bar{y}_{5.}$	$\bar{y}_{1.}$	$\bar{y}_{3.}$	$\bar{y}_{6.}$	$\bar{y}_{4.}$
14.50	16.75	19.84	21.12	22.90	23.20

En este caso, queda claro que los resultados con los procedimientos de Tukey y Duncan son muy similares. El procedimiento de Tukey no detectó ninguna diferencia entre 2 y 5; mientras que el de Duncan sí lo hizo.

13.7 Comparación de los tratamientos con un control

En muchos problemas científicos y de ingeniería, no nos interesa hacer inferencias acerca de todas las comparaciones posibles entre las medias de los tratamientos, del tipo $\mu_i - \mu_j$. En vez de ello, es frecuente que el experimento dicte la necesidad de comparar simultáneamente cada *tratamiento* con un *control*. Un procedimiento de prueba desarrollado por C. W. Dunnett determina diferencias significativas entre cada media de tratamiento y el control, con un solo nivel conjunto de significancia, α. Para ilustrar el procedimiento de Dunnet, se considerarán los datos experimentales de la tabla 13.6, para la clasificación de un solo factor, donde se estudió el efecto de

Tabla 13.6: Producto de una reacción

Control	**Catalizador 1**	**Catalizador 2**	**Catalizador 3**
50.7	54.1	52.7	51.2
51.5	53.8	53.9	50.8
49.2	53.1	57.0	49.7
53.1	52.5	54.1	48.0
52.7	54.0	52.5	47.2
$\bar{y}_{0.} = 51.44$	$\bar{y}_{1.} = 53.50$	$\bar{y}_{2.} = 54.04$	$\bar{y}_{3.} = 49.38$

tres catalizadores sobre el producto de una reacción. Como control se emplea un cuarto tratamiento, no catalizador.

En general, se desea probar las k hipótesis

$$\left.\begin{array}{l} H_0\colon\ \mu_0 = \mu_i \\ H_1\colon\ \mu_0 \neq \mu_i \end{array}\right\} \qquad i = 1, 2, \ldots, k,$$

donde μ_0 representa la respuesta media para la población de medidas en que se utiliza el control. Se espera que sigan siendo válidas las suposiciones habituales del análisis de varianza, como se mencionó en la sección 13.3. Para probar la hipótesis nula especificada con H_0 contra las alternativas bilaterales para una situación experimental donde existen k tratamientos, sin incluir el control, y n observaciones por tratamiento, primero calculamos los valores

$$d_i = \frac{\bar{y}_{i.} - \bar{y}_{0.}}{\sqrt{2s^2/n}}, \qquad i = 1, 2, \ldots, k.$$

Al igual que antes, la varianza muestral s^2 se obtiene a partir del error cuadrático medio del análisis de varianza. Ahora, la región crítica para rechazar H_0, con el nivel de significancia α, se establece con la desigualdad

$$|d_i| > d_{\alpha/2}(k, v),$$

donde v es el número de grados de libertad para el error cuadrático medio. Los valores de la cantidad $d_{\alpha/2}(k, v)$ para una prueba de dos colas, están dados en la tabla A.14 para $\alpha = 0.05$ y $\alpha = 0.01$, para diversos valores de k y v.

Ejemplo 13.5: Para los datos de la tabla 13.6, pruebe la hipótesis que compara cada catalizador con el control, usando alternativas bilaterales. Como nivel de significancia conjunto elija $\alpha = 0.05$.

Solución: El error cuadrático medio con 16 grados de libertad se obtiene de la tabla de análisis de varianza, con todos los $k + 1$ tratamientos. El error cuadrático medio está dado por

$$s^2 = \frac{36.812}{16} = 2.30075,$$

y

$$\sqrt{\frac{2s^2}{n}} = \sqrt{\frac{(2)(2.30075)}{5}} = 0.9593.$$

Entonces,

$$d_1 = \frac{53.50 - 51.44}{0.9593} = 2.147,$$
$$d_2 = \frac{54.04 - 51.44}{0.9593} = 2.710,$$
$$d_3 = \frac{49.38 - 51.44}{0.9593} = -2.147.$$

De la tabla A.14, el valor crítico para $\alpha = 0.05$ resulta ser

$$d_{0.025}\,(3,\,16) = 2.59.$$

Como $|d_1| < 2.59$ y $|d_3| < 2.59$, se concluye que tan sólo la respuesta media para el catalizador 2 es significativamente distinta de la respuesta media de la reacción que utiliza el control.

Muchas aplicaciones prácticas imponen la necesidad de una prueba de una cola para comparar tratamientos con un control. En efecto, si un farmacólogo desea comparar varias dosis de un medicamento con el efecto de reducir el nivel de colesterol, y su control consiste en no usar ninguna dosis, sería de interés determinar si cada una de éstas produce una reducción significativamente mayor que la del control. En la tabla A.15 se presentan los valores críticos de $d_\alpha(k,\, v)$ para alternativas de una cola.

Ejercicios

13.11 Considere los datos del ejercicio de repaso 13.58 de la página 568. Efectúe pruebas de significancia sobre los siguientes contrastes:

a) B contra A, C y D;

b) C contra A y D;

c) A contra D.

13.12 El Departamento de Alimentación y Nutrición Humanas, del Instituto Politécnico y Universidad Estatal de Virginia, realizó el estudio *Loss of Nitrogen Through Sweat by Preadolescent Boys Consuming Three Levels of Dietary Protein*, para determinar la pérdida de nitrógeno por transpiración con varios niveles dietéticos de proteínas. En el experimento se utilizaron 12 hombres preadolescentes cuyas edades iban de 7 años 8 meses a 9 años 8 meses, y a quienes se les juzgó estar saludables. Cada muchacho estuvo sujeto a una de tres dietas controladas en las cuales consumía 29, 54 u 84 gramos de proteínas por día. Los siguientes datos representan la pérdida de nitrógeno del cuerpo a través de la transpiración, en miligramos, recabados durante los dos días últimos del periodo de experimentación:

Nivel de proteínas		
29 Gramos	**54 Gramos**	**84 Gramos**
190	318	390
266	295	321
270	271	396
	438	399
	402	

a) Ejecute un análisis de varianza con un nivel de significancia de 0.05, para demostrar que las pérdidas medias de nitrógeno a través de la transpiración son diferentes con los distintos niveles de proteínas.

b) Emplee un contraste de un grado de libertad con $\alpha = 0.05$ para comparar la pérdida media de nitrógeno por la transpiración, en muchachos que consumen 29 gramos de proteínas por día contra quienes consumen 54 y 84 gramos.

13.13 El propósito del estudio *The Incorporation of a Chelating Agent into a Flame Retardant Finish of a Cotton Flannelette and the Evaluation of Selected Fabric Properties*, llevado a cabo por el Instituto Politécnico y Universidad Estatal de Virginia, fue evaluar el

uso de un agente quelante como parte del acabado retardante de fuego de la franela algodón, determinando sus efectos en la inflamabilidad después de lavar la tela en condiciones específicas. Se prepararon dos baños, uno con celulosa de carboximetilo y otro sin ella. Se lavaron 12 piezas de tela 5 veces en el baño I; y otras 12 en el mismo tipo de baño pero 10 veces. Esto se repitió con 24 piezas adicionales de tela en el baño II. Después de los lavados, se midieron las longitudes quemadas de la tela, así como los tiempos de combustión. Por conveniencia, se definieron los siguientes tratamientos:

Tratamiento 1: 5 lavados en el baño I,

Tratamiento 2: 5 lavados en el baño II,

Tratamiento 3: 10 lavados en el baño I,

Tratamiento 4: 10 lavados en el baño II.

Los registros del tiempo de combustión, en segundos, son los siguientes:

Tratamiento			
1	**2**	**3**	**4**
13.7	6.2	27.2	18.2
23.0	5.4	16.8	8.8
15.7	5.0	12.9	14.5
25.5	4.4	14.9	14.7
15.8	5.0	17.1	17.1
14.8	3.3	13.0	13.9
14.0	16.0	10.8	10.6
29.4	2.5	13.5	5.8
9.7	1.6	25.5	7.3
14.0	3.9	14.2	17.7
12.3	2.5	27.4	18.3
12.3	7.1	11.5	9.9

a) Efectúe un análisis de varianza con un nivel de significancia de 0.01, y determine si hay diferencias significativas entre las medias de los tratamientos.

b) Use contrastes de un solo grado de libertad con $\alpha = 0.01$ para comparar el tiempo medio de combustión del tratamiento 1 contra el tratamiento 2, y también del tratamiento 3 contra el 4.

13.14 Emplee la prueba de Tukey con un nivel de significancia de 0.05, para analizar las medias de 5 marcas distintas de los comprimidos para el dolor de cabeza del ejercicio 13.4, en la página 521.

13.15 Para los datos del ejercicio de repaso 13.58 de la página 568, lleve a cabo la prueba de Tukey con un nivel de significancia de 0.01, para determinar cuáles laboratorios difieren, en promedio, en sus análisis.

13.16 Se realizó una investigación para determinar la fuente de reducción del producto de cierto reactivo químico. Se sabía que la pérdida de producto ocurría en el licor madre, es decir, el material eliminado en la etapa de filtración. Se intuía que mezclas distintas del material original podrían ocasionar reducciones diferentes del producto en la etapa de licor madre. A continuación se presentan los resultados de la reducción porcentual para tres lotes de cada una de cuatro mezclas seleccionadas.

Mezcla			
1	**2**	**3**	**4**
25.6	25.2	20.8	31.6
24.3	28.6	26.7	29.8
27.9	24.7	22.2	34.3

a) Haga un análisis de varianza con nivel de significancia de $\alpha = 0.05$.

b) Utilice la prueba de Duncan de rango múltiple para determinar qué mezclas difieren.

c) Resuelva el inciso *b*) usando la prueba de Tukey.

13.17 En el estudio efectuado en el río Jackson, denominado *An Evaluation of the Removal Method for Estimating Benthic Populations and Diversity*, efectuado por el Instituto Politécnico y Universidad Estatal de Virginia, se emplearon 5 procedimientos distintos de muestreo para determinar el total de especies. Se seleccionaron 12 muestras al azar y los 5 procedimientos de muestreo se repitieron 4 veces. Se registraron los conteos de especies, como sigue:

Procedimiento de muestreo				
Disminución	**De Hess modificado**	**Surber**	**Remoción del sustrato, de Kicknet**	**Kicknet**
85	75	31	43	17
55	45	20	21	10
40	35	9	15	8
77	67	37	27	15

a) ¿Hay diferencia significativa en el conteo promedio de especies para los distintos procedimientos de muestreo? En su conclusión use un valor *P*.

b) Emplee una prueba de Tukey con $\alpha = 0.05$ para determinar cuáles son los procedimientos de muestreo que difieren.

13.18 Los datos siguientes son valores de presión (psi) en un resorte de torsión para valores distintos del ángulo entre las vueltas del resorte en posición libre.

Ángulo (°)					
67	**71**	**75**		**79**	**83**
83	84	86	87	89	90
85	85	87	87	90	92
	85	88	88	90	
	86	88	88	91	
	86	88	89		
	87	90			

Para el experimento, calcule un análisis de varianza de un solo factor, y dé su conclusión acerca del efecto que tiene el ángulo sobre la presión en el resorte. (C. R. Hicks, *Fundamental Concepts in the Design of Experiments*, Holt, Rinehart y Winston, Nueva York, 1973).

13.19 En el siguiente experimento de biología se emplearon 4 concentraciones de cierto producto químico para mejorar el crecimiento de cierto tipo de planta du-

rante el transcurso del tiempo. Se utilizaron cinco plantas con cada concentración, y se midió su crecimiento, en centímetros. Se obtuvieron los datos siguientes y también se aplicó un control (ausencia de producto químico)

	Concentración			
Control	**1**	**2**	**3**	**4**
6.8	8.2	7.7	6.9	5.9
7.3	8.7	8.4	5.8	6.1
6.3	9.4	8.6	7.2	6.9
6.9	9.2	8.1	6.8	5.7
7.1	8.6	8.0	7.4	6.1

Utilice una prueba bilateral de Dunnet con un nivel de significancia de 0.05 para comparar de manera simultánea las concentraciones con el control.

13.20 La tabla siguiente (A. Hald, *Statistical Theory with Engineering Applications*, John Wiley & Sons, New York, 1952) proporciona las resistencias a la tensión de desviaciones a partir de la 340, para conductores extraídos de nueve cables que deben usarse para una red de alto voltaje. Cada cable está constituido por 12 conductores. Se desea saber si las resistencias medias de los conductores en los nueve cables son las mismas. Si los cables son diferentes, ¿cuáles son los que difieren? En su análisis de varianza utilice un valor *P*.

Cable	**Resistencia a la tensión**											
1	5	−13	−5	−2	−10	−6	−5	0	−3	2	−7	−5
2	−11	−13	−8	8	−3	−12	−12	−10	5	−6	−12	−10
3	0	−10	−15	−12	−2	−8	−5	0	−4	−1	−5	−11
4	−12	4	2	10	−5	−8	−12	0	−5	−3	−3	0
5	7	1	5	0	10	6	5	2	0	−1	−10	−2
6	1	0	−5	−4	−1	0	2	5	1	−2	6	7
7	−1	0	2	1	−4	2	7	5	1	0	−4	2
8	−1	0	7	5	10	8	1	2	−3	6	0	5
9	2	6	7	8	15	11	−7	7	10	7	8	1

13.21 La información de salida de la figura 13.4 de la página 536 presenta la prueba de Duncan usando PROC GLM en *SAS*, para los datos de agregados del ejemplo 13.1. Saque conclusiones acerca de comparaciones por pares con el empleo de resultados de la prueba de Duncan.

13.22 La estructura financiera de una empresa consiste en la forma en que los activos de ésta se dividen entre propios y de deuda, y el apalancamiento financiero se refiere al porcentaje de activos financiados con endeudamiento. En el artículo *The Effect of Financial Leverage on Return*, Tai Ma, del Instituto Politécnico y Universidad Estatal de Virginia, afirma que es posible utilizar el apalancamiento financiero para incrementar la tasa de rendimiento sobre el capital. Dicho de otra manera, los accionistas pueden recibir rendimientos más elevados sobre el capital propio con la misma cantidad de inversión, si usan apalancamiento financiero. Los siguientes datos muestran las tasas de rendimiento sobre el capital con el uso de 3 niveles distintos de apalancamiento financiero, así como un nivel de control (deuda igual a cero) para 24 empresas seleccionadas al azar.

Apalancamiento financiero			
Control	**Bajo**	**Medio**	**Alto**
2.1	6.2	9.6	10.3
5.6	4.0	8.0	6.9
3.0	8.4	5.5	7.8
7.8	2.8	12.6	5.8
5.2	4.2	7.0	7.2
2.6	5.0	7.8	12.0

Fuente: Standard & Poor's *Machinery Industry Survey*, 1975.

a) Haga el análisis de varianza con un nivel de significancia de 0.05.

b) Use una prueba de Dunnet con un nivel de significancia de 0.01, para determinar si las tasas medias de rendimiento sobre el capital propio, con los niveles bajo, medio y alto de apalancamiento financiero, son mayores que con el nivel de control.

13.23 Se sospecha que la temperatura ambiente en que operan las baterías afecta su vida. Se probaron 30 baterías homogéneas, seis por cada una de cinco temperaturas, y los datos se presentan a continuación (vida activada, en segundos). Analice e interprete los datos. (C. R. Hicks, *Fundamental Concepts in Design of Experiments*, Holt, Rinehart y Winston, Nueva York, 1973.)

Temperatura (°C)				
0	**25**	**50**	**75**	**100**
55	60	70	72	65
55	61	72	72	66
57	60	72	72	60
54	60	68	70	64
54	60	77	68	65
56	60	77	69	65

13.24 Realice la prueba de Duncan para comparaciones por pares con los datos del ejercicio 13.8 de la página 523. Comente los resultados.

13.8 Comparación de un conjunto de tratamientos por bloques

En la sección 13.2 estudiamos la idea de formar bloques, es decir, de aislar conjuntos de unidades experimentales que fueran razonablemente homogéneas para asignarles tratamientos al azar. Ésta es una extensión del concepto de "formar pares" que se

```
                    The GLM Procedure
          Duncan's Multiple Range Test for moisture
NOTE: This test controls the Type I comparisonwise error rate,
      not the experimentwise error rate.
             Alpha                          0.05
             Error Degrees of Freedom         25
             Error Mean Square          4960.813
  Number of Means          2          3          4          5
  Critical Range       83.75      87.97      90.69      92.61

Means with the same letter are not significantly different.
        Duncan Grouping          Mean      N    aggregate
                         A        610.67      6    5
                         A
                         A        610.50      6    3
                         A
                         A        569.33      6    2
                         A
                         A        553.33      6    1

                         B        465.17      6    4
```

Figura 13.4: Salida de *SAS* para los ejercicios 13.21.

analizó en los capítulos 9 y 10, y se hace para reducir el error experimental, ya que las unidades en los mismos bloques tienen características que son más comunes que las unidades localizadas en diferentes bloques.

El lector no debería considerar a los bloques como un segundo factor, aunque esa sea una forma tentadora de visualizar el diseño. La realidad es que el factor principal (los tratamientos) aún lleva el peso mayor del experimento. Las unidades experimentales todavía son la fuente del error, igual que en el diseño completamente al azar. Con la formación de bloques tan sólo se trata a dichas unidades de manera más sistemática. De ese modo, se dice que la aleatoriedad tiene restricciones. Por ejemplo, para un experimento químico diseñado para determinar si hay una diferencia en la reacción media producida por cuatro catalizadores, las muestras de los materiales que tienen que probarse se extraen de los mismos lotes de materias primas, a la vez que se mantienen constantes otras condiciones como la temperatura y concentración de los reactivos. En este caso, la hora del día en que se efectúan los experimentos podría representar las unidades experimentales, y si el experimentador considera que es posible que haya un efecto leve del tiempo, haría aleatoria la asignación de los catalizadores a los experimentos, de manera que se contrarreste la tendencia posible. Este tipo de estrategia experimental es el **diseño completamente aleatorio**. Como otro ejemplo de dicho diseño, considere un experimento para comparar cuatro métodos para medir una propiedad física en particular de una sustancia fluida. Suponga que el proceso de muestreo es destructivo, es decir, que una vez que se ha medido una muestra de la sustancia usando un método, ya no puede medirse con ningún otro. Se decidió que con cada método habrían de tomarse 5 mediciones, por lo que se seleccionaron al azar 20 muestras de un lote grande y se utilizaron en el

experimento para comparar los cuatro dispositivos de medición. Las unidades experimentales son las muestras seleccionadas al azar. Cualquier variación de una muestra a otra aparecerá en la variación del error, según se mida con s^2 en el análisis.

¿Cuál es el propósito de formar bloques?

Si la variación debida a la heterogeneidad de unidades experimentales fuera tan grande que la sensibilidad de detectar diferencias en el tratamiento se redujera a un valor inflado de s^2, un plan mejor sería "bloquear" la variación debida a dichas unidades, para reducir así la variación externa a aquella considerada por bloques más pequeños o más homogéneos. Por ejemplo, suponga que en la ilustración anterior de los catalizadores, se supiera *a priori* que existe en definitiva un efecto significativo diario sobre el producto, y que es posible medir el producto para cuatro catalizadores en un día específico. En vez de asignar los 4 catalizadores a las 20 corridas de prueba completamente al azar, se eligen, por decir algo, 5 días y se prueba cada uno de los cuatro catalizadores en cada día, asignando al azar éstos a las corridas dentro de los días. De esta manera, se eliminaría del análisis la variación diaria y, en consecuencia, el error experimental, que aún incluye cualquier tendencia temporal *dentro de los días*, representa con más precisión la variación probabilística. Se hace referencia a cada día como un **bloque**.

La manera más directa de los diseños aleatorios de bloques es aquella donde se asigna al azar un tratamiento por vez a cada bloque. Un plan experimental así se denomina **diseño por bloques completamente aleatorio**, y cada bloque constituye una sola réplica de los tratamientos.

13.9 Diseños por bloques completamente aleatorios

Un plan clásico del diseño por bloques completamente aleatorio (BCA) con tres mediciones en cuatro bloques, es el siguiente:

Bloque 1	Bloque 2	Bloque 3	Bloque 4
t_2	t_1	t_3	t_2
t_1	t_3	t_2	t_1
t_3	t_2	t_1	t_3

Las t denotan la asignación de cada uno de tres tratamientos a los bloques. Por supuesto, la asignación verdadera de los tratamientos a las unidades dentro de los bloques se hace al azar. Una vez que ha finalizado el experimento, los datos se registran en el arreglo de 3×4 que se presenta a continuación:

Tratamiento	**Bloque:**	**1**	**2**	**3**	**4**
1		y_{11}	y_{12}	y_{13}	y_{14}
2		y_{21}	y_{22}	y_{23}	y_{24}
3		y_{31}	y_{32}	y_{33}	y_{34}

donde y_{11} representa la respuesta que se obtiene con usando el tratamiento 1 en el bloque 1, y_{12} es la respuesta por usar el tratamiento 1 en el bloque $2, \ldots,$ y y_{34} es la respuesta por emplear el tratamiento 3 en el bloque 4.

Ahora vamos a generalizar y a considerar el caso de k tratamientos asignados a b bloques. Los datos se resumen como se observa en el arreglo rectangular de $k \times b$ de la tabla 13.7. Se supondrá que las y_{ij}, $i = 1, 2, \ldots, b$, son valores de variables aleatorias independientes que tienen distribuciones normales con medias μ_{ij} y varianza común σ^2.

Tabla 13.7: Arreglo de $k \times b$ para el Diseño de BCA

	Bloque:							
Tratamiento	**1**	**2**	$\cdots$	j	$\cdots$	b	**Total**	**Media**
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1b}	$T_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2b}	$T_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{ib}	$T_{i.}$	$\bar{y}_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kb}	$T_{k.}$	$\bar{y}_{k.}$
Total	$T_{.1}$	$T_{.2}$	$\cdots$	$T_{.j}$	$\cdots$	$T_{.b}$	$T_{..}$	
Media	$\bar{y}_{.1}$	$\bar{y}_{.2}$	$\cdots$	$\bar{y}_{.j}$	$\cdots$	$\bar{y}_{.b}$		$\bar{y}_{..}$

Sea que μ_i represente el promedio (en vez del total) de las b medias poblacionales para el i-ésimo tratamiento. Es decir,

$$\mu_{i.} = \frac{1}{b}\sum_{j=1}^{b} \mu_{ij}.$$

De manera similar, el promedio de las medias poblacionales para el j-ésimo bloque, μ_j, está definido por

$$\mu_{.j} = \frac{1}{k}\sum_{i=1}^{k} \mu_{ij},$$

y el promedio de las bk medias poblacionales, μ, está definido por

$$\mu = \frac{1}{bk}\sum_{i=1}^{k}\sum_{j=1}^{b} \mu_{ij}.$$

Para determinar si parte de la variación de nuestras observaciones se debe a diferencias entre los tratamientos, se considera la prueba

Hipótesis de medias iguales de los tratamientos	H_0': $\mu_{1.} = \mu_{2.} = \cdots = \mu$, H_1': No todas las μ_i son iguales.

Modelo para el diseño BCA

Cada observación puede escribirse en la forma siguiente:

$$y_{ij} = \mu_{ij} + \epsilon_{ij},$$

donde ϵ_{ij} mide la desviación del valor observado y_{ij} de la media poblacional μ_{ij}. La forma preferida de esta ecuación se obtiene al sustituir

$$\mu_{ij} = \mu + \alpha_i + \beta_j,$$

donde α_i es, como antes, el efecto del i-ésimo tratamiento, y β_j es el efecto el j-ésimo bloque. Se supone que el tratamiento y los efectos de los bloques son aditivos. Por lo tanto, puede escribirse

$$\boldsymbol{y}_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}.$$

Observe que el modelo se parece al de clasificación de un solo factor; la diferencia esencial es la introducción del efecto de bloque β_j. El concepto básico se parece mucho al de la clasificación de un solo factor, excepto que en el análisis debe tenerse en cuenta el efecto adicional debido a los bloques, ya que ahora la variación *en dos direcciones* se controla de manera sistemática. Y si imponemos las restricciones de que

$$\sum_{i=1}^{k} \alpha_i = 0 \qquad \text{y} \qquad \sum_{j=1}^{b} \beta_j = 0,$$

entonces

$$\mu_{i.} = \frac{1}{b}\sum_{j=1}^{b}(\mu + \alpha_i + \beta_j) = \mu + \alpha_i$$

y

$$\mu_{.j} = \frac{1}{k}\sum_{i=1}^{k}(\mu + \alpha_i + \beta_j) = \mu + \beta_j.$$

La hipótesis nula de que las medias de los k tratamientos μ_{i} son iguales y, por ello, iguales a μ, ahora es **equivalente a probar las hipótesis:**

H_0': $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$,
H_1': Al menos una de las α_i no es igual a cero.

Cada una de las pruebas sobre los tratamientos se basará en comparar estimadores independientes de la varianza común poblacional σ^2. Dichos estimadores se obtendrán con el desglose de la suma total de cuadrados de los datos en tres componentes, usando la siguiente identidad:

Teorema 13.3:

Identidad de la suma de cuadrados

$$\sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = b\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 + k\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

La demostración se deja como ejercicio para el lector.

La identidad de la suma de cuadrados se representa simbólicamente con la ecuación

$$SST = SSA + SSB + SSE.$$

donde

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 \quad = \text{suma total de cuadrados,}$$

$$SSA = b\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 \quad = \text{suma de los cuadrados de los tratamientos,}$$

$$SSB = k\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 \quad = \text{suma de los cuadrados de los bloques,}$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \quad = \text{suma de los cuadrados de los errores.}$$

Al seguir el procedimiento bosquejado en el teorema 13.2, donde se interpreta la suma de cuadrados como funciones de las variables aleatorias independientes, Y_{11}, $Y_{12}, \ldots, Y_{kb}$, puede demostrarse que los valores esperados de las sumas de los cuadrados de los tratamientos, los bloques y los errores están dadas por

$$E(SSA) = (k-1)\sigma^2 + b\sum_{i=1}^{k}\alpha_i^2,$$

$$E(SSB) = (b-1)\sigma^2 + k\sum_{j=1}^{b}\beta_j^2,$$

$$E(SSE) = (b-1)(k-1)\sigma^2.$$

Como en el caso del problema de un solo factor, tenemos que el cuadrado de la media del tratamiento es

$$s_1^2 = \frac{SSA}{k-1}.$$

Si los efectos del tratamiento $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, entonces s_1^2 es un estimador insesgado de σ^2. Sin embargo, si los efectos de los tratamientos no son todos iguales a cero, se tiene que

Media cuadrática esperada del tratamiento

$$E\left(\frac{SSA}{k-1}\right) = \sigma^2 + \frac{b}{k-1}\sum_{i=1}^{k}\alpha_i^2,$$

y s_1^2 sobrestima σ^2. Un segundo estimador de σ^2, con base en $b-1$ grados de libertad, es

$$s_2^2 = \frac{SSB}{b-1}.$$

El estimador s_2^2 es uno insesgado de σ^2 si los efectos de los bloques $\beta_1 = \beta_2 = \cdots =$

$\beta_b = 0$. Si no todos los efectos de los bloques son iguales a cero, entonces,

$$E\left(\frac{SSB}{b-1}\right) = \sigma^2 + \frac{k}{b-1}\sum_{j=1}^{b}\beta_j^2,$$

y s_2^2 sobrestimará σ^2. Un tercer estimador de σ^2, con base en $(k-1)(b-1)$ grados de libertad e independiente de s_1^2 y s_2^2, es

$$s^2 = \frac{SSE}{(k-1)(b-1)},$$

que es insesgado sin que importe la verdad o falsedad de cualquier hipótesis nula.

Para probar la hipótesis nula de que los efectos de los tratamientos son iguales a cero, se calcula la razón $f_1 = s_1^2/s^2$, que es un valor de la variable aleatoria F_1 que tiene una distribución F con $k-1$ y $(k-1)(b-1)$ grados de libertad, cuando la hipótesis nula es verdadera. La hipótesis nula se rechaza con el nivel de significancia α cuando

$$f_1 > f_\alpha[k-1, (k-1)(b-1)].$$

En la práctica, primero calculamos *SST*, *SSA* y *SSB* y, después, utilizando la identidad de la suma de cuadrados, se obtiene *SSE* mediante una resta. Los grados de libertad asociados con *SSE*, por lo general, también se obtienen por sustracción; es decir,

$$(k-1)(b-1) = kb - 1 - (k-1) - (b-1).$$

En la tabla 13.8 se resumen los cálculos necesarios en un problema de análisis de varianza para un problema de diseño de bloques completamente aleatorio.

Tabla 13.8: Análisis de varianza para el diseño de bloques completamente aleatorios

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
Tratamientos	*SSA*	$k-1$	$s_1^2 = \frac{SSA}{k-1}$	$f_1 = \frac{s_1^2}{s^2}$
Bloques	*SSB*	$b-1$	$s_2^2 = \frac{SSB}{b-1}$	
Error	*SSE*	$(k-1)(b-1)$	$s^2 = \frac{SSE}{(k-1)(b-1)}$	
Total	*SST*	$kb-1$		

Ejemplo 13.6: Están en consideración cuatro máquinas diferentes, M_1, M_2 y M_3, para ensamblar un producto específico. Se decidió que para comparar las máquinas deben utilizarse 6 operadores distintos en un experimento por bloques completamente aleatorios. Las máquinas se asignan al azar a cada operador. La operación de las máquinas requiere destreza física, y se anticipa que habrá una diferencia en la velocidad con que los operadores trabajan con las máquinas (véase la tabla 13.9). Se registró la cantidad de tiempo (en segundos) que tomó ensamblar el producto:
Pruebe la hipótesis H_0 de que con un nivel de significancia de 0.05, las máquinas se desempeñan con la misma velocidad media.

Tabla 13.9: Tiempo, en segundos, para ensamblar el producto

	Operador						
Máquina	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
1	42.5	39.3	39.6	39.9	42.9	43.6	247.8
2	39.8	40.1	40.5	42.3	42.5	43.1	248.3
3	40.2	40.5	41.3	43.4	44.9	45.1	255.4
4	41.3	42.2	43.5	44.2	45.9	42.3	259.4
Total	163.8	162.1	164.9	169.8	176.2	174.1	1010.9

Solución: H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ (los efectos de las máquinas son iguales a cero)
H_1: al menos una de las α_i no es igual a cero.

Para realizar el análisis que aparece en la tabla 13.10 se emplean las fórmulas de la suma de cuadrados y los grados de libertad que se muestran en la página 540. El valor $f = 3.34$ es significativo con $P = 0.048$. Si se emplea $\alpha = 0.05$ como al menos una aproximación burda, se concluye que las máquinas no se desempeñan con la misma velocidad media.

Tabla 13.10: Análisis de varianza para los datos de la tabla 13.9

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Medias al cuadrado**	***f* calculada**
Máquinas	15.93	3	5.31	3.34
Operadores	42.09	5	8.42	
Error	23.84	15	1.59	
Total	81.86	23		

Comentarios adicionales acerca de la formación de bloques

En el capítulo 10 presentamos un procedimiento para comparar las medias cuando las observaciones estaban por *pares*. El procedimiento implicaba "restar" el efecto debido a la paridad homogénea para así trabajar con las diferencias. Éste es un caso especial de diseño por bloques completamente aleatorio con $k = 2$ tratamientos. Las n unidades homogéneas a las cuales fueron asignados los tratamientos adoptaban el papel de bloques.

Si hay heterogeneidad en las unidades experimentales, el experimentador no debería cometer el error de creer que siempre tiene ventajas reducir el error experimental al utilizar bloques pequeños homogéneos. La verdad es que hay circunstancias en las que no es deseable formar bloques. El propósito de reducir la varianza del error es incrementar la *sensibilidad* de la prueba para detectar diferencias en las medias de los tratamientos. Esto se refleja en la potencia del procedimiento de prueba. (En la sección 13.13 se analiza con mayor amplitud la potencia del procedimiento de prueba del análisis de varianza.) La potencia para detectar ciertas diferencias entre las medias de los tratamientos se incrementa con una disminución de la varianza del error. Sin embargo, la potencia también se ve afectada por los grados de libertad con los que se estima la varianza, y la formación de bloques reduce los grados de libertad

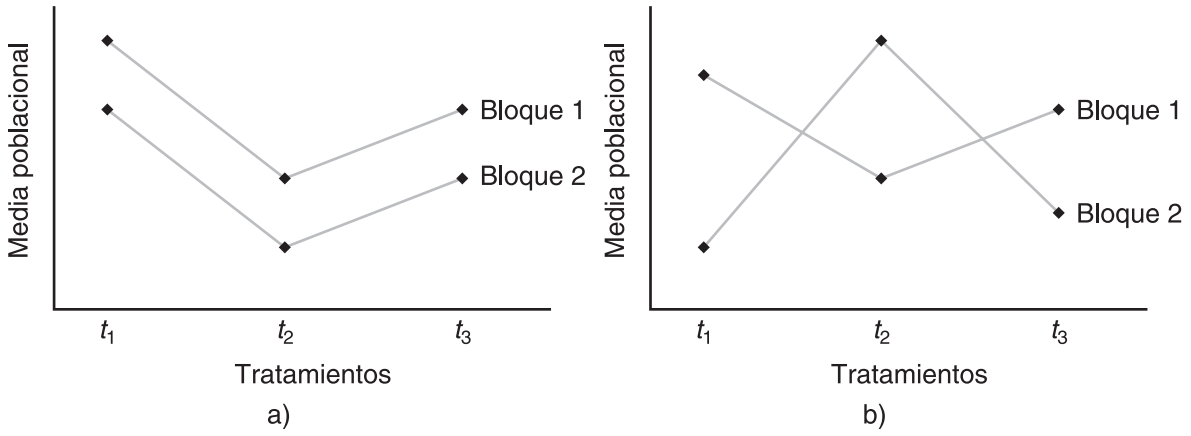


Figura 13.5: Medias poblacionales para a) resultados aditivos y b) efectos de la interacción.

de que se dispone para la clasificación de un solo factor, desde $k(b-1)$ hasta $(k-1)$ $(b-1)$. De modo que si no hubiera una reducción significativa de la varianza del error, con la formación de bloques podría perderse potencia.

Interacción entre bloques y tratamientos

Otra suposición importante que está implícita en la escritura del modelo, para un diseño por bloques completamente aleatorio, es que se supone que los efectos de los bloques y el tratamiento son aditivos. Esto equivale a decir que

$$\mu_{ij} - \mu_{ij'} = \mu_{i'j} - \mu_{i'j'} \quad \text{o bien,} \quad \mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'},$$

para cada valor de i, i', j y j'. Es decir, la diferencia entre las medias poblacionales para los bloques j y j' es la misma para cada tratamiento, y la diferencia entre las medias poblacionales para los tratamientos i e i' es la misma para cada bloque. Las líneas paralelas de la figura 13.5a) ilustran un conjunto de respuestas medias para las cuales los efectos del tratamiento y los bloques son aditivas; mientras que las líneas que se intersecan, en la figura 13.5b), muestran una situación en que dichos efectos **interactúan**. En relación con el ejemplo 13.6, si el operador 3 es en promedio 0.5 segundos más rápido que el operador 2 cuando utiliza la máquina 1, entonces el operador 3 en promedio será 0.5 más rápido que el operador 2 cuando se empleen las máquinas 2, 3 o 4. En muchos experimentos, no se cumple la suposición de aditividad y el análisis de la sección 13.9 llevaría a conclusiones erróneas. Por ejemplo, suponga que el operador 3 es 0.5 segundos más rápido, en promedio, que el operador 2 si emplea la máquina 1; pero 0.2 segundos más lento, en promedio, que el operador 2 si utiliza la máquina 2. En ese caso, los operadores y las máquinas estarían interactuando.

El análisis de la tabla 13.9 sugiere la posibilidad de la interacción. Ésta puede ser real o deberse al error experimental. El análisis del ejemplo 13.6 se basó en la suposición de que la interacción aparente se debía por completo al error experimental. Si la variabilidad total de nuestros datos se debiera en parte al efecto de la interacción, esa fuente de variación formaría parte de la suma de los cuadrados de los errores, **lo que ocasionaría que el error cuadrático medio sobrestimara σ^2**, con lo que

se incrementaría la probabilidad de cometer un error del tipo II. De hecho, se habría adoptado un modelo incorrecto. Si $(\alpha\beta)_{ij}$ denotara el efecto de la interacción del i-ésimo tratamiento y el j-ésimo bloque, un modelo más adecuado tendría la forma siguiente:

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij},$$

al que se impondrían las restricciones adicionales

$$\sum_{i=1}^{k}(\alpha\beta)_{ij} = \sum_{j=1}^{b}(\alpha\beta)_{ij} = 0.$$

Ahora, es fácil comprobar que

$$E\left[\frac{SSE}{(b-1)(k-1)}\right] = \sigma^2 + \frac{1}{(b-1)(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2.$$

Así, el error cuadrático medio es visto como un **estimador insesgado de σ^2 cuando se ha ignorado la interacción existente**. En este momento, parece necesario llegar a un procedimiento para detectar la interacción en aquellos casos en que se sospecha que exista. Tal procedimiento requiere que se disponga de un estimador insesgado e independiente de σ^2. Por desgracia, el diseño por bloques aleatorios no conduce a dicha prueba, a menos que se modifique el planteamiento inicial. En el capítulo 14 se estudia ese tema en forma extensa.

13.10 Métodos gráficos y comprobación del modelo

En varios capítulos de este libro se hace referencia a procedimientos gráficos para mostrar datos y resultados analíticos. En los primeros, se usaron gráficas de tallo y hojas y de caja y extensión, como ayudas visuales para resumir muestras. Se emplearon diagnósticos similares para entender mejor los datos de dos problemas de muestreo en los capítulos 9 y 10. En el capítulo 9 se introdujo el concepto de graficar los residuos (ordinarios y studentizados) para detectar trasgresiones de las suposiciones estándar. En los últimos años, gran parte de la atención dedicada al análisis de datos se ha centrado en los **métodos gráficos**. Al igual que en la regresión, el análisis de varianza lleva por sí mismo a gráficas que ayudan a resumir los datos, así como a detectar trasgresiones de los supuestos. Por ejemplo, una gráfica sencilla de las observaciones crudas alrededor de la media de cada tratamiento proporciona al analista una noción de la variabilidad entre las medias muestrales y dentro de las muestras. La figura 13.6 ilustra una de tales gráficas para los datos de agregados que se presentan en la tabla 13.1. De acuerdo con la apariencia de la gráfica se obtiene incluso la visión de cuáles agregados (si los hubiera) se apartan de los demás. Es evidente que el agregado 4 se aleja de los otros. También queda claro que los agregados 3 y 5 forman un grupo homogéneo, así como los agregados 1 y 2.

Como en el caso de la regresión, los residuos son de ayuda en el análisis de varianza para dar un diagnóstico sobre trasgresiones de los supuestos. Para formar los residuos, tan sólo necesitamos considerar el modelo del problema con un solo factor,

$$y_{ij} = \mu_i + \epsilon_{ij}.$$

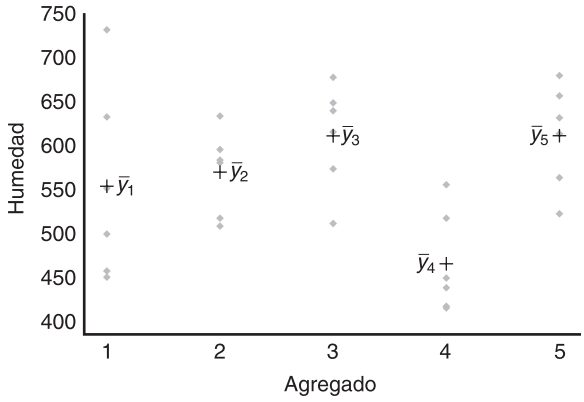


Figura 13.6: Gráfica de los datos alrededor de la media, con los datos de los agregados de la tabla 13.1.

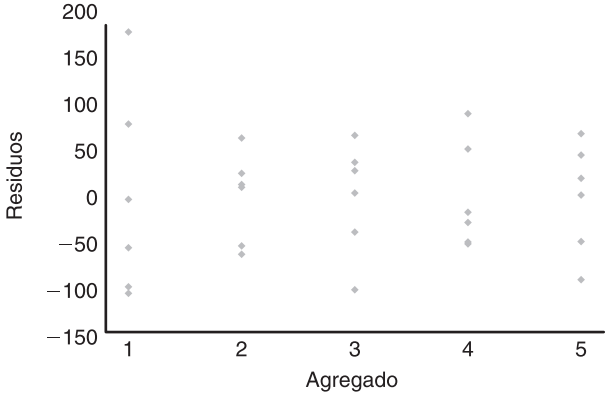


Figura 13.7: Gráficas de los residuos para cinco agregados, con los datos de la tabla 13.1.

Es inmediata la determinación de que la estimación de μ_i es $\bar{y}_{i.}$. Por lo tanto, el ij-ésimo residuo es $\bar{y}_{i.} - \bar{y}_{..}$. Esto se extiende fácilmente al modelo de bloques aleatorios por completo. Es instructivo graficar los residuos para cada agregado, con la finalidad de tener alguna perspectiva sobre la suposición de varianza homogénea. Dicha gráfica se muestra en la figura 13.7.

En ciertas situaciones, las tendencias de las gráficas revelan dificultades, en particular cuando la trasgresión de una suposición específica se manifiesta en la gráfica. En el caso de la figura 13.7, los residuos parecen indicar que las varianzas *dentro de los tratamientos* son razonablemente homogéneas, excepto la del agregado 1. Hay cierta evidencia gráfica de que la varianza del agregado 1 es más grande que las del resto.

¿Qué es un residuo para un diseño de BCA?

La formación de bloques completamente aleatorios es otra situación experimental en la cual una gráfica hace que el analista se sienta cómodo con una "imagen ideal" o

con la detección de dificultades. Hay que recordar que el modelo para bloques completamente aleatorios es

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \qquad i = 1, \ldots, k, \qquad j = 1, \ldots, b,$$

con las restricciones impuestas

$$\sum_{i=1}^{k} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0.$$

Para determinar qué es lo que en realidad constituye un residuo, considere que

$$\alpha_i = \mu_{i.} - \mu, \qquad \beta_j = \mu_{.j} - \mu$$

y que μ es estimada por $\bar{y}_{..}$, $\mu_{i.}$ es estimada por $\bar{y}_{i.}$, y que $\mu_{.j}$ es estimada por $\bar{y}_{.j}$. Como resultado, el *valor ajustado* o pronosticado, por $\hat{y}_{ij}$ está dado por

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..},$$

y, entonces, el residuo en la observación (i, j) está dado por

$$y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}.$$

Observe que $\hat{y}_{ij}$, el valor ajustado, es un estimador de la media μ_{ij}. Esto es consistente con la partición de la variabilidad dada en el teorema 13.3, en la que la suma de los errores al cuadrado es

$$SSE = \sum_{i} \sum_{j} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2.$$

Las técnicas visuales en la formación de bloques completamente aleatorios implican graficar los residuos por separado para cada tratamiento y bloque. Si la suposición de varianza homogénea se cumple, el analista debería esperar una variabilidad aproximadamente igual. El lector seguramente recordará que en el capítulo 12 se estudiaron gráficas de los residuos, en las cuales éstos se empleaban con el objetivo de detectar si el modelo era inadecuado. En el caso de los bloques aleatorios por completo, la falla seria del modelo podría estar relacionada con la suposición de aditividad (es decir, no hay interacción). Si no está presente ninguna interacción, debe surgir un patrón aleatorio.

Considere los datos del ejemplo 13.6, en los cuales los tratamientos son cuatro máquinas y los bloques son seis operadores. Las figuras 13.8 y 13.9 muestran las gráficas de los residuos para tratamientos separados y bloques separados. La figura 13.10 presenta una gráfica de los residuos contra los valores ajustados. La figura 13.8 revela que la varianza del error podría no ser la misma para todas las máquinas. Lo mismo sería válido para la varianza del error de cada uno de los seis operadores. Sin embargo, son dos residuos inusualmente grandes los que parecen producir la dificultad. La figura 13.10 revela una gráfica de residuos que dan evidencia razonable de un comportamiento aleatorio. Sin embargo, sobresalen los dos residuos grandes ya detectados.

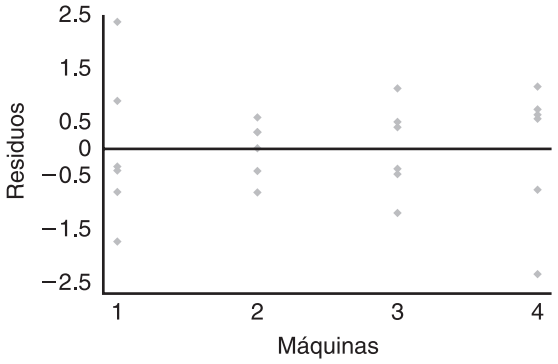


Figura 13.8: Gráfica de los residuos para las cuatro máquinas de los datos del ejemplo 13.6.

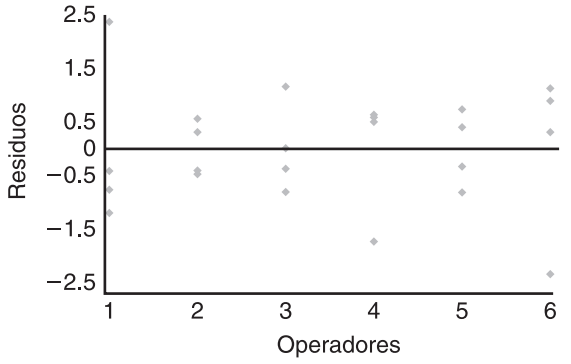


Figura 13.9: Gráfica de los residuos para los seis operadores para los datos del ejemplo 13.6.

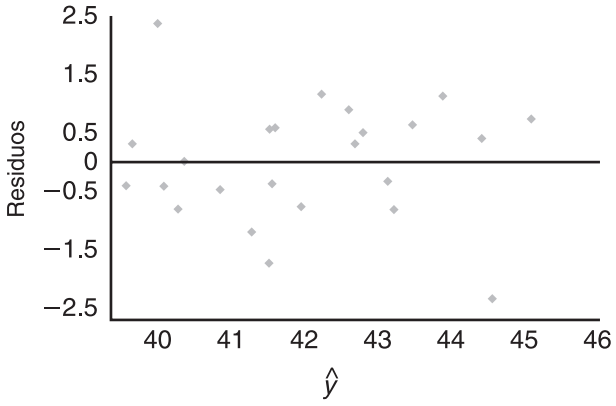


Figura 13.10: Los residuos graficados contra los valores ajustados para los datos del ejemplo 13.6.

13.11 Transformaciones de los datos en el análisis de varianza

En el capítulo 11 dimos considerable atención a la transformación de la respuesta y en situaciones para las que se ajustaba un modelo lineal a un conjunto de datos. Es evidente que se aplican los mismos conceptos a la regresión lineal múltiple, aunque ello no se analizó en el capítulo 12. En el estudio del modelado con regresión, se hizo énfasis en las transformaciones de y que producirían un modelo que se ajustaría mejor a los datos que aquel en el que la y entraba linealmente. Por ejemplo, si la estructura del "tiempo" es de naturaleza exponencial, entonces una transformación logarítmica de y linealiza la estructura, con lo que se espera tener más éxito cuando se use la respuesta transformada.

Si bien el propósito fundamental de transformar los datos que se ha perseguido hasta este momento ha sido mejorar el ajuste del modelo, hay otras razones para transformar o reexpresar la respuesta y, y muchas de ellas se relacionan con las suposiciones que se hacen (de las que depende la validez del análisis). Una suposición

muy importante en el análisis de varianza es la de la varianza homogénea que se estudió al inicio de la sección 13.4. Se supone una **varianza común σ^2**. Si la varianza difiriera mucho de un tratamiento a otro, y se llevara a cabo el ANOVA estándar que se estudia en este capítulo (y otros posteriores), los resultados estarían muy equivocados. En otras palabras, el análisis de varianza no es **robusto** respecto de la suposición de varianza homogénea. Como se ha dicho hasta el momento, se trata del motivo principal para graficar los residuos, según se analizó en la sección última y se ilustró con las figuras 13.8, 13.9 y 13.10. Dichas gráficas permiten detectar problemas debidos a una varianza no homogénea. Sin embargo, ¿qué hay que hacer al respecto? ¿Cómo se les enfrenta?

¿De dónde proviene la varianza no homogénea?

Con frecuencia, aunque no siempre, la varianza no homogénea en el ANOVA existe debido a la distribución de las respuestas. Ahora, por supuesto, se acepta la normalidad de la respuesta. Pero hay ciertas situaciones en las que se necesitan pruebas sobre las medias aun cuando la distribución de la respuesta sea una de esas distribuciones que no son la normal y que se estudiaron en los capítulos 5 y 6 (Poisson, logarítmica normal, exponencial y gamma, entre otras). Los problemas del tipo ANOVA existen con datos de conteo, tiempo de operación antes del fallo, etcétera.

En los capítulos 5 y 6 se demostró que, además del caso de la normal, la varianza de una distribución con frecuencia será función de la media, es decir, $\sigma_i^2 = g(\mu_i)$. Por ejemplo, en el caso de la distribución de Poisson, $Var(Y_i) = \mu_i = \sigma_i^2$ (es decir, la *varianza es igual a la media*). En el caso de la exponencial, $Var(Y_i) = \sigma_i^2 = \mu_i^2$ (es decir, la *varianza es igual al cuadrado de la media*). Para el caso de la logarítmica normal, una transformación logarítmica produce una distribución normal con varianza constante σ^2.

Para obtener la varianza de una función no lineal se utilizan los mismos conceptos empleados en el capítulo 4, como ayuda para determinar la naturaleza de la *transformación estabilizadora de la varianza* $g(y_i)$. Al recordar la expansión en series de Taylor de primer orden de $g(y_i)$ alrededor de $y_i = \mu_i$ cuando $g'(\mu_i) = \left[\frac{\partial g(y_i)}{\partial y_i}\right]_{y_i=\mu_i}$, la función de transformación $g(y)$ debe ser independiente de μ para que baste como la transformación estabilizadora de la varianza. De lo anterior

$$Var[g(y_i)] = [g'(\mu_i)]^2 \sigma_i^2.$$

Como resultado, $g(y_i)$ debe ser tal que $g'(\mu_i) \propto \frac{1}{\sigma}$. Así, si se sospecha que la respuesta tiene una distribución de Poisson, $\sigma_i = \mu_i^{1/2}$, de modo que $g'(\mu_i) \propto \frac{1}{\mu_i^{1/2}}$. Entonces, la transformación estabilizadora de la varianza se vuelve $g(y_i) = y_i^{1/2}$. Con esta ilustración y manipulaciones similares para las distribuciones exponencial y gamma, se obtiene lo siguiente

Distribución	Transformaciones estabilizadoras de la varianza
Poisson	$g(y) = y^{1/2}$
Exponencial	$g(y) = \ln y$
Gamma	$g(y) = \ln y$

13.12 Cuadrados latinos (opcional)

El diseño por bloques completamente aleatorios es muy eficaz para reducir el error experimental al eliminar una fuente de variación. Otro diseño que es particularmente útil para controlar dos fuentes de variación, al tiempo que se reduce el número requerido de combinaciones de tratamientos, se denomina **cuadrados latinos**. Suponga el lector que hay interés en los rendimientos de 4 variedades de trigo utilizando 4 fertilizantes durante un periodo de 4 años. El número total de combinaciones de tratamientos para un diseño por completo aleatorio sería 64. Al seleccionar el mismo número de categorías para los tres criterios de clasificación, podría elegirse un diseño de cuadrados latinos y realizar el análisis de varianza empleando los resultados de solo 16 combinaciones de tratamientos. Un cuadrado latino común, seleccionado al azar de todos los cuadrados de 4×4 posibles, es el siguiente:

	Columna			
Renglón	**1**	**2**	**3**	**4**
1	A	B	C	D
2	D	A	B	C
3	C	D	A	B
4	B	C	D	A

Las cuatro letras, A, B, C y D, representan las 4 variedades de trigo a que se alude como **tratamientos**. Los renglones y las columnas, representados por los 4 fertilizantes y los 4 años, respectivamente, son las dos fuentes de variación que se desea controlar. Ahora se observa que cada tratamiento ocurre exactamente una vez en cada renglón y cada columna. Con este arreglo balanceado, el análisis de varianza permite separar la variación debida a los distintos fertilizantes y los años diferentes, de la suma de errores al cuadrado, y con ello obtener una prueba más exacta para las diferencias en los rendimientos de las cuatro variedades de trigo. Cuando existe interacción entre cualesquiera fuentes de variación, los valores f dejan de ser válidos en el análisis de varianza. En ese caso, el diseño por cuadrados latinos sería inadecuado.

Generalización al cuadrado latino

Ahora generalizaremos y consideraremos un cuadrado latino de $r \times r$, donde y_{ijk} denota una observación en el i-ésimo renglón y en la j-ésima columna para la k-ésima letra. Observe que una vez que se especifican la i y la j para un cuadrado latino en particular, en forma automática se conoce la letra determinada por k. Por ejemplo, en el cuadrado latino de 4×4 anterior, cuando $i = 2$ y $j = 3$, se tiene que $k = \text{B}$. Entonces, k es función de i y j. Si α_i y β_j son los efectos del i-ésimo renglón y la j-ésima columna, τ_k el efecto del k-ésimo tratamiento, μ la media general, y ϵ_{ijk} el error aleatorio, entonces,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \epsilon_{ijk},$$

a la que se imponen las restricciones

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \tau_k = 0.$$

Igual que antes, se acepta que las y_{ijk} son valores de variables aleatorias independientes que tienen distribuciones normales con medias

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \tau_k$$

y varianza común σ^2. Las hipótesis a probar son las siguientes:

H_0: $\tau_1 = \tau_2 = \cdots = \tau_r = 0$,
H_1: Al menos una de las τ_i no es igual a cero.

Esta prueba se basará en la comparación de estimadores independientes de σ^2, obtenidos con la descomposición de la suma total de cuadrados de nuestros datos en cuatro componentes, usando la siguiente identidad. En el ejercicio 13.37 de la página 554 se pide al lector que dé la demostración.

Teorema 13.4: **Identidad de la suma de cuadrados**

$$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2 = r \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 + r \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$$
$$+ r \sum_k (\bar{y}_{..k} - \bar{y}_{...})^2 + \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2$$

Simbólicamente, la identidad de la suma de cuadrados se escribe así:

$$SST = SSR + SSC + SSTr + SSE,$$

donde SSR y SSC se denominan suma de cuadrados del **renglón** y suma de cuadrados de la **columna**, respectivamente; $SSTr$ recibe el nombre de suma de cuadrados del **tratamiento**, y SSE es la suma de cuadrados del **error**. Se hace la partición de los grados de libertad de acuerdo con la identidad

$$r^2 - 1 = (r-1) + (r-1) + (r-1) + (r-1)(r-2).$$

Al dividir cada una de las sumas de cuadrados del lado derecho de la identidad entre su número correspondiente de grados de libertad, se obtienen los cuatro estimadores independientes

$$s_1^2 = \frac{SSR}{r-1}, \qquad s_2^2 = \frac{SSC}{r-1}, \qquad s_3^2 = \frac{SSTr}{r-1}, \qquad s^2 = \frac{SSE}{(r-1)(r-2)}$$

de σ^2. Si se interpretan las sumas de cuadrados como funciones de variables aleatorias independientes, no es difícil comprobar que

$$E(S_1^2) = E\left[\frac{SSR}{r-1}\right] = \sigma^2 + \frac{r}{r-1}\sum_i \alpha_i^2,$$

$$E(S_2^2) = E\left[\frac{SSC}{r-1}\right] = \sigma^2 + \frac{r}{r-1}\sum_j \beta_j^2,$$

$$E(S_3^2) = E\left[\frac{SSTr}{r-1}\right] = \sigma^2 + \frac{r}{r-1}\sum_k \tau_k^2,$$

$$E(S^2) = E\left[\frac{SSE}{(r-1)(r-2)}\right] = \sigma^2.$$

El análisis de varianza (tabla 13.11) indica la prueba F adecuada para los tratamientos.

Tabla 13.11: Análisis de varianza para un cuadrado latino de $r \times r$

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	**f calculada**
Renglones	SSR	$r-1$	$s_1^2 = \frac{SSR}{r-1}$	
Columnas	SSC	$r-1$	$s_2^2 = \frac{SSC}{r-1}$	
Tratamientos	$SSTr$	$r-1$	$s_3^2 = \frac{SSTr}{r-1}$	$f = \frac{s_3^2}{s^2}$
Error	SSE	$(r-1)(r-2)$	$s^2 = \frac{SSE}{(r-1)(r-2)}$	
Total	SST	r^2-1		

Ejemplo 13.7: Para ilustrar el análisis de un cuadrado latino, volveremos a analizar el experimento en que las letras A, B, C y D representan 4 variedades de trigo; los renglones representan 4 fertilizantes distintos; y las columnas se refieren a 4 años diferentes. Los datos de la tabla 13.12 son las producciones de las cuatro variedades de trigo, medidas en kilogramos por parcela. Se supone que las distintas fuentes de variación no interactúan. Con un nivel de significancia de 0.05, pruebe la hipótesis H_0: No hay diferencia en las producciones promedio de las 4 variedades de trigo.

Tabla 13.12: Producciones de trigo (kilogramos por parcela)

Tratamiento con fertilizante	**1981**	**1982**	**1983**	**1984**
t_1	A: 70	B: 75	C: 68	D: 81
t_2	D: 66	A: 59	B: 55	C: 63
t_3	C: 59	D: 66	A: 39	B: 42
t_4	B: 41	C: 57	D: 39	D: 55

Solución:

H_0: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$,

H_1: Al menos una de las τ_i no es igual a cero.

Se emplean la suma de cuadrados y los grados de libertad que se muestran en la tabla 13.11. Las fórmulas de la suma de cuadrados aparecen en el teorema 13.4. En este caso, por supuesto, la tabla del análisis de varianza (tabla 13.13) debe reflejar la variabilidad que se debe al fertilizante, a los años y a los tipos de tratamiento. El valor $f = 2.02$ es sobre 3 y 6 grados de libertad. Es evidente que el valor p de aproximadamente 0.2 es muy grande como para concluir que las variedades de trigo afectan de manera significativa su producción.

Ejercicios

13.25 Demuestre que el cálculo de la fórmula de SSB para el análisis de varianza del diseño por bloques completamente aleatorio, es equivalente al término correspondiente en la identidad del teorema 13.3.

13.26 Para el diseño por bloques completamente aleatorios con k tratamientos

Tabla 13.13: Análisis de varianza para los datos de la tabla 13.12

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**	**Valor *P***
Fertilizante	1557	3	519.000		
Año	418	3	139.333		
Tratamientos	264	3	88.000	2.02	0.21
Error	261	6	43.500		
Total	2500	15			

y b bloques, demuestre que

$$E(SSB) = (b-1)\sigma^2 + k\sum_{j=1}^{b}\beta_j^2.$$

13.27 Se utilizaron cuatro clases de fertilizante f_1, f_2, f_3 y f_4, para estudiar el rendimiento en el cultivo de frijol. El suelo se dividió en 3 bloques, cada uno de los cuales contiene 4 parcelas homogéneas. A continuación se presentan los rendimientos en kilogramos por parcela, así como los tratamientos correspondientes:

Bloque 1	Bloque 2	Bloque 3
$f_1 = 42.7$	$f_3 = 50.9$	$f_4 = 51.1$
$f_3 = 48.5$	$f_1 = 50.0$	$f_2 = 46.3$
$f_4 = 32.8$	$f_2 = 38.0$	$f_1 = 51.9$
$f_2 = 39.3$	$f_4 = 40.2$	$f_3 = 53.5$

a) Realice un análisis de varianza con un nivel de significancia de 0.05, y utilice el modelo de bloques aleatorios por completo.

b) Emplee contrastes de un solo grado de libertad y un nivel de significancia de 0.01, para comparar los fertilizantes (f_1, f_3) contra (f_2, f_4), y f_1 contra f_3. Saque conclusiones.

13.28 Se comparan tres variedades de patatas en cuanto a su rendimiento. El experimento se efectuó con la asignación aleatoria de cada variedad a 3 parcelas de igual tamaño, en 4 ubicaciones diferentes. Se registraron los siguientes rendimientos para las variedades A, B y C, en 100 kilogramos por parcela:

Ubicación 1	Ubicación 2	Ubicación 3	Ubicación 4
B : 13	C : 21	C : 9	A : 11
A : 18	A : 20	B : 12	C : 10
C : 12	B : 23	A : 14	B : 17

Realice un análisis de varianza por bloques aleatorios con el objetivo de probar la hipótesis de que no hay diferencia en los rendimientos de las tres variedades de patatas. Utilice un nivel de significancia de 0.05. Obtenga sus conclusiones.

13.29 Los siguientes datos son los porcentajes de aditivos externos, medidos por 5 analistas, de 3 marcas distintas de mermelada de fresas, A, B y C.

Analista 1	Analista 2	Analista 3	Analista 4	Analista 5
B: 2.7	C: 7.5	B: 2.8	A: 1.7	C: 8.1
C: 3.6	A: 1.6	A: 2.7	B: 1.9	A: 2.0
A: 3.8	B: 5.2	C: 6.4	C: 2.6	B: 4.8

Ejecute el análisis de varianza y pruebe la hipótesis, con un nivel de significancia de 0.05, de que el porcentaje de aditivos externos es el mismo para las tres marcas de mermelada. ¿Cuál de ellas parece tener menos aditivos?

13.30 Los siguientes datos representan las calificaciones finales obtenidas por 5 estudiantes en matemáticas, inglés, francés y biología:

	Materia			
Estudiante	**Matemáticas**	**Inglés**	**Francés**	**Biología**
1	68	57	73	61
2	83	94	91	86
3	72	81	63	59
4	55	73	77	66
5	92	68	75	87

Pruebe la hipótesis de que los cursos tienen la misma dificultad. En las conclusiones use un valor P y analice lo que descubra.

13.31 En el estudio *The Periphyton of the South River, Virginia: Mercury Concentration, Productivity, and Autotropic Index Studies*, efectuado por el Departamento de Ciencias e Ingeniería Ambientales, del Instituto Politécnico y Universidad Estatal de Virginia, se midió la concentración total de mercurio en sólidos totales en perifitón en seis estaciones distintas y en seis días diferentes. Se registraron los datos siguientes:

	Estación					
Fecha	***CA***	***CB***	***E*1**	***E*2**	***E*3**	***E*4**
8 de abril	0.45	3.24	1.33	2.04	3.93	5.93
23 de junio	0.10	0.10	0.99	4.31	9.92	6.49
1 de julio	0.25	0.25	1.65	3.13	7.39	4.43
8 de julio	0.09	0.06	0.92	3.66	7.88	6.24
15 de julio	0.15	0.16	2.17	3.50	8.82	5.39
23 de julio	0.17	0.39	4.30	2.91	5.50	4.29

Determine si entre las estaciones la media del contenido de mercurio es significativamente distinta. Use un valor P y analice sus hallazgos.

13.32 Las instalaciones para generar energía nuclear producen gran cantidad de calor que, en general, se descarga a cuerpos de agua. Ese calor eleva la temperatura del líquido, lo cual da como resultado una mayor concentración de clorofila *a* que, a la vez, alarga la temporada de crecimiento. Para estudiar este efecto, se tomaron muestras de agua en forma mensual en 3 estaciones, durante un periodo de 12 meses. La estación A es la que se ubica más cerca de una descarga potencial de agua caliente, la estación C es la más lejana de la descarga, y la estación B se encuentra entre las estaciones A y C. Se registraron las siguientes concentraciones de clorofila *a*.

	Estación		
Mes	**A**	**B**	**C**
Enero	9.867	3.723	4.410
Febrero	14.035	8.416	11.100
Marzo	10.700	20.723	4.470
Abril	13.853	9.168	8.010
Mayo	7.067	4.778	34.080
Junio	11.670	9.145	8.990
Julio	7.357	8.463	3.350
Agosto	3.358	4.086	4.500
Septiembre	4.210	4.233	6.830
Octubre	3.630	2.320	5.800
Noviembre	2.953	3.843	3.480
Diciembre	2.640	3.610	3.020

Realice un análisis de varianza y pruebe la hipótesis de que, con un nivel de significancia de 0.05, no hay diferencia en las concentraciones medias de clorofila *a* en las 3 estaciones.

13.33 En un estudio realizado por el Departamento de Salud y Educación Física del Instituto Politécnico y Universidad Estatal de Virginia, se asignaron 3 dietas durante 3 días a cada uno de 6 sujetos, con diseño por bloques aleatorio. Los sujetos, que desempeñan el papel de bloques, recibieron las siguientes 3 dietas, en orden aleatorio:

Dieta 1: grasas mixtas y carbohidratos,
Dieta 2: muchas grasas,
Dieta 3: muchos carbohidratos.

Al terminar el periodo de tres días, se puso a cada sujeto en una banda caminadora y se midió el tiempo, en segundos, en el que quedaban exhaustos. Se registraron los siguientes datos:

		Sujeto					
		1	**2**	**3**	**4**	**5**	**6**
Dieta	1	84	35	91	57	56	45
	2	91	48	71	45	61	61
	3	122	53	110	71	91	122

Efectúe un análisis de varianza en el que aparte la dieta, a los sujetos y la suma de errores al cuadrado. Utilice un valor P para determinar si existe diferencia significativa entre las dietas.

13.34 El personal forestal utiliza arsénico orgánico como arboricida. Un problema grande de salud lo constituye la cantidad de arsénico que ingresa al cuerpo cuando se le expone a dicho arboricida. Es importante que la cantidad de exposición se determine rápido, de manera que pueda retirarse del trabajo a los empleados con niveles elevados de arsénico. En un experimento descrito en el artículo "A Rapid Method for the Determination of Arsenic Concentrations in Urine at Field Locations", publicado en *Amer. Ind. Hyg. Assoc. J.* (vol. 37, 1976), especímenes de orina procedentes de 4 personas del servicio forestal fueron divididos por igual en tres muestras, de manera que pudiera analizarse el arsénico en cada individuo en un laboratorio universitario, por un químico que utilizaba un sistema portátil, y por un empleado forestal que había recibido una capacitación breve. Se registraron los siguientes niveles de arsénico, en partes por millón:

	Analista		
Individuo	**Empleado**	**Químico**	**Laboratorio**
1	0.05	0.05	0.04
2	0.05	0.05	0.04
3	0.04	0.04	0.03
4	0.15	0.17	0.10

Realice un análisis de varianza y pruebe la hipótesis de que con los tres métodos de análisis no hay diferencia en los niveles de arsénico, con un nivel de significancia de 0.05.

13.35 Los científicos del Departamento de Patología Vegetal en el Tecnológico de Virginia Tech, realizaron un experimento en el que se aplicaron 5 tratamientos diferentes a 6 localidades distintas de un huerto de manzanas, para determinar si había diferencias significativas en el crecimiento según el tratamiento. Los tratamientos 1 a 4 representan distintos herbicidas, y el 5 es un control. El periodo de crecimiento fue de mayo a noviembre de 1982, y el crecimiento nuevo, medido en centímetros, para muestras seleccionadas de 6 ubicaciones en el huerto, se registraron como sigue:

	Ubicaciones					
Tratamiento	**1**	**2**	**3**	**4**	**5**	**6**
1	455	72	61	215	695	501
2	622	82	444	170	437	134
3	695	56	50	443	701	373
4	607	650	493	257	490	262
5	388	263	185	103	518	622

Lleve a cabo un análisis de varianza que separe el tratamiento, la ubicación y la suma de errores al cuadrado. Determine si hay diferencias significativas entre las medias de los tratamientos. Obtenga un valor P.

13.36 En el artículo "Self-Control and Therapist Control in the Behavioral Treatment of Overweight

Women", publicado en *Behavioral Research and Therapy* (vol. 10, 1972), se estudiaron dos tratamientos de reducción y otro de control, para observar sus efectos en el cambio del peso en mujeres obesas. Los dos tratamientos reductores involucrados fueron, respectivamente, un programa autoinducido de reducción de peso, y otro controlado por el terapeuta. Se asignó a cada uno de 10 sujetos a los tres programas de tratamiento en orden al azar, y se midió la pérdida de peso. Se registraron los siguientes cambios en el peso:

	Tratamiento		
Sujeto	**Control**	**Autoinducido**	**Con terapeuta**
1	1.00	−2.25	−10.50
2	3.75	−6.00	−13.50
3	0.00	−2.00	0.75
4	−0.25	−1.50	−4.50
5	−2.25	−3.25	−6.00
6	−1.00	−1.50	4.00
7	−1.00	−10.75	−12.25
8	3.75	−0.75	−2.75
9	1.50	0.00	−6.75
10	0.50	−3.75	−7.00

Realice un análisis de varianza y pruebe la hipótesis, con un nivel de significancia de 0.01, de que no hay diferencia en la media de las pérdidas de peso con los 3 tratamientos. ¿Qué tratamiento fue el mejor?

13.37 Compruebe la identidad de la suma de cuadrados del teorema 13.4, de la página 550.

13.38 Para el diseño con el cuadrado latino de r X r, demuestre que

$$E(SSTr) = (r-1)\sigma^2 + r\sum_k \tau_k^2.$$

13.39 El departamento de matemáticas de una universidad grande quiere evaluar las habilidades didácticas de 4 profesores. Para eliminar cualesquiera efectos debidos a los horarios y cursos distintos de matemáticas a lo largo del día, se decidió realizar un experimento utilizando el diseño del cuadrado latino, en el que las letras A, B, C y D representaban a los 4 diferentes profesores. Cada uno de ellos enseñó una parte de cada uno de cuatro cursos programados en 4 horarios distintos del día. Los datos siguientes muestran las calificaciones asignadas a los maestros por 16 estudiantes de capacidad aproximadamente igual. Utilice un nivel de significancia de 0.05 para probar la hipótesis de que los distintos profesores no tienen ningún efecto en las calificaciones.

	Curso			
Horario	**Álgebra**	**Geometría**	**Estadística**	**Cálculo**
1	A: 84	B: 79	C: 63	D: 97
2	B: 91	C: 82	D: 80	A: 93
3	C: 59	D: 70	A: 77	B: 80
4	D: 75	A: 91	B: 75	C: 68

13.40 Una empresa de manufactura desea investigar los efectos de 5 aditivos para el color en el tiempo de preparación de una nueva mezcla de concreto. Se esperan variaciones en los tiempos de preparación debido a los cambios diarios en temperatura y humedad, así como a los distintos trabajadores que preparan los moldes de prueba. Para eliminar esas fuentes extrañas de variación, se diseñó un cuadrado latino de 5 × 5, en el cual las letras A, B, C, D y E representan los 5 aditivos. En la tabla que sigue se presentan los tiempos de preparación, en horas, para los 25 moldes.

	Día				
Trabajador	**1**	**2**	**3**	**4**	**5**
1	D: 10.7	E: 10.3	B: 11.2	A: 10.9	C: 10.5
2	E: 11.3	C: 10.5	D: 12.0	B: 11.5	A: 10.3
3	A: 11.8	B: 10.9	C: 10.5	D: 11.3	E: 7.5
4	B: 14.1	A: 11.6	E: 11.0	C: 11.7	D: 11.5
5	C: 14.5	D: 11.5	A: 11.5	E: 12.7	B: 10.9

Con un nivel de significancia de 0.05, ¿es posible decir que los aditivos para el color no tienen efecto alguno en el tiempo de preparación de la mezcla de concreto?

13.41 En el libro *Design of Experiments for the Quality Improvement*, publicado por la Japanese Standards Association (1989), se describe un estudio sobre la cantidad de tinta que se requiere para obtener el mejor color para cierto tipo de tela. Se administraron en dos plantas diferentes las tres cantidades de tinta: $\frac{1}{3}$% wof (es decir, $\frac{1}{3}$% del peso de la tela), 1% wof y 3% wof. Después se observó la densidad del color de una tela cuatro veces para cada nivel de tinta aplicada en cada planta.

	Cantidad de tinta					
	1/3%		**1%**		**3%**	
Planta 1	5.2	6.0	12.3	10.5	22.4	17.8
	5.9	5.9	12.4	10.9	22.5	18.4
Planta 2	6.5	5.5	14.5	11.8	29.0	23.2
	6.4	5.9	16.0	13.6	29.7	24.0

Con un nivel de significancia de 0.05, realice un análisis de varianza para probar la hipótesis de que no hay diferencia en la densidad de color de una tela para los tres niveles de tinta. Considere a las plantas como bloques.

13.42 Se realizó un experimento con la finalidad de comparar tres tipos de materiales para recubrir alambres de cobre. El propósito del recubrimiento consiste en eliminar los "defectos" del alambre. Se asignaron al azar 10 especímenes distintos de cinco milímetros de longitud, para que recibieran cada proceso de recubrimiento, y los 30 especímenes se sujetaron a cierto tipo de uso abrasivo. Se midió el número de defectos en cada uno y se obtuvieron los siguientes resultados:

Material												
1				**2**				**3**			
6	8	4	5	3	3	5	4	12	8	7	14
7	7	9	6	2	4	4	5	18	6	7	18
7	8			4	3			8	5		

Suponga que se acepta que es aplicable un proceso de Poisson, por lo que el modelo es $Y_{ij} = \mu_i + \epsilon_{ij}$, donde μ_i es la media de la distribución de Poisson, y $\sigma_{Yij} = \mu_i$.

a) Realice tanto una transformación apropiada de los datos como un análisis de varianza.

b) Determine si hay evidencia suficiente o no para preferir un material de recubrimiento sobre los demás. Muestre cualesquiera hallazgos que sugieran una conclusión.

c) Haga una gráfica de los residuos y coméntela.

d) Mencione el propósito de la transformación de los datos.

e) ¿Qué otra suposición se hace en este caso, la cual quizá no se satisfaga por completo en la transformación?

f) Comente el inciso *e*) después de elaborar una gráfica de probabilidad normal sobre los residuos.

13.13 Modelos de efectos aleatorios

A lo largo de este capítulo estudiamos los procedimientos de análisis de varianza en los que el objetivo principal es estudiar el efecto sobre ciertas respuestas de ciertos tratamientos fijos o predeterminados. Los experimentos en los que los tratamientos o los niveles de tratamiento son preseleccionados por el experimentador, a diferencia de aquellos que se eligen al azar, se denominan **experimentos de efectos fijos** o **experimentos del modelo I**. Para el modelo de efectos fijos, las inferencias se hacen sólo sobre aquellos tratamientos particulares que se usó en el experimento.

Con frecuencia es importante que el experimentador sea capaz de hacer inferencias acerca de una población de tratamientos usando un experimento en el que los tratamientos empleados se eligieron al azar de entre la población. Por ejemplo, un biólogo quizás esté interesado en saber si hay o no una varianza significativa en cierta característica fisiológica debida a un tipo de animal. Los tipos de animales que en realidad se usan en el experimento se eligen al azar y representan los efectos del tratamiento. Un químico podría estar interesado en estudiar el efecto de los laboratorios de análisis sobre el análisis químico de una sustancia. No le preocupa ningún laboratorio en particular, sino que se ocupa de una población grande de ellos. Así, puede seleccionar al azar un grupo de laboratorios y asignar muestras a cada uno para que las analice. Entonces, la inferencia estadística implicaría **1.** probar si los laboratorios contribuyen o no a una varianza diferente de cero en los resultados de los análisis, y **2.** estimar la varianza debida a los laboratorios y a la varianza dentro de éstos.

Modelo y suposiciones para el modelo de efectos aleatorios

El **modelo de efectos aleatorios** de un solo factor, que con frecuencia recibe el nombre de **modelo II**, se denota como el modelo de efectos fijos pero sus términos toman significados diferentes. La respuesta

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

ahora es un valor de la variable aleatoria

$$Y_{ij} = \mu + A_i + E_{ij},$$

con $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, n$, donde las Ai tienen distribución normal y son independientes con media igual a cero y varianza σ_α^2, y son independientes de las E_{ij}. Al igual que para el modelo de efectos fijos, las E_{ij} también tienen distribución normal y son independientes, con media igual a cero y varianza σ^2. Observe que para

un experimento del modelo II, la variable aleatoria $\sum_{i=1}^{k} A_i$ adopta el valor $\sum_{i=1}^{\kappa} \alpha_i$; y ya no se aplica la restricción de que la suma de estas α_i sea igual a cero.

Teorema 13.5: Para el modelo del análisis de varianza de un solo factor, de efectos aleatorios,

$$E(SSA) = (k-1)\sigma^2 + n(k-1)\sigma_\alpha^2 \qquad \text{y} \qquad E(SSE) = k(n-1)\sigma^2.$$

La tabla 13.14 muestra los cuadrados esperados de la media para un experimento tanto del modelo I como del modelo II. Los cálculos para un experimento del modelo II se ejecutan exactamente de la misma forma que para el experimento del modelo I. Es decir, la suma de cuadrados, los grados de libertad y las columnas de la media cuadrática en la tabla del análisis de varianza, son las mismas para ambos modelos.

Tabla 13.14: Cuadrados esperados de la media para el experimento de un solo factor

Fuente de variación	**Grados de libertad**	**Cuadrados de la media**	**Cuadrados esperados de la media**	
			Modelo I	**Modelo II**
Tratamientos	$k-1$	s_1^2	$\sigma^2 + \frac{n}{k-1}\sum_i \alpha_i^2$	$\sigma^2 + n\sigma_\alpha^2$
Error	$k(n-1)$	s^2	σ^2	σ^2
Total	$nk-1$			

Para el modelo de efectos aleatorio, las hipótesis de que todos los efectos del tratamiento son iguales a cero se escriben como sigue:

Hipótesis para un experimento del modelo II

$$H_0\colon\ \sigma_\alpha^2 = 0,$$
$$H_1\colon\ \sigma_\alpha^2 \neq 0.$$

Esta hipótesis indica que los tratamientos diferentes no contribuyen en absoluto a la variabilidad de la respuesta. De la tabla 13.14, es evidente que tanto s_1^2 como s^2 son estimadores de σ^2 cuando H_0 es verdad, y que la razón

$$f = \frac{s_1^2}{s^2}$$

es un valor de la variable aleatoria F que tiene la distribución F con $k-1$ y $k(n-1)$ grados de libertad. Con un nivel de significancia α, se rechaza la hipótesis nula cuando

$$f > f_\alpha[k-1, k(n-1)].$$

En muchos estudios científicos y de ingeniería, el interés no se centra en la prueba F. El científico sabe que el efecto aleatorio, en efecto, es significativo. Lo que es más importante es la estimación de los diversos componentes de la varianza. Esto produce un sentido de *ordenación* en términos de cuáles factores producen la máxima variabilidad, y de cuánto. En el contexto presente, resulta de interés cuantificar cuánto más grande es el *componente de la varianza de un solo factor*, que el producido por el azar (variación aleatoria).

Estimación de los componentes de la varianza

La tabla 13.14 también se utiliza para estimar los **componentes de la varianza** σ^2 y σ_α^2. Como s_1^2 estima $\sigma^2 + n\sigma_\alpha^2$ y s^2 estima σ^2,

$$\hat{\sigma}^2 = s^2, \qquad \hat{\sigma}_\alpha^2 = \frac{s_1^2 - s^2}{n}.$$

Ejemplo 13.8: Los datos de la tabla 13.15 representan observaciones del producto de un proceso químico, usando 5 lotes de materia prima seleccionados al azar.

Tabla 13.15: Datos para el ejemplo 13.8

Lote:	**1**	**2**	**3**	**4**	**5**	
	9.7	10.4	15.9	8.6	9.7	
	5.6	9.6	14.4	11.1	12.8	
	8.4	7.3	8.3	10.7	8.7	
	7.9	6.8	12.8	7.6	13.4	
	8.2	8.8	7.9	6.4	8.3	
	7.7	9.2	11.6	5.9	11.7	
	8.1	7.6	9.8	8.1	10.7	
Total	55.6	59.7	80.7	58.4	75.3	329.7

Demuestre que el componente de la varianza del lote es significativamente mayor que cero, y obtenga su estimador.

Solución: Las sumas total, del lote y de los cuadrados del error, son

$$SST = 194.64, \qquad SSA = 72.60, \qquad SSE = 194.64 - 72.60 = 122.04.$$

En la tabla 13.16 se presentan estos resultados, con los cálculos restantes.

Tabla 13.16: Análisis de la varianza para el ejemplo 13.8

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
Lotes	72.60	4	18.15	4.46
Error	122.04	30	4.07	
Total	194.64	34		

La razón f es significativa con un nivel $\alpha = 0.05$, lo cual quiere decir que se rechaza la hipótesis de un componente del lote igual a cero. Una estimación del componente de la varianza del lote es

$$\hat{\sigma}_\alpha^2 = \frac{18.15 - 4.07}{7} = 2.01.$$

Observe que aun cuando el **componente de la varianza del lote** es significativamente distinta de cero, cuando se compara contra el estimador de σ^2, es decir $\hat{\sigma}^2 = MSE = 4.07$, parece como si el componente de varianza del lote no fuera apreciablemente grande.

Diseño aleatorio por bloques, con bloques al azar

En un experimento completo con bloques aleatorios, donde los bloques representen días, es concebible que el experimentador quiera que los resultados se apliquen no sólo a los días reales utilizados en el análisis, sino a cada día del año. Entonces, él seleccionaría al azar los días en que se haría el experimento, así como los tratamientos y el uso del modelo de efectos aleatorios

$$Y_{ij} = \mu + A_i + B_j + \epsilon_{ij}, \qquad i = 1, 2, \ldots, k, \qquad \text{y} \qquad j = 1, 2, \ldots, b,$$

con las A_i, B_j y ϵ_{ij} que son variables aleatorias independientes con medias igual a cero y varianzas σ_α^2, σ_β^2 y σ^2, respectivamente. Los cuadrados esperados de la media para un modelo II de diseño por bloques por completo aleatorios se obtienen, usando el mismo procedimiento que para el problema de un solo factor, y se presentan junto con aquellos para un experimento del modelo I, como se aprecia en la tabla 13.17.

Tabla 13.17: Cuadrados esperados de la media para un diseño por bloques completamente aleatorio

Fuente de variación	**Grados de libertad**	**Cuadrados de la media**	**Cuadrados esperados de la media**	
			Modelo I	**Modelo II**
Tratamientos	$k-1$	s_1^2	$\sigma^2 + \frac{b}{k-1}\sum_i \alpha_i^2$	$\sigma^2 + b\sigma_\alpha^2$
Bloques	$b-1$	s_2^2	$\sigma^2 + \frac{k}{b-1}\sum_j \beta_j^2$	$\sigma^2 + k\sigma_\beta^2$
Error	$(k-1)(b-1)$	s^2	σ^2	σ^2
Total	$kb-1$			

Otra vez, los cálculos para las sumas individuales de cuadrados y los grados de libertad son idénticos para aquellos del modelo de efectos fijos. Las hipótesis

$$H_0\text{: } \sigma_\alpha^2 = 0,$$
$$H_1\text{: } \sigma_\alpha^2 \neq 0,$$

se obtienen calculando

$$f = \frac{s_1^2}{s^2},$$

y H_0 se rechaza cuando $f > f_\alpha[k-1, (b-1)(k-1)]$.

Los estimadores insesgados de los componentes de la varianza son

$$\hat{\sigma}^2 = s^2, \qquad \hat{\sigma}_\alpha^2 = \frac{s_1^2 - s^2}{b}, \qquad \hat{\sigma}_\beta^2 = \frac{s_2^2 - s^2}{k}.$$

Para el diseño del cuadrado latino, el modelo de efectos aleatorios se escribe como

$$Y_{ijk} = \mu + A_i + B_j + T_k + \epsilon_{ijk},$$

para $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, r$, y $k = A, B, C, \ldots$, con A_i, B_j, T_k y ϵ_{ijk} que son variables aleatorias independientes con medias iguales a cero y varianzas σ_α^2, σ_β^2 , σ_τ^2

Tabla 13.18: Cuadrados esperados de la media para un diseño de cuadrado latino

Fuente de variación	Grados de libertad	Cuadrados de la media	Cuadrados esperados de la media	
			Modelo I	Modelo II
Renglones	$r-1$	s_1^2	$\sigma^2 + \frac{r}{r-1}\sum_i \alpha_i^2$	$\sigma^2 + r\sigma_\alpha^2$
Columnas	$r-1$	s_2^2	$\sigma^2 + \frac{r}{r-1}\sum_j \beta_j^2$	$\sigma^2 + r\sigma_\beta^2$
Tratamientos	$r-1$	s_3^2	$\sigma^2 + \frac{r}{r-1}\sum_k \tau_k^2$	$\sigma^2 + r\sigma_\tau^2$
Error	$(r-1)(r-2)$	s^2	σ^2	σ^2
Total	r^2-1			

y σ^2, respectivamente. La obtención de los cuadrados esperados de la media para un diseño de cuadrado latino del modelo II es directa, y para fines de comparación los presentamos en la tabla 13.18 junto con aquellos para un experimento del modelo I.

Las pruebas de hipótesis que conciernen a los diversos componentes de la varianza se efectúan calculando las razones de cuadrados de la media apropiados, como se indica en la tabla 13.18, y se comparan con los valores f correspondientes de la tabla A.6.

13.14 Potencia de las pruebas del análisis de varianza

Como señalamos antes, es frecuente que el investigador se vea obstaculizado por el problema de no saber qué tan *grande* elegir una muestra. En la planeación de un diseño aleatorio por completo de un solo factor con n observaciones por tratamiento, el objetivo principal es probar la hipótesis de igualdad de las medias de los tratamientos.

H_0: $\alpha_1 = \alpha_2 = \cdots \alpha_k = 0$,

H_1: Al menos una de las α_i no es igual a cero.

Sin embargo, con demasiada frecuencia, la varianza del error experimental σ^2 es tan grande que el procedimiento de prueba será insensible a las diferencias reales entre las k medias de los tratamientos. En la sección 13.3, los valores esperados de los cuadrados de las medias para el modelo de un solo factor, están dados por

$$E(S_1^2) = E\left(\frac{SSA}{k-1}\right) = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k}\alpha_i^2, \quad E(S^2) = E\left(\frac{SSE}{k(n-1)}\right) = \sigma^2.$$

Así, para una desviación dada de la hipótesis nula H_0, medida por

$$\frac{n}{k-1}\sum_{i=1}^{k}\alpha_i^2,$$

los valores grandes de σ^2 disminuyen la probabilidad de obtener un valor $f = s_1^2/s^2$ que se encuentre en la región crítica de la prueba. La sensibilidad de la prueba describe la capacidad del procedimiento para detectar diferencias en las medias poblacionales y se mide por la potencia de la prueba (véase la sección 10.2), que es tan

sólo $1 - \beta$, donde β es la probabilidad de aceptar una hipótesis falsa. Entonces, se puede interpretar la potencia de nuestras pruebas de análisis de varianza como la probabilidad de que el estadístico F se halle en la región crítica cuando, de hecho, la hipótesis nula sea falsa y las medias del tratamiento sí difieran. Para la prueba de análisis de varianza de un solo factor, la potencia, $1 - \beta$, es

$$\begin{aligned} 1 - \beta &= P\left[\frac{S_1^2}{S^2} > f_\alpha(v_1, v_2) \text{ cuando } H_1 \text{ es verdadera}\right] \\ &= P\left[\frac{S_1^2}{S^2} > f_\alpha(v_1, v_2) \text{ cuando } \sum_{i=1}^{k} \alpha_i^2 = 0\right]. \end{aligned}$$

El término $f_\alpha(v_1, v_2)$ es, desde luego, el punto crítico de la cola superior de la distribución F con v_1 y v_2 grados de libertad. Para valores dados de $\sum_{i=1}^{k} \alpha_i^2/(k-1)$ y σ^2, la potencia se incrementa con el uso de un tamaño de muestra n más grande. El problema se convierte en uno de diseño del experimento con un valor de n tal que se cumplan los requerimientos de potencia. Por ejemplo, podría requerirse que para valores específicos de $\sum_{i=1}^{k} \alpha_i^2 \neq 0$ y σ^2, se rechace la hipótesis con una probabilidad de 0.9. Cuando la potencia de la prueba es baja, limita con severidad el alcance de las inferencias que se pueden hacer a partir de los datos experimentales.

El caso de los efectos fijos

En el análisis de varianza, la potencia depende de la distribución de la razón F con la hipótesis alternativa de que las medias del tratamiento difieran. Por lo tanto, en el caso del modelo de efectos fijos de un solo factor, se requiere la distribución de S_1^2/S^2 cuando, en realidad,

$$\sum_{i=1}^{k} \alpha_i^2 \neq 0.$$

Por supuesto, cuando la hipótesis nula es verdadera, $\alpha_i = 0$ para $i = 1, 2, \ldots, k$, y el estadístico sigue la distribución F con $k - 1$ y $N - k$ grados de libertad. Si $\sum_{i=1}^{k} \alpha_i^2 \neq 0$, la razón sigue una **distribución *F* no central**.

La variable aleatoria fundamental de la F no central se denota con F'. Sea $f_\alpha(v_1, v_2, \lambda)$ un valor de F' con parámetros v_1, v_2 y λ. Los parámetros v_1 y v_2 de la distribución son los grados de libertad asociados con S_1^2 y S^2, respectivamente, y λ se denomina el **parámetro de no centralidad**. Cuando $\lambda = 0$, la F no central tan sólo se reduce a la distribución F común con v_1 y v_2 grados de libertad.

Para efectos fijos, el análisis de varianza de un solo factor con tamaños de muestra $n_1, n_2, \ldots, n_k$, se define como

$$\lambda = \frac{1}{2\sigma^2} \sum_{i=1}^{k} n_i \alpha_i^2.$$

Si se dispone de tablas de la F no central, la potencia para detectar una alternativa específica se obtiene al evaluar la probabilidad siguiente:

$$1-\beta = P\left[\frac{S_1^2}{S^2} > f_\alpha(k-1, N-k) \text{ cuando } \lambda = \frac{1}{2\sigma^2}\sum_{i=1}^{k} n_i\alpha_i^2\right]$$
$$= P(F' > f_\alpha(k-1, N-k)].$$

Aunque la F no central normalmente está definida en términos de λ, para fines de tabulación es más conveniente trabajar con

$$\phi^2 = \frac{2\lambda}{v_1+1}.$$

La tabla A.16 muestra gráficas de la potencia del análisis de varianza como función de ϕ para distintos valores de v_1, v_2 y el nivel de significancia α. Estas **gráficas de potencia** se emplean no sólo para los modelos de efectos fijos estudiados en este capítulo, sino también para los modelos multifactoriales del capítulo 14. Ahora resta dar un procedimiento con el cual pueda encontrarse el parámetro de no centralidad λ y, por lo tanto, ϕ para estos casos de efectos fijos.

El parámetro de no centralidad λ se escribe, en términos de los **valores esperados del cuadrado de la media en el numerador** de la razón F, en el análisis de varianza. Se tiene que

$$\lambda = \frac{v_1[E(S_i^2)]}{2\sigma^2} - \frac{v_1}{2}$$

por lo que

$$\phi^2 = \frac{[E(S_i^2)-\sigma^2]}{\sigma^2}\frac{v_1}{v_1+1}.$$

En la tabla 13.19 se presentan las expresiones para λ y ϕ^2 para el modelo de un solo factor, el diseño con bloques completamente aleatorios y el diseño con el cuadrado latino.

Tabla 13.19: Parámetro de no centralidad λ y ϕ^2 para el modelo de efectos fijos

	Clasificación de un solo factor	**Bloque completamente aleatorio**	**Cuadrado latino**
λ:	$\frac{1}{2\sigma^2}\sum_i n_i\alpha_i^2$	$\frac{b}{2\sigma^2}\sum_i \alpha_i^2$	$\frac{r}{2\sigma^2}\sum_k \tau_k^2$
ϕ^2:	$\frac{1}{k\sigma^2}\sum_i n_i\alpha_i^2$	$\frac{b}{k\sigma^2}\sum_i \alpha_i^2$	$\frac{1}{\sigma^2}\sum_k \tau_k^2$

En la tabla A.16 observe que para valores dados de v_1 y v_2, la potencia de la prueba se incrementa con valores crecientes de ϕ. El valor de λ depende, por supuesto, de σ^2, y en un problema práctico es frecuente que se necesite sustituir el error cuadrático medio como un estimador para determinar ϕ^2.

Ejemplo 13.9: En un experimento de bloques aleatorios se van a comparar 4 tratamientos en 6 bloques, lo cual da como resultado 15 grados de libertad para el error. ¿Son suficientes

6 bloques, si la potencia de la prueba para detectar diferencias entre las medias del tratamiento, con un nivel de significancia de 0.05, debe ser por lo menos de 0.8 cuando las medias verdaderas sean $\mu_{1.} = 5.0$, $\mu_{2.} = 7.0$, $\mu_{3.} = 4.0$ y $\mu_{4.} = 4.0$? Un estimador de σ^2 para usarse en el cálculo de la potencia está dado por $\hat{\sigma}^2 = 2.0$.

Solución: Hay que recordar que las medias del tratamiento están dadas por $\mu_{i.} = \mu + \alpha_i$. Si se considera la restricción de que $\sum_{i=1}^{4} \alpha_i = 0$, entonces,

$$\mu = \frac{1}{4} \sum_{i=1}^{4} \mu_{i.} = 5.0,$$

y se tiene que $\alpha_1 = 0$, $\alpha_2 = 2.0$, $\alpha_3 = -1.0$, y $\alpha_4 = -1.0$. Por lo tanto,

$$\phi^2 = \frac{b}{k\sigma^2} \sum_{i=1}^{4} \alpha_i^2 = \frac{(6)(6)}{(4)(2)} = 4.5,$$

de la que se obtiene $\phi = 2.121$. Con la tabla A.16 se encuentra que la potencia es aproximadamente de 0.89, por lo que se satisfacen los requerimientos para ella. Esto significa que si el valor de $\sum_{i=1}^{4} \alpha_i^2 = 6$ y $\sigma^2 = 2.0$, el uso de seis bloques dará como resultado el rechazo de la hipótesis de que las medias del tratamiento son iguales con una probabilidad de 0.89.

El caso de los efectos aleatorios

En el caso de efectos fijos, el cálculo de la potencia requiere que se utilice la distribución F no central. Ése no es el caso en el modelo de efectos aleatorios. En realidad, la potencia se calcula de forma muy sencilla usando las tablas F estándar. Por ejemplo, considere el modelo de efectos aleatorios de un solo factor, n observaciones por tratamiento, con las hipótesis

$$H_0\colon\ \sigma_\alpha^2 = 0,$$
$$H_1\colon\ \sigma_\alpha^2 \neq 0.$$

Cuando H_1 es verdadera, la razón

$$f = \frac{SSA/\left[(k-1)(\sigma^2 + n\sigma_\alpha^2)\right]}{SSE/\left[k(n-1)\sigma^2\right]} = \frac{s_1^2}{s^2(1 + n\sigma_\alpha^2/\sigma^2)}$$

es un valor de la variable aleatoria F que tiene distribución F con $k-1$ y $k(n-1)$ grados de libertad. Entonces, el problema se convierte en determinar la probabilidad de rechazar H_0 con la condición de que el componente de la varianza del tratamiento verdadera sea $\sigma_\alpha^2 \neq 0$. Entonces, se tiene

$$\begin{aligned} 1 - \beta &= P\left\{\frac{S_1^2}{S^2} > f_\alpha[k-1, k(n-1)] \text{ cuando } H_1 \text{ es verdadera}\right\} \\ &= P\left\{\frac{S_1^2}{S^2(1 + n\sigma_\alpha^2/\sigma^2)} > \frac{f_\alpha[k-1, k(n-1)]}{1 + n\sigma_\alpha^2/\sigma^2}\right\} \\ &= P\left\{F > \frac{f_\alpha[k-1, k(n-1)]}{1 + n\sigma_\alpha^2/\sigma^2}\right\}. \end{aligned}$$

Observe que conforme n se incrementa, el valor $f_\alpha[k-1, k(n-1)]/(1+n\sigma_\alpha^2/\sigma^2)$ se aproxima a cero, lo cual da como resultado un aumento en la potencia de la prueba. En la figura 13.11 se muestra una ilustración de la potencia para esta clase de situación. El área con sombra *más clara* es el nivel de significancia α, en tanto que la de sombra más *oscura* es la *potencia* de la prueba.

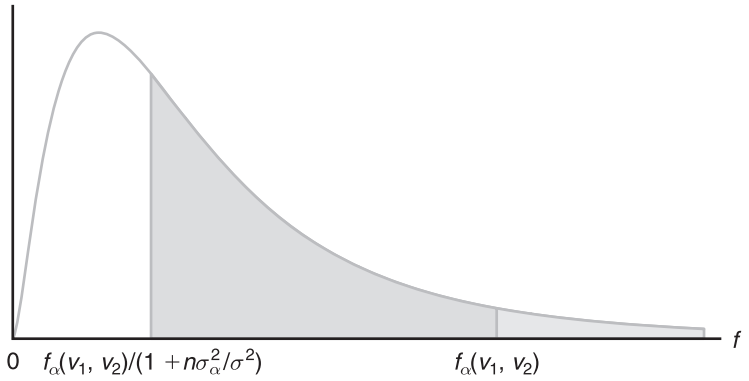


Figura 13.11: Potencia para el análisis de varianza de un solo factor de efectos aleatorios.

Ejemplo 13.10: Suponga que en un problema de un solo factor hay interés en probar la significancia del componente de varianza σ_α^2. En el experimento deben usarse cuatro tratamientos, con 5 observaciones por tratamiento. ¿Cuál será la probabilidad de rechazar la hipótesis $\sigma_\alpha^2 = 0$, cuando en realidad el componente de la varianza del tratamiento es $(3/4)\ \sigma^2$?

Solución: Con un nivel de significancia de $\alpha = 0.05$, se tiene

$$1-\beta = P\left[F > \frac{f_{0.05}(3,16)}{1+(5)(3)/4}\right] = P\left[F > \frac{f_{0.05}(3,16)}{4.75}\right] = P\left(F > \frac{3.24}{4.75}\right)$$
$$= P(F > 0.682) = 0.58.$$

Por lo tanto, el procedimiento de prueba detecta un componente de varianza de $(3/4)\sigma^2$ sólo alrededor del 58% de las veces.

13.15 Estudio de caso

Se pidió al personal del Departamento de Química del Tecnológico de Virginia que analizara un conjunto de datos que se obtuvo para comparar 4 métodos distintos de análisis del aluminio, de cierta mezcla inflamadora sólida. Para considerar un rango amplio de laboratorios de análisis, se utilizaron 5 de ellos en el experimento. Estos laboratorios se seleccionaron porque, en general, son proclives para realizar esa clase de análisis. Se asignaron al azar 20 muestras de material inflamador que contenían 2.70% de aluminio, 4 a cada laboratorio, y se dieron instrucciones acerca de cómo efectuar los análisis químicos utilizando los cuatro métodos. Los datos que se obtuvieron son los siguientes:

	Laboratorio					
Método	**1**	**2**	**3**	**4**	**5**	**Media**
A	2.67	2.69	2.62	2.66	2.70	2.668
B	2.71	2.74	2.69	2.70	2.77	2.722
C	2.76	2.76	2.70	2.76	2.81	2.758
D	2.65	2.69	2.60	2.64	2.73	2.662

Los laboratorios no se consideran efectos aleatorios ya que no fueron seleccionados al azar de entre una población más grande de ellos. Se analizaron los datos como un diseño de bloques por completo aleatorios. Se presentan gráficas de los datos para determinar si es apropiado un modelo aditivo del tipo

$$y_{ij} = \mu + m_i + l_j + \epsilon_{ij}$$

en otras palabras, un modelo con efectos aditivos. El bloque aleatorio no es adecuado cuando existe interacción entre los laboratorios y los métodos. Considere la gráfica de la figura 13.12. Aunque es un poco difícil de interpretar porque cada punto es una sola observación, parece que no hay interacción apreciable entre los métodos y los laboratorios.

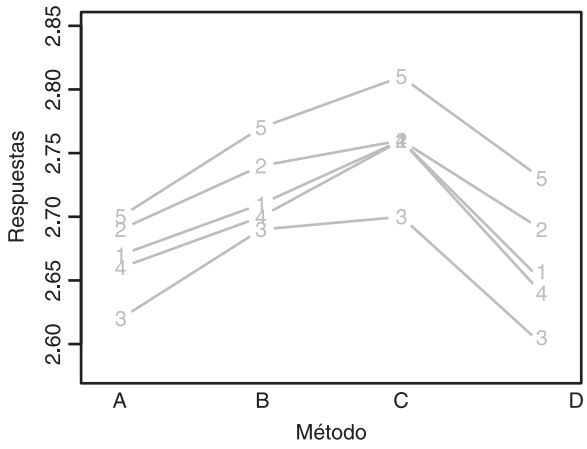


Figura 13.12: Gráfica de interacción para los datos del estudio de caso.

Gráficas de residuos

Las gráficas de residuos se usaron como indicaciones de diagnóstico de la suposición de una varianza homogénea. La figura 13.13 presenta una gráfica de residuos contra los métodos de análisis. La variabilidad que se ilustra en los residuos parece ser bastante homogénea. Para que sea completa, en la figura 13.14 se muestra una gráfica de probabilidad normal de los residuos.

Las gráficas de residuos no presentan dificultad con la suposición de errores normales ni con la de varianza homogénea. Para hacer el análisis de varianza se utilizó el *SAS* PROC GLM. La figura 13.15 muestra la salida por computadora comentada.

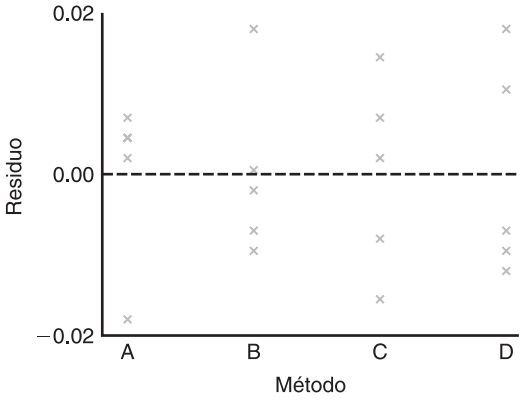


Figura 13.13: Gráfica de residuos contra el método para los datos del estudio de caso.

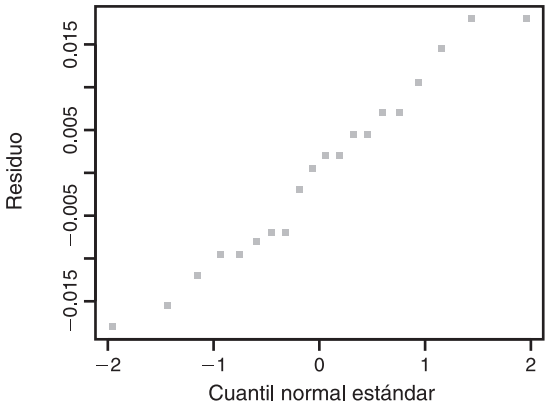


Figura 13.14: Gráfica de probabilidad normal de residuos para los datos del caso de estudio.

Los valores f y P calculados sí indican una diferencia significativa entre los métodos de análisis. A este análisis puede seguir un análisis de comparación múltiple, para determinar en dónde se hallan las diferencias entre los métodos.

Ejercicios

13.43 Los datos siguientes muestran el efecto de cuatro operadores, elegidos al azar, sobre la producción de una máquina específica:

Operador			
1	**2**	**3**	**4**
175.4	168.5	170.1	175.2
171.7	162.7	173.4	175.7
173.0	165.0	175.7	180.1
170.5	164.1	170.7	183.7

a) Desarrolle un análisis de varianza del modelo II con un nivel de significancia de 0.05.

b) Calcule un estimador para el componente de varianza del operador y para el del error experimental.

13.44 Si se supone un modelo de efectos aleatorios, demuestre que

$$E(SSB) = (b-1)\sigma^2 + k(b-1)\sigma_\beta^2$$

para el diseño por bloques completamente aleatorios.

13.45 Se efectúa un experimento en el cual tienen que compararse 4 tratamientos en 5 bloques. Se generaron los siguientes datos:

	Bloque				
Tratamiento	**1**	**2**	**3**	**4**	**5**
1	12.8	10.6	11.7	10.7	11.0
2	11.7	14.2	11.8	9.9	13.8
3	11.5	14.7	13.6	10.7	15.9
4	12.6	16.5	15.4	9.6	17.1

a) Suponiendo un modelo de efectos aleatorios, pruebe la hipótesis de que no hay diferencia entre las medias del tratamiento, con un nivel de significancia de 0.05.

b) Calcule estimadores de los componentes que tienen el tratamiento y el bloque en la varianza.

13.46 Suponga un modelo de efectos aleatorios y demuestre que

$$E(SSTr) = (r-1)(\sigma^2 + r\sigma_r^2)$$

para el diseño de cuadrado latino.

13.47 *a*) Utilizando un enfoque de regresión para el diseño por bloques completamente aleatorios, obtenga las ecuaciones normales $\mathbf{Ab} = \mathbf{g}$ en forma matricial.

b) Demuestre que

$$R(\beta_1, \beta_2, \ldots, \beta_b \mid \alpha_1, \alpha_2, \ldots, \alpha_k) = SSB.$$

```
                    The GLM Procedure
                Class Level Information
              Class         Levels    Values
              Method             4    A B C D
              Lab                5    1 2 3 4 5
            Number of Observations Read          20
            Number of Observations Used          20
Dependent Variable: Response
                         Sum of
Source          DF      Squares   Mean Square  F Value   Pr > F
Model            7   0.05340500    0.00762929    42.19   <.0001
Error           12   0.00217000    0.00018083
Corrected Total 19   0.05557500

R-Square     Coeff Var      Root MSE    Response Mean
0.960954      0.497592      0.013447         2.702500

Source          DF   Type III SS   Mean Square  F Value    Pr > F
Method           3    0.03145500    0.01048500    57.98    <.0001
Lab              4    0.02195000    0.00548750    30.35    <.0001

Observation       Observed          Predicted           Residual
          1     2.67000000         2.66300000         0.00700000
          2     2.71000000         2.71700000        -0.00700000
          3     2.76000000         2.75300000         0.00700000
          4     2.65000000         2.65700000        -0.00700000
          5     2.69000000         2.68550000         0.00450000
          6     2.74000000         2.73950000         0.00050000
          7     2.76000000         2.77550000        -0.01550000
          8     2.69000000         2.67950000         0.01050000
          9     2.62000000         2.61800000         0.00200000
         10     2.69000000         2.67200000         0.01800000
         11     2.70000000         2.70800000        -0.00800000
         12     2.60000000         2.61200000        -0.01200000
         13     2.66000000         2.65550000         0.00450000
         14     2.70000000         2.70950000        -0.00950000
         15     2.76000000         2.74550000         0.01450000
         16     2.64000000         2.64950000        -0.00950000
         17     2.70000000         2.71800000        -0.01800000
         18     2.77000000         2.77200000        -0.00200000
         19     2.81000000         2.80800000         0.00200000
         20     2.73000000         2.71200000         0.01800000
```

Figura 13.15: Salida del *SAS* para los datos del caso de estudio.

13.48 En el ejercicio 13.43, si hubiera interés en probar la significancia del componente de varianza del operador, ¿se tendría muestras suficientemente grandes para garantizar un componente de varianza significativa, con una probabilidad de hasta 0.95, si la σ_α^2 verdadera fuera de 1.5 σ^2? Si no fuera así, ¿cuántas corridas serían necesarias para cada operador? Utilice un nivel de significancia de 0.05.

13.49 Si en el ejercicio 13.45 se acepta un modelo de efectos fijos y se utiliza una prueba con un nivel $\alpha = 0.05$, ¿cuántos bloques serían necesarios para que se aceptara la hipótesis de igualdad de las medias del tratamiento, con una probabilidad de 0.1, cuando en verdad se tuviera que

$$\frac{1}{\sigma^2}\sum_{i=1}^{4}\alpha_i^2 = 2.0?$$

13.50 Compruebe los valores dados para λ y ϕ^2 en la tabla 13.19, para el diseño por bloques completamente aleatorios.

13.51 Al probar las muestras de sangre de un paciente para detectar anticuerpos del VIH, un espectrómetro determina la densidad óptica de cada muestra. La densidad óptica se mide como la absorbencia de la luz de cierta longitud de onda. La muestra de sangre es positiva si excede cierto valor límite que se determina con muestras de control para esa corrida. A los investigadores les interesa comparar la variabilidad del laboratorio para los valores de control positivo. Los datos representan valores de control positivo para 10 pruebas distintas en 4 laboratorios seleccionados al azar.

a) Escriba un modelo adecuado para este experimento.

b) Estime el componente de varianza del laboratorio y la varianza dentro de los laboratorios.

	Laboratorio			
Experimento	**1**	**2**	**3**	**4**
1	0.888	1.065	1.325	1.232
2	0.983	1.226	1.069	1.127
3	1.047	1.332	1.219	1.051
4	1.087	0.958	0.958	0.897
5	1.125	0.816	0.819	1.222
6	0.997	1.015	1.140	1.125
7	1.025	1.071	1.222	0.990
8	0.969	0.905	0.995	0.875
9	0.898	1.140	0.928	0.930
10	1.018	1.051	1.322	0.775

13.52 De 5 "vaciados" de metales se han obtenido 5 muestras de núcleos, y cada una se analizó para la cantidad de cierto elemento traza. Los siguientes son los datos para los 5 vaciados seleccionadas al azar.

	Vaciado				
Núcleo	**1**	**2**	**3**	**4**	**5**
1	0.98	0.85	1.12	1.21	1.00
2	1.02	0.92	1.68	1.19	1.21
3	1.57	1.16	0.99	1.32	0.93
4	1.25	1.43	1.26	1.08	0.86
5	1.16	0.99	1.05	0.94	1.41

a) La intención es que los vaciados sean idénticos. Así, pruebe que el componente de varianza del vaciado es igual a cero. Saque conclusiones.

b) Realice un ANOVA completo junto con un estimador de la varianza dentro del vaciado.

13.53 Una compañía textil produce cierta tela en un número grande de telares. Los administradores querrían que los telares fueran homogéneos, de manera que su tela tuviera resistencia uniforme. Se sospecha que hay variación significativa en la resistencia entre los telares. Considere los datos siguientes para los cuatro telares seleccionados al azar. Cada observación es una determinación de la resistencia de la tela expresada en libras por pulgada cuadrada.

Telar			
1	**2**	**3**	**4**
99	97	94	93
97	96	95	94
97	92	90	90
96	98	92	92

a) Escriba un modelo para el experimento.

b) ¿El componente de la varianza del telar difiere significativamente de cero?

c) Haga comentarios sobre la sospecha.

Ejercicios de repaso

13.54 El Centro de Consultoría en Estadística, junto con el Departamento de Bosques, del Instituto Politécnico y Universidad Estatal de Virginia, llevó a cabo un análisis. Se aplicó cierto tratamiento a un conjunto de tocones de árbol. Se empleó el producto químico Garlon con la finalidad de regenerar las raíces de los tocones. Se usó un aerosol con cuatro niveles de concentración de Garlon. Después de cierto tiempo, se observó la altura de los retoños. Trate los siguientes datos como un análisis de varianza de un solo factor. Haga pruebas para saber si la concentración de Garlon tiene un efecto significativo sobre la altura de los retoños. Emplee $\alpha = 0.05$.

Nivel de Garlon			
1	**2**	**3**	**4**
2.87	3.27	2.39	3.05
2.31	2.66	1.91	0.91
3.91	3.15	2.89	2.43
2.04	2.00	1.89	0.01

13.55 Considere los datos agregados del ejemplo 13.1. Efectúe una prueba de Bartlett para determinar si hay heterogeneidad en la varianza entre los agregados.

13.56 En 1983 el Departamento de Ciencia de Lácteos, del Instituto Politécnico y Universidad Estatal de Virginia, realizó un experimento para estudiar el efecto de las raciones alimenticias, con diferentes fuentes de proteínas, sobre el promedio de producción de leche de las vacas. En el experimento se utilizaron cinco raciones. Se empleó un cuadrado latino 5×5, en el que los renglones representaban vacas diferentes; y las columnas, periodos de lactación distintos. El Centro de Consulta Estadística del Tecnológico de Virginia analizó los siguientes datos, expresados en kilogramos.

	Periodos de lactación				
Vacas	**1**	**2**	**3**	**4**	**5**
1	A: 33.1	C: 30.7	D: 28.7	E: 31.4	B: 28.9
2	B: 34.4	D: 28.7	E: 28.8	A: 22.3	C: 22.3
3	C: 26.4	E: 24.9	A: 20.0	B: 18.7	D: 15.8
4	D: 34.6	A: 28.8	B: 31.9	C: 31.0	E: 30.9
5	E: 33.9	B: 28.0	C: 22.7	D: 21.3	A: 19.0

Con un nivel de significancia de 0.01, ¿es posible concluir que las raciones con fuentes distintas de proteínas tienen un efecto en el promedio diario de producción de leche de las vacas?

13.57 En un proceso químico se utilizaron 3 catalizadores, con un control (no catalizador) incluido. Se tienen los datos siguientes de la producción del proceso:

	Catalizador		
Control	**1**	**2**	**3**
74.5	77.5	81.5	78.1
76.1	82.0	82.3	80.2
75.9	80.6	81.4	81.5
78.1	84.9	79.5	83.0
76.2	81.0	83.0	82.1

Use una prueba de Dunnett con un nivel de significancia de $\alpha = 0.01$ para determinar si se obtiene una producción significativamente más alta con catalizador que sin él.

13.58 Se emplean 4 laboratorios para efectuar análisis químicos. A ellos se envían muestras del mismo material para que las analicen, como parte del estudio, con la finalidad de determinar si en promedio dan o no los mismos resultados. Los resultados analíticos para los 4 laboratorios son los siguientes:

Laboratorio			
A	**B**	**C**	**D**
58.7	62.7	55.9	60.7
61.4	64.5	56.1	60.3
60.9	63.1	57.3	60.9
59.1	59.2	55.2	61.4
58.2	60.3	58.1	62.3

a) Utilice una prueba de Bartlett para demostrar que las varianzas dentro de los laboratorios no son diferentes en forma significativa, con un nivel de $\alpha = 0.05$.

b) Realice una gráfica de probabilidad normal de los residuos.

13.59 Emplee una prueba de Bartlett con un nivel de significancia de 0.01 para probar la homogeneidad de las varianzas del ejercicio 13.9 de la página 523.

13.60 Use una prueba de Cochran con un nivel de significancia de 0.01 para probar la homogeneidad de las varianzas del ejercicio 13.6 de la página 522.

13.61 Emplee una prueba de Bartlett con un nivel de significancia de 0.05 para probar la homogeneidad de las varianzas en el ejercicio 13.9 de la página 523.

13.62 Se diseñó un experimento para el personal del Departamento de Ciencia Animal, en el Instituto Politécnico y Universidad Estatal de Virginia, con el propósito de estudiar el tratamiento con urea y amoniaco acuoso de la espiga del trigo. El propósito era mejorar el valor nutricional para las ovejas macho. Los tratamientos dietéticos son: control; urea en la alimentación; espiga tratada con amoniaco; espiga tratada con urea. En el experimento se emplearon 24 ovejas y se separaron de acuerdo con su peso relativo. En cada grupo homogéneo había seis ovejas. Cada una de éstas recibió cada una de las 4 dietas en orden aleatorio. Para cada una de las 24 ovejas se midió el porcentaje de materia seca digerida. Los siguientes son los datos:

	Grupo por peso (bloque)					
Dieta	**1**	**2**	**3**	**4**	**5**	**6**
Control	32.68	36.22	36.36	40.95	34.99	33.89
Urea en la alimentación	35.90	38.73	37.55	34.64	37.36	34.35
Tratada con amoniaco	49.43	53.50	52.86	45.00	47.20	49.76
Tratada con urea	46.58	42.82	45.41	45.08	43.81	47.40

a) Use un tipo de análisis por bloques aleatorios para probar las diferencias entre las dietas. Use $\alpha = 0.05$.

b) Utilice la prueba de Dunnett para comparar las 3 dietas con el control. Emplee $\alpha = 0.05$.

c) Haga una gráfica de probabilidad normal de los residuos.

13.63 En el conjunto de datos que se analizó para el personal del Departamento de Bioquímica, del Instituto Tecnológico y Universidad Estatal de Virginia, se dieron 3 dietas a un grupo de ratas con el objetivo de estudiar el efecto de cada una sobre el cinc dietético residual en el torrente sanguíneo. Se asignaron al azar 5 ratas preñadas a cada grupo dietético, y a cada una

se le dio la dieta en el día 22 del embarazo. Se midió la cantidad de cinc, en partes por millón. Los datos son los que siguen:

Dieta	1	0.50	0.42	0.65	0.47	0.44
	2	0.42	0.40	0.73	0.47	0.69
	3	1.06	0.82	0.72	0.72	0.82

Determine si hay diferencia significativa en el cinc dietético residual entre las tres dietas. Use $\alpha = 0.05$. Lleve a cabo un ANOVA de un solo factor.

13.64 Se hizo un estudio para comparar el rendimiento de la gasolina para 3 marcas competidoras. Se seleccionaron al azar 4 modelos diferentes de automóvil de tamaño variable. A continuación se presentan los datos, en millas por galón. El orden de prueba es aleatorio para cada modelo.

	Marca de gasolina		
Modelo	**A**	**B**	**C**
A	32.4	35.6	38.7
B	28.8	28.6	29.9
C	36.5	37.6	39.1
D	34.4	36.2	37.9

a) Analice la necesidad de utilizar más de un solo modelo de automóvil.

b) Considere el ANOVA de la siguiente salida del SAS que se observa en la figura 13.16. ¿Es importante la marca de la gasolina?

c) ¿Qué marca de gasolina seleccionaría usted? Consulte el resultado de la prueba de Duncan.

```
                          The GLM Procedure
Dependent Variable: MPG
                              Sum of
 Source             DF       Squares   Mean Square   F Value    Pr > F
 Model               5  153.2508333    30.6501667     24.66    0.0006
 Error               6    7.4583333     1.2430556
 Corrected Total    11  160.7091667

 R-Square     Coeff Var      Root MSE      MPG Mean
 0.953591      3.218448      1.114924      34.64167

 Source             DF   Type III SS   Mean Square  F Value    Pr > F
 Model               3   130.3491667    43.4497222    34.95    0.0003
 Brand               2    22.9016667    11.4508333     9.21    0.0148
                Duncan's Multiple Range Test for MPG
 NOTE: This test controls the Type I comparisonwise error rate, not
 the experimentwise error rate.

                 Alpha                          0.05
                 Error Degrees of Freedom          6
                 Error Mean Square          1.243056

                 Number of Means             2          3
                 Critical Range          1.929      1.999
       Means with the same letter are not significantly different.
              Duncan Grouping           Mean      N    Brand
                          A          36.4000      4    C
                          A
                      B   A          34.5000      4    B
                      B
                      B              33.0250      4    A
```

Figura 13.16: Impresión de SAS para el Ejercicio de Repaso 13.64.

```
                          The GLM Procedure

Dependent Variable: gasket
                                 Sum of
 Source                DF       Squares    Mean Square  F Value   Pr > F
 Model                  5    1.68122778     0.33624556    76.52   <.0001
 Error                 12    0.05273333     0.00439444
 Corrected Total       17    1.73396111
 R-Square      Coeff Var      Root MSE    gasket Mean
 0.969588       1.734095      0.066291        3.822778

 Source                DF  Type III SS    Mean Square  F Value    Pr > F
 material               2   0.81194444     0.40597222    92.38    <.0001
 machine                1   0.10125000     0.10125000    23.04    0.0004
 material*machine       2   0.76803333     0.38401667    87.39    <.0001
  Level of     Level of           ------------gasket-----------
  material     machine       N             Mean          Std Dev
  cork         A             3       4.32666667       0.06658328
  cork         B             3       3.91333333       0.09291573
  plastic      A             3       3.94666667       0.06027714
  plastic      B             3       3.47666667       0.05507571
  rubber       A             3       3.42000000       0.06000000
  rubber       B             3       3.85333333       0.05507571

  Level of           ------------gasket-----------
  material     N             Mean          Std Dev
  cork         6       4.12000000       0.23765521
  plastic      6       3.71166667       0.26255793
  rubber       6       3.63666667       0.24287171

  Level of           ------------gasket-----------
  machine      N             Mean          Std Dev
  A            9       3.89777778       0.39798800
  B            9       3.74777778       0.21376259
```

Figura 13.17: Salida del *SAS* para el ejercicio de repaso 13.65.

13.65 Una compañía que troquela empaques de hojas de caucho, plástico y corcho desea comparar el número medio de empaques producidos por hora para 3 tipos de material. Como bloques se eligieron al azar a 2 máquinas troqueladoras. Los datos representan el número de empaques (en miles) producidos por hora. En la figura 13.17 se observa la salida del análisis.

	Material		
Máquina	**Corcho**	**Caucho**	**Plástico**
A	4.31 4.27 4.40	3.36 3.42 3.48	4.01 3.94 3.89
B	3.94 3.81 3.99	3.91 3.80 3.85	3.48 3.53 3.42

a) ¿Por qué habría que elegir las máquinas troqueladoras como bloques?

b) Grafique las 6 medias para las combinaciones de máquinas y materiales.

c) ¿Hay un material que sea mejor?

d) ¿Existe interacción entre los tratamientos y los bloques? Si es así, diga si la interacción ocasiona alguna dificultad seria para obtener una conclusión adecuada. Explique su respuesta.

13.66 Se realizó un experimento para comparar 3 tipos de pintura, con la finalidad de determinar si hay evidencia de diferencias en sus calidades de uso, y se expusieron a acciones abrasivas y se midieron las horas hasta que la abrasión se notó. Se usaron 6 especímenes para cada tipo de pintura. Los datos son los siguientes:

Tipo de pintura								
1			2			3		
158	97	282	515	264	544	317	662	213
315	220	115	525	330	525	536	175	614

a) Efectúe un análisis de varianza para determinar si la evidencia sugiere que la calidad de uso de las 3 pinturas es diferente. En sus conclusiones utilice un valor *P*.

b) Si se encuentran diferencias significativas, diga cuáles son. ¿Hay una pintura que destaque? Analice sus hallazgos.

c) Haga todos los análisis gráficos que necesite para determinar si son válidas las suposiciones que se hicieron en el inciso *a*). Analice sus hallazgos.

d) Suponga que se determina que los datos para cada tratamiento siguen una distribución exponencial. ¿Sugiere esto un análisis alternativo? Si fuera así, hágalo y dé sus conclusiones.

13.67 Se utilizaron 4 localidades diferentes del noroeste para colectar mediciones del ozono, en partes por millón. Se colectaron cantidades de ozono en 5 muestras en cada localidad.

Localidad			
1	2	3	4
0.09	0.15	0.10	0.10
0.10	0.12	0.13	0.07
0.08	0.17	0.08	0.05
0.08	0.18	0.08	0.08
0.11	0.14	0.09	0.09

a) ¿Hay información suficiente que sugiera que existen diferencias en los niveles medios de ozono de una localidad a otra? Guíese usando un valor *P*.

b) Si se encuentran diferencias significativas en el inciso *a*), caracterice su naturaleza. Emplee cualesquiera métodos que haya aprendido.

13.16 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

Al igual que en otros procedimientos de capítulos anteriores, el análisis de varianza es razonablemente robusto con respecto a la suposición de normalidad; pero lo es menos en cuanto a la de varianza homogénea.

La prueba de Bartlett para igual varianza es débil en extremo con respecto a la normalidad.

Capítulo 14

Experimentos factoriales (dos o más factores)

14.1 Introducción

Considere una situación en la que haya interés por estudiar el efecto de **dos factores**, A y B, sobre alguna respuesta. Por ejemplo, en un experimento químico nos gustaría variar en forma simultánea la presión de reacción y el tiempo de reacción, y estudiar el efecto que cada uno tiene sobre el producto. En un experimento biológico resulta de interés estudiar el efecto que tienen el tiempo de secado y la temperatura sobre la cantidad de sólidos (porcentaje por peso) que queda en las muestras de levadura. Igual que en el capítulo 13, el término **factor** se utiliza en un sentido general para denotar cualquier característica del experimento que pueda variar de un ensayo a otro, como la temperatura, el tiempo o la presión. Los **niveles** de un factor se definen como los valores reales que se utilizan en el experimento.

Para cada uno de estos casos, es importante determinar no sólo si cada uno de los dos factores tiene influencia en la respuesta, sino también si hay interacción significativa entre ellos. Hasta donde concierne a la terminología, el experimento descrito aquí es de dos factores, y el diseño experimental podría ser ya sea uno completamente aleatorio donde las distintas combinaciones de tratamiento se asignan al azar a todas las unidades experimentales, o un diseño por bloques completamente aleatorio donde las combinaciones de factores se asignan al azar a los bloques. En el caso del ejemplo de la levadura, las distintas combinaciones de tratamiento de temperatura y tiempo de secado se asignarían al azar a las muestras de levadura, si se empleara un diseño por completo aleatorio.

En este capítulo se amplían a dos y tres factores muchos de los conceptos que se estudian en el capítulo 13. El objetivo principal de este material es el empleo del diseño completamente aleatorio con un *experimento factorial*. Un experimento factorial con dos factores implica ensayos experimentales (o uno solo) con todas las combinaciones de factores. Por ejemplo, en el caso de la temperatura y tiempo de secado con, digamos, tres niveles de cada uno y $n = 2$ corridas por cada una de las nueve combinaciones, tendríamos un *diseño factorial de dos factores completamente aleatorio*. Ninguno de ellos es un obstáculo; nos interesa la manera en que cada uno influye en el porcentaje de sólidos en las muestras, y si interactúan o no. Entonces, el biólogo

dispondría de 18 muestras físicas de material que constituirían unidades experimentales. Luego, éstas se asignarían al azar a las 18 combinaciones (nueve combinaciones de tratamiento, cada una de ellas por duplicado).

Antes de entrar en detalles analíticos, sumas de cuadrados y demás, sería interesante que el lector observe la clara conexión entre lo que hemos descrito y la situación con el problema de un solo factor. Considere el experimento de la levadura. La explicación de los grados de libertad ayuda a que el lector o el analista visualicen la ampliación del método. En un inicio deberían verse a las 9 combinaciones de tratamientos como si representaran un factor con 9 niveles (8 grados de libertad). Así, un vistazo inicial a los grados de libertad arroja lo siguiente:

Combinaciones de tratamiento	8
Error	9
Total	17

Efectos principales e interacción

En realidad, el experimento podría analizarse como se describe en la tabla anterior. Sin embargo, es probable que la prueba F para las combinaciones no dé al analista la información que desea, es decir, el papel de la temperatura y del tiempo de secado. Tres tiempos de secado tienen asociados 2 grados de libertad, y a tres temperaturas se asocian también 2 grados de libertad. Los factores principales, temperatura y tiempo de secado, reciben el nombre de **efectos principales**, los cuales representan 4 de los 8 grados de libertad para las *combinaciones de factores*. Los 4 grados de libertad adicionales se asocian con la *interacción* entre los dos factores. Como resultado, el análisis incluye

Combinaciones		8
Temperatura	2	
Tiempo de secado	2	
Interacción	4	
Error		9
Total		17

Del capítulo 13 es necesario recordar que en un análisis de varianza los factores pueden verse como fijos o aleatorios, en función del tipo de inferencia que se desea hacer y de la manera en que se eligen los niveles. Aquí, se debe considerar los efectos fijos, los aleatorios e incluso los casos en que los efectos son mixtos. Conforme avancemos en estos temas pondremos mayor atención en los cuadrados esperados de las medias. En la siguiente sección nos centraremos en el concepto de interacción.

14.2 Interacción en el experimento de dos factores

En el modelo de bloques aleatorizados que se estudió en forma previa, se supuso que en cada bloque se tomaba una observación de cada tratamiento. Si la suposición del modelo es correcta, es decir, si los bloques y los tratamientos son los únicos efectos reales y la interacción no existe, el valor esperado del error cuadrático de la media es la varianza, σ^2, del error experimental. Sin embargo, suponga que existe interacción entre los tratamientos y los bloques, como lo indica el modelo

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$$

de la sección 13.9. El valor esperado del error cuadrático de la media se había dado como

$$E\left[\frac{SSE}{(b-1)(k-1)}\right] = \sigma^2 + \frac{1}{(b-1)(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2.$$

Los efectos del tratamiento y los bloques no aparecen en el error cuadrático esperado de la media, aunque los efectos de la interacción sí. Entonces, si en el modelo hay interacción, el error cuadrático de la media refleja variación debida al error experimental más una contribución de la interacción y, para este plan experimental, no hay forma de separarlos.

La interacción y la interpretación de los efectos principales

Desde el punto de vista del experimentador, parecería necesario llegar a una prueba significativa sobre la existencia de interacción, al separar la variación del error verdadero de aquel que se debe a la interacción. Los efectos principales, A y B, adoptan un significado distinto en presencia de la interacción. En el ejemplo biológico anterior, el efecto que el tiempo de secado tiene sobre la cantidad de sólidos que quedan en la levadura muy bien podrían depender de la temperatura a la que se expusieron las muestras. En general, habría situaciones experimentales en las cuales el factor A tuviera un efecto positivo sobre la respuesta en un nivel del factor B; en tanto que con un nivel distinto de éste, el efecto de A sería negativo. Aquí se usa el término **efecto positivo** para indicar que el producto o la respuesta se incrementan conforme los niveles de un factor dado aumentan de acuerdo con cierto orden definido. En el mismo sentido, un **efecto negativo** corresponde a una disminución del producto para niveles crecientes del factor.

Por ejemplo, considere los datos siguientes de temperatura (factor A con niveles t_1, t_2 y t_3, en orden creciente) y tiempo de secado d_1, d_2 y d_3 (también en orden creciente). La respuesta se expresa en porcentaje de sólidos. Estos datos son hipotéticos por completo, y se dan para ilustrar un aspecto.

	B			
A	d_1	d_2	d_3	**Total**
t_1	4.4	8.8	5.2	18.4
t_2	7.5	8.5	2.4	18.4
t_3	9.7	7.9	0.8	18.4
Total	21.6	25.2	8.4	55.2

Es evidente que el efecto de la temperatura es positivo sobre el porcentaje de sólidos con el tiempo de secado bajo d_1, pero negativo para el tiempo alto d_3. Esta **interacción clara** entre la temperatura y el tiempo de secado tiene un interés notorio para el biólogo; pero, con base en los totales de las respuestas para las temperaturas t_1, t_2 y t_3, la suma de los cuadrados de la temperatura, SSA, produciría un valor de cero. Entonces, se dice que la presencia de la interacción **enmascara** el efecto de la temperatura. Por ello, si se considera el efecto medio de la temperatura, promediado para el tiempo de secado, **no existe efecto alguno**. Entonces, esto define el efecto principal. Pero, por supuesto, es probable que esto no sea pertinente para el biólogo.

Antes de sacar cualesquiera conclusiones finales de las pruebas de significancia sobre los efectos principales y los de la interacción, el **experimentador primero debería observar si la prueba para la interacción es significativa o no**. Si

la interacción no es significativa, entonces los resultados de las pruebas sobre los efectos principales carecen de significado. No obstante, si la interacción debe ser significativa, entonces únicamente tienen significado aquellas pruebas sobre los efectos principales que resultan significativos. En presencia de interacción, los efectos principales no significativos bien podrían ser resultado del enmascaramiento, y dictan la necesidad de observar la influencia de cada factor a niveles fijos del otro.

Representación gráfica de la interacción

La presencia de interacción, así como su influencia científica, puede interpretarse bien usando **gráficas de interacción**. Éstas dan con claridad una visión panorámica de la tendencia de los datos para mostrar el efecto que tiene cambiar un factor, conforme se pasa de un nivel a otro del segundo factor. La figura 14.1 ilustra la interacción tan marcada entre la temperatura y el tiempo de secado. La interacción se revela en la falta de paralelismo entre las líneas.

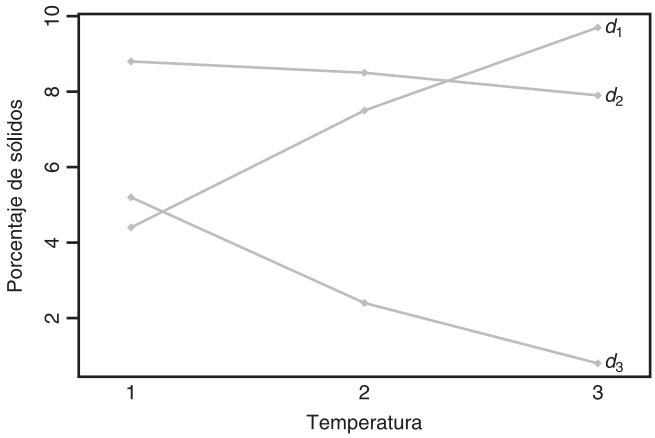


Figura 14.1: Gráfica de la interacción para los datos de temperatura y de tiempo de secado.

El *efecto relativamente fuerte de la temperatura* sobre el porcentaje de sólidos con el tiempo de secado más bajo se refleja en la marcada pendiente de d_1. Con el tiempo de secado medio, d_2, la temperatura tiene muy poco efecto; mientras que con el tiempo de secado alto d_3 la pendiente negativa indica un efecto negativo de la temperatura. Las gráficas de la interacción como las que se muestran dan al científico una interpretación rápida y significativa de la interacción que existe. Debe ser claro que el **paralelismo** en las gráficas indica una **ausencia de interacción**.

Necesidad de observaciones múltiples

En el experimento de dos factores, la interacción y el error experimental sólo se separan si se hacen observaciones múltiples con las distintas combinaciones de tratamiento. Para máxima eficiencia, con cada combinación debe haber el mismo número n de observaciones. Éstas deben ser repeticiones verdaderas, no únicamente medi-

ciones verdaderas. Por ejemplo, en la ilustración de la levadura, si para cada combinación de temperatura y tiempo de secado se tomaran $n = 2$ observaciones, habría dos muestras distintas y no sólo mediciones repetidas sobre la misma muestra. Esto permitiría que la variabilidad debida a las unidades experimentales apareciera en el "error", de manera que la variación no sólo fuera error de medición.

14.3 Análisis de varianza de dos factores

Para presentar las fórmulas generales para el análisis de varianza de un experimento de dos factores que utiliza observaciones repetidas en un diseño por completo aleatorio, debe considerarse el caso de n repeticiones de las combinaciones del tratamiento, determinadas por a niveles del factor A y b niveles del factor B. Las observaciones pueden clasificarse usando un arreglo rectangular, donde los renglones representan los niveles del factor A; y las columnas, los del factor B. Cada combinación de tratamiento define una celda del arreglo. Así, se tienen ab celdas, cada una de las cuales contiene n observaciones. Se denota con y_{ijk} la k-ésima observación tomada en el i-ésimo nivel del factor A y el j-ésimo nivel del factor B; en la tabla 14.1 se muestran las abn observaciones.

Tabla 14.1: Experimento con dos factores con n repeticiones

	B					
A	**1**	**2**	$\cdots$	b	**Total**	**Media**
1	y_{111}	y_{121}	$\cdots$	y_{1b1}	$Y_{1..}$	$\bar{y}_{1..}$
	y_{112}	y_{122}	$\cdots$	y_{1b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{11n}	y_{12n}	$\cdots$	y_{1bn}		
2	y_{211}	y_{221}	$\cdots$	y_{2b1}	$Y_{2..}$	$\bar{y}_{2..}$
	y_{212}	y_{222}	$\cdots$	y_{2b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{21n}	y_{22n}	$\cdots$	y_{2bn}		
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
a	y_{a11}	y_{a21}	$\cdots$	y_{ab1}	$Y_{a..}$	$\bar{y}_{a..}$
	y_{a12}	y_{a22}	$\cdots$	y_{ab2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{a1n}	y_{a2n}	$\cdots$	y_{abn}		
Total	$Y_{.1.}$	$Y_{.2.}$	$\cdots$	$Y_{.b.}$	$Y_{...}$	
Media	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\cdots$	$\bar{y}_{.b.}$		$\bar{y}_{...}$

Las observaciones en la celda (ij)-ésima constituyen una muestra aleatoria de tamaño n de una población que se supone tiene distribución normal con media μ_{ij} y varianza σ^2. Se supone que todas las ab poblaciones tienen la misma varianza σ^2.

Se definen los símbolos siguientes, que son de utilidad y algunos de los cuales se utilizaron en la tabla 14.1:

$Y_{ij.}$ = suma de las observaciones en la (ij)-ésima celda,
$Y_{i..}$ = suma de las observaciones para el i-ésimo nivel del factor A,
$Y_{.j.}$ = suma de las observaciones para el j-ésimo nivel del factor B,
$Y_{...}$ = suma de todas las abn observaciones,
$\bar{y}_{ij.}$ = media de las observaciones en la (ij)-ésima celda,
$\bar{y}_{i..}$ = media de las observaciones para el i-ésimo nivel del factor A,
$\bar{y}_{.j.}$ = media de las observaciones para el j-ésimo nivel del factor B,
$\bar{y}_{...}$ = media de todas las abn observaciones.

A diferencia de la situación para un solo factor, que se cubrió con amplitud en el capítulo 13, en éste supondremos que las **poblaciones**, de las que se toman n observaciones independientes con distribución idéntica, son **combinaciones** de los factores. Asimismo, se supondrá siempre que de cada combinación de factores se toma un número igual (n) de observaciones. En los casos en que los tamaños de las muestras por combinación son desiguales, los cálculos son más complicados, aunque los conceptos son transferibles.

Modelo e hipótesis para el problema con dos factores

Cada observación de la tabla 14-1 puede escribirse en la siguiente forma:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

donde ϵ_{ijk} mide las desviaciones, con respecto de la media μ_{ij}, de los valores y_{ijk} observados en la (ij)-ésima celda. Si $(\alpha\beta)_{ij}$ denota el efecto de la interacción del i-ésimo nivel del factor A y el j-ésimo nivel del factor B, α_i el efecto del i-ésimo nivel del factor A, β_j el efecto del j-ésimo nivel del factor B, y μ la media conjunta, escribimos

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

y, entonces,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

a las que se imponen las restricciones

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0, \qquad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0.$$

Las tres hipótesis por probar son las siguientes:

1. H_0': $\alpha_1 = \alpha_2 = \cdots \ \alpha_a = 0$,
H_1': Al menos una de las α_i no es igual a cero.
2. H_0'': $\beta_1 = \beta_2 = \cdots \ \beta_b = 0$,
H_1'': Al menos una de las β_j no es igual a cero.

3. H_0''': $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots\ (\alpha\beta)_{ab} = 0$,
H_1''': Al menos una de las $(\alpha\beta)_{ij}$ no es igual a cero.

Se alerta al lector del problema del enmascaramiento de los efectos principales cuando la interacción es un contribuyente de importancia en el modelo. Se recomienda que, en primer lugar, se considere el resultado de la prueba de interacción y, luego, la interpretación de la prueba del efecto principal; la naturaleza de la conclusión científica depende de si hay interacción. Si ésta no existe, entonces pueden probarse las hipótesis 1 y 2, y la interpretación es muy sencilla. No obstante, si se descubre que hay interacción la interpretación sería más complicada, como se vio al analizar el tiempo de secado y la temperatura en la sección previa. La estructura de las pruebas de hipótesis 1, 2 y 3 se estudiarán en las secciones siguientes. En el análisis del ejemplo 14.1 se tratará la interpretación de los resultados.

Las pruebas de hipótesis anteriores se basarán en la comparación de estimadores independientes de σ^2 proporcionados por la separación de la suma total de cuadrados de los datos en cuatro componentes, mediante la siguiente identidad.

Partición de la variabilidad en el caso de dos factores

Teorema 14.1: ***Identidad de la suma de cuadrados***

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2$$
$$+n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2$$

Simbólicamente, la identidad de la suma de cuadrados se escribe así

$$SST = SSA + SSB + SS(AB) + SSE,$$

donde SSA y SSB se denominan la suma de cuadrados para los efectos principales A y B, respectivamente, $SS(AB)$ recibe el nombre de suma de cuadrados de la interacción para A y B, y SSE es la suma de errores al cuadrado. La partición de los grados de libertad se efectúa de acuerdo con la identidad

$$abn - 1 = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1).$$

Formación de los cuadrados de la media

Si dividimos cada una de las sumas de los cuadrados en el lado derecho de la identidad de la suma de cuadrados entre su número correspondiente de grados de libertad, obtenemos los cuatro estadísticos

$$S_1^2 = \frac{SSA}{a-1}, \qquad S_2^2 = \frac{SSB}{b-1}, \qquad S_3^2 = \frac{SS(AB)}{(a-1)(b-1)}, \qquad S^2 = \frac{SSE}{ab(n-1)}.$$

Todos estos estimadores de la varianza son estimadores independientes de σ^2, a condición de que no haya efectos α_i, β_j ni, por supuesto, $(\alpha\beta)_{ij}$. Si la suma de cua-

drados se interpreta como funciones de las variables aleatorias independientes y_{111}, $y_{112}, \ldots, y_{abn}$, no es difícil comprobar que

$$E(S_1^2) = E\left[\frac{SSA}{a-1}\right] = \sigma^2 + \frac{nb}{a-1}\sum_{i=1}^{a}\alpha_i^2,$$

$$E(S_2^2) = E\left[\frac{SSB}{b-1}\right] = \sigma^2 + \frac{na}{b-1}\sum_{j=1}^{b}\beta_j^2,$$

$$E(S_3^2) = E\left[\frac{SS(AB)}{(a-1)(b-1)}\right] = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2,$$

$$E(S^2) = E\left[\frac{SSE}{ab(n-1)}\right] = \sigma^2,$$

para las cuales se observa de inmediato que los cuatro estimadores de σ^2 son insesgados cuando H_0', H_0'' y H_0''' son verdaderas.

Para probar la hipótesis H_0' de que los efectos de los factores A son todos iguales a cero, se calcula la siguiente razón:

Prueba F para el factor A

$$f_1 = \frac{s_1^2}{s^2},$$

que es un valor de la variable aleatoria F_1 con distribución F con $a-1$ grados de libertad y $ab(n-1)$ grados de libertad cuando H_0' es verdadera. La hipótesis nula se rechaza al nivel de significancia α cuando $f_1 > f_\alpha[a-1, ab(n-1)]$. Asimismo, para probar la hipótesis H_0'' de que los efectos del factor B todos son iguales a cero, se calcula la razón

Prueba F para el factor B

$$f_2 = \frac{s_2^2}{s^2},$$

que es un valor de la variable aleatoria F_2 con distribución F con $b-1$ y $ab(n-1)$ grados de libertad cuando H_0'' es verdadera. Esta hipótesis se rechaza con el nivel de significancia α cuando $f_2 > f_\alpha[b-1, ab(n-1)]$. Por último, para probar la hipótesis H_0''' que los efectos de la interacción son todos iguales a cero, se calcula la razón que sigue:

Prueba F para la interacción

$$f_3 = \frac{s_3^2}{s^2},$$

que es un valor de la variable aleatoria F_3 que tiene distribución F con $(a-1)(b-1)$ y $ab(n-1)$ grados de libertado cuando H_0''' es verdadera. Concluimos que hay interacción cuando $f_3 > f_\alpha[(a-1)(b-1), ab(n-1)]$.

Como se indicó en la sección 14.2, se recomienda interpretar la prueba para la interacción antes de tratar de sacar inferencias sobre los efectos principales. Si la interacción no es significativa, es claro que hay evidencia de que las pruebas sobre los efectos principales son interpretables. El rechazo de la hipótesis 1 de la página 578 implica que las medias de la respuesta a los niveles del factor A son distintos en forma significativa; mientras que el rechazo de la hipótesis 2 implica una condición similar para las medias a los niveles del factor B. Sin embargo, una interacción significativa podría muy bien implicar que los datos deberían analizarse en forma algo distinta —**quizá con la observación del efecto del factor *A* a los niveles del factor *B***, y así sucesivamente.

Los cálculos del problema de análisis de la varianza para un experimento de dos factores con n repeticiones, por lo general, se resumen como se ilustra en la tabla 14.2.

Tabla 14.2: Análisis de varianza para el experimento de dos factores con n repeticiones

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
Efecto principal				
A	SSA	$a-1$	$s_1^2 = \frac{SSA}{a-1}$	$f_1 = \frac{s_1^2}{s^2}$
B	SSB	$b-1$	$s_2^2 = \frac{SSB}{b-1}$	$f_2 = \frac{s_2^2}{s^2}$
Interacciones de dos factores				
AB	$SS(AB)$	$(a-1)(b-1)$	$s_3^2 = \frac{SS(AB)}{(a-1)(b-1)}$	$f_3 = \frac{s_3^2}{s^2}$
Error	SSE	$ab(n-1)$	$s^2 = \frac{SSE}{ab(n-1)}$	
Total	SST	$abn-1$		

Ejemplo 14.1: En un experimento que se realice para determinar cuál de tres sistemas de misiles distintos es preferible, se midió la tasa de combustión del propulsor para 24 arranques estáticos. Se emplearon 4 tipos de combustible diferentes. El experimento generó observaciones duplicadas de las tasas de combustión con cada combinación de los tratamientos.

Los datos, ya codificados, se dan en la tabla 14.3. Pruebe las siguientes hipótesis: *a*) $H_0^{'}$: no hay diferencia en las tasas medias de combustión del propulsor cuando se emplean diferentes sistemas de misiles, *b*) $H_0^{''}$: no existe diferencia en las tasas medias de combustión de los cuatro tipos de propulsor, *c*) $H_0^{'''}$: no hay interacción entre los distintos sistemas de misiles y los tipos diferentes de propulsor.

Tabla 14.3: Tasas de combustión del propulsor

Sistema de misiles	**Tipo de propulsor**			
	b_1	b_2	b_3	b_4
a_1	34.0	30.1	29.8	29.0
	32.7	32.8	26.7	28.9
a_2	32.0	30.2	28.7	27.6
	33.2	29.8	28.1	27.8
a_3	28.4	27.3	29.7	28.8
	29.3	28.9	27.3	29.1

Solución:

1. *a*) $H_0^{'}$: $\alpha_1 = \alpha_2 = \alpha_3 = 0$.
b) $H_0^{''}$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.
c) $H_0^{'''}$: $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots\ (\alpha\beta)_{34} = 0$.

2. *a*) $H_1^{'}$: Al menos una de las α_i no es igual a cero.
b) $H_1^{''}$: Al menos una de las β_j no es igual a cero.
c) $H_1^{'''}$: Al menos una de las $(\alpha\beta)_{ij}$ no es igual a cero.

Se utiliza la fórmula de la suma de cuadrados que se describió en el teorema 14.1. En la tabla 14.4 se presenta el análisis de varianza.

Tabla 14.4: Análisis de varianza para los datos de la tabla 14.3

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
Sistema de misiles	14.52	2	7.26	5.84
Tipo de propulsor	40.08	3	13.36	10.75
Interacción	22.16	6	3.69	2.97
Error	14.91	12	1.24	
Total	92.68	23		

Se remite al lector al procedimiento GLM (modelos lineales generales) del *SAS* para el análisis de los datos de tasa de combustión, que aparece en la figura 14.2. Observe la forma en que al principio se prueba el "modelo" (11 grados de libertad), y por separado se prueban el sistema, el tipo y el sistema por tipo de interacción. La prueba f sobre el modelo ($P = 0.0030$) prueba la acumulación de los dos efectos principales y la interacción.

a) Rechace H_0' y concluya que los distintos sistemas de misiles generan tasas medias diferentes de combustión del propulsor. El valor P es de aproximadamente 0.017.

b) Rechace H_0'' y concluya que las tasas medias de combustión del propulsor no son las mismas para los cuatro tipos de propulsor. El valor P es más pequeño que 0.0010.

c) La interacción es casi insignificante al nivel 0.05, pero el valor P de aproximadamente 0.0512 indicaría que la interacción debe tomarse en serio.

```
                          The GLM Procedure
Dependent Variable: rate
                             Sum of
Source                DF       Squares   Mean Square   F Value   Pr > F
Model                 11   76.76833333    6.97893939      5.62   0.0030
Error                 12   14.91000000    1.24250000
Corrected Total       23   91.67833333

R-Square     Coeff Var      Root MSE     rate Mean
0.837366      3.766854      1.114675      29.59167

Source                DF   Type III SS   Mean Square  F Value    Pr > F
system                 2   14.52333333    7.26166667     5.84    0.0169
type                   3   40.08166667   13.36055556    10.75    0.0010
system*type            6   22.16333333    3.69388889     2.97    0.0512
```

Figura 14.2: Salida de *SAS* del análisis de los datos de la combustión del propulsor de la tabla 14.3.

En este momento, debe establecerse algún tipo de interpretación de la interacción. Debe hacerse énfasis en que la significancia estadística de un efecto principal

tan sólo implica que las *medias marginales son significativamente distintas*. Sin embargo, considere la tabla 14.5 de los promedios con dos factores.

Tabla 14.5: Interpretación de la interacción

	b_1	b_2	b_3	b_4	**Promedio**
a_1	33.35	31.45	28.25	28.95	30.50
a_2	32.60	30.00	28.40	27.70	29.68
a_3	28.85	28.10	28.50	28.95	28.60
Promedio	31.60	29.85	28.38	28.53	

Es evidente que hay más información importante en el cuerpo de la tabla —tendencias que son inconsistentes con aquella que describen los promedios marginales. Es claro que la tabla 14.5 sugiere que el efecto del tipo de propulsor depende del sistema que se utiliza. Por ejemplo, para el sistema 3, el efecto del tipo de propulsor no parece ser importante, aunque tiene un efecto grande si se emplea ya sea el sistema 1 o el 2. Esto explica la interacción "significativa" entre esos dos factores. Más adelante se harán más descubrimientos sobre esta interacción.

Ejemplo 14.2: En relación con el ejemplo 14.1, elija dos contrastes ortogonales a la partición de la suma de cuadrados para el sistema de misiles en componentes con un solo grado de libertad, para usarlos en la comparación de los sistemas 1 y 2 con el 3, y el sistema 1 contra el sistema 2.

Solución: El contraste para comparar los sistemas 1 y 2 con el 3 es

$$\omega_1 = \mu_{1.} + \mu_{2.} - 2\mu_{3.}.$$

Un segundo contraste, ortogonal a ω_1, para comparar el sistema 1 con el 2, está dado por $\omega_2 = \mu_{1.} - \mu_{2.}$. Las sumas de los cuadrados con un solo grado de libertad son

$$SS\omega_1 = \frac{[244.0 + 237.4 - (2)(228.8)]^2}{(8)[(1)^2 + (1)^2 + (-2)^2]} = 11.80,$$

y

$$SS\omega_2 = \frac{(244.0 - 237.4)^2}{(8)[(1)^2 + (-1)^2]} = 2.72.$$

Observe que $SS\omega_1 + SS\omega_2 = SSA$, como se esperaba. Los valores f calculados correspondientes a ω_1 y ω_2 son, respectivamente,

$$f_1 = \frac{11.80}{1.24} = 9.5 \qquad \text{y} \qquad f_2 = \frac{2.72}{1.24} = 2.2.$$

Al comparar con el valor crítico $f_{0.05}(1, 12) = 4.755$, se encuentra que f_1 es significativo. En realidad, el valor P es menor que 0.01. Así, el primer contraste indica que se rechaza la hipótesis

$$H_0\colon \ \frac{1}{2}(\mu_{1.} + \mu_{2.}) = \mu_{3.}$$

Como $f_2 < 4.75$, las tasas medias de combustión de los sistemas primero y segundo no son diferentes en forma significativa.

Efecto de la interacción significativa en el ejemplo 14.1

Si es verdadera la hipótesis de que en el ejemplo 14.1 no hay interacción, podríamos hacer las comparaciones *generales* del ejemplo 14.2 relacionado con los sistemas de misiles, en vez de las comparaciones separadas para cada propulsor. De manera similar, es posible realizar comparaciones generales entre los propulsores, en vez de comparar por separado cada sistema de misiles. Por ejemplo, podrían compararse los propulsores 1 y 2 con el 3 y 4, y también el 1 contra el 2. Las razones f resultantes, cada una con 1 y 12 grados de libertad, resultan ser de 24.86 y 7.41, respectivamente, y ambas son muy significativas al nivel 0.05.

De los promedios de los propulsores, parece haber evidencia de que el 1 ofrece la mayor tasa media de combustión. Un experimentador prudente sería cauteloso al sacar concusiones generales en un problema como éste, donde la razón f de la interacción está apenas por debajo del valor crítico 0.05. Por ejemplo, la evidencia conjunta, 31.60 contra 29.85 sobre el promedio para los dos propulsores, indica con claridad que el 1 es superior al 2, en términos de una mayor tasa de combustión. Sin embargo, si nos restringimos al sistema 3, donde tenemos un promedio de 28.85 para el propulsor 1 en oposición a 28.10 para el 2, parece haber una diferencia mínima o incluso ninguna entre estos dos propulsores. De hecho, parecería asimismo que hay una estabilización de las tasas de combustión para los distintos propulsores si se opera con el sistema 3. Es claro que existe evidencia conjunta que indica que el sistema 1 da una tasa de consumo mayor que el sistema 3; pero parece que esta conclusión no se mantiene si nos restringimos al propulsor 4.

El analista puede hacer una prueba t sencilla con el empleo de las tasas de consumo del sistema 3, con la finalidad de recabar evidencias concluyentes de que la interacción *produce dificultades considerables en la obtención de conclusiones generales sobre los efectos principales*. Considere una comparación del propulsor 1 contra el 2 únicamente usando el sistema 3. Se toma prestado un estimador de σ^2 del análisis conjunto, es decir, se utiliza $s^2 = 1.24$ con 12 grados de libertad, y se emplea

$$|t| = \frac{0.75}{\sqrt{2s^2/n}} = \frac{0.75}{\sqrt{1.24}} = 0.67,$$

que no está nada cerca de ser significativa. Esta ilustración sugiere que, en presencia de interacción, debería tenerse cautela con la interpretación estricta de los efectos principales.

Análisis gráfico para el problema de dos factores del ejemplo 14.1

Muchos de los mismos tipos de ilustraciones gráficas que se sugirió emplear en los problemas con un factor también se aplican en el caso de dos factores. Las gráficas en dos dimensiones de las medias de las celdas, o de las combinaciones de tratamientos, brindan un panorama de la presencia de las interacciones entre los dos factores. Además, una gráfica de los residuos contra los valores ajustados bien podría dar una indicación acerca de si se cumple o no la suposición de la varianza homogénea. Por supuesto, es frecuente que una trasgresión de la suposición de varianza homogénea implique un aumento en la varianza del error, conforme *los números de la respuesta se vuelven más grandes*. Como resultado, esta gráfica podría resaltar la trasgresión.

La figura 14.3 muestra la gráfica de las medias de las celdas para el caso del propulsor de los sistemas de misiles, que se ilustraron en el ejemplo 14.1. Observe (gráficamente en este caso) cuánta falta de paralelismo hay. Vea lo aplanado de la

parte de la figura que corresponde al efecto del propulsor en el sistema 3. Esto ilustra la interacción entre los factores. La figura 14.4 muestra la gráfica de los residuos contra los valores ajustados para los mismos datos. No hay dificultad visible con la suposición de la varianza homogénea.

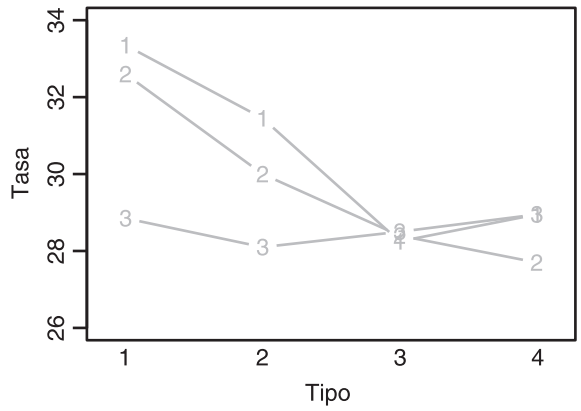


Figura 14.3: Gráfica de las medias de las celdas para los datos del ejemplo 14.1. Los números representan sistemas de misiles.

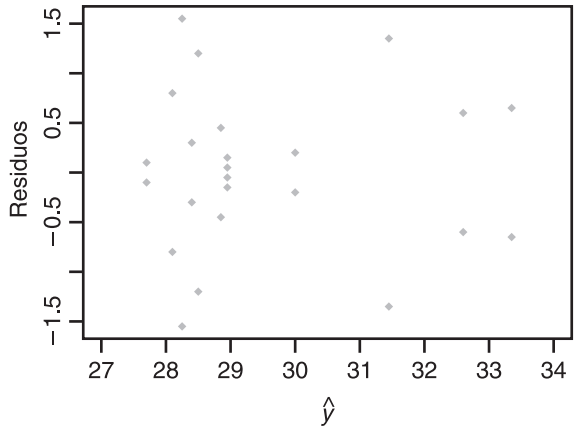


Figura 14.4: Gráfica de los residuos de los datos del ejemplo 14.1.

Ejemplo 14.3: Un ingeniero eléctrico investiga un proceso de grabar con plasma que se emplea en la fabricación de semiconductores. Es de interés estudiar los efectos de dos factores, la tasa de flujo (A) del gas C_2F_6, y la potencia aplicada al cátodo (B). La respuesta es la tasa de grabado. Cada factor se opera a tres niveles y se hacen 2 corridas experimentales sobre la tasa de grabado, para cada una de las 9 combinaciones. El planteamiento es el de un diseño aleatorio por completo. En la tabla 14.6 se presentan los datos. La tasa de grabado se expresa en A°/min.

Los niveles de los factores están en orden ascendente, donde el nivel 1 es el más bajo y el 3 el más alto.

Tabla 14.6: Datos para el ejemplo 14.3

Tasa de flujo del C_2F_6	**Potencia suministrada**		
	1	**2**	**3**
1	288	488	670
	360	465	720
2	385	482	692
	411	521	724
3	488	595	761
	462	612	801

a) Elabore una tabla de análisis de varianza y saque conclusiones, empezando con la prueba sobre la interacción.

b) Haga pruebas sobre los efectos principales y saque conclusiones.

Solución: En la figura 14.5 se muestra una salida del *SAS*. De esa salida se concluye lo siguiente.

```
                          The GLM Procedure
Dependent Variable: etchrate
                                 Sum of
Source                     DF       Squares   Mean Square   F Value   Pr > F
Model                       8   379508.7778    47438.5972     61.00   <.0001
Error                       9     6999.5000      777.7222
Corrected Total            17   386508.2778

R-Square     Coeff Var      Root MSE    etchrate Mean
0.981890      5.057714      27.88767          551.3889

Source                     DF   Type III SS   Mean Square   F Value   Pr > F
c2f6                        2    46343.1111    23171.5556     29.79   0.0001
power                       2   330003.4444   165001.7222    212.16   <.0001
c2f6*power                  4     3162.2222      790.5556      1.02   0.4485
```

Figura 14.5: Salida del *SAS* para el ejemplo 14.3.

a) El valor P para la prueba de interacción es 0.04485. Se concluye que la interacción no es significativa.

b) Existe diferencia significativa en la tasa media de grabado para los 3 niveles de la tasa de flujo del C_2F_6. Una prueba de Duncan muestra que la tasa media de grabado para el nivel 3 es significativamente mayor que para el nivel 2, y la tasa para el nivel 2 es significativamente mayor que para el nivel 1. Véase la figura 14.6*a*).

Con base en el nivel de potencia al cátodo, hay diferencia significativa en la tasa media de grabado. Una prueba de Duncan revela que la tasa de grabado para el nivel 3 es significativamente más alta que para el 2, y que la tasa para el nivel 2 es significativamente más alta en que para el 1. Véase la figura 14.6*b*).

Duncan Grouping	Mean	N	c2f6
A	619.83	6	3
B	535.83	6	2
C	498.50	6	1

(a)

Duncan Grouping	Mean	N	power
A	728.00	6	3
B	527.17	6	2
C	399.00	6	1

(b)

Figura 14.6: *a*) Salida de *SAS* para el ejemplo 14.3 (prueba de Duncan sobre la tasa de flujo del gas); *b*) Salida de *SAS* para el ejemplo 14.3 (prueba de Duncan sobre la potencia).

Ejercicios

14.1 Se realizó un experimento para estudiar el efecto de la temperatura y el tipo de horno sobre la vida de un componente particular que está a prueba. En el experimento se utilizaron 4 tipos de horno y 3 niveles de temperatura. Se asignaron cuatro piezas al azar, 2 para cada combinación de tratamientos, y se registraron los resultados siguientes.

Temperatura (grados)	Horno O_1	O_2	O_3	O_4
500	227	214	225	260
	221	259	236	229
550	187	181	232	246
	208	179	198	273
600	174	198	178	206
	202	194	213	219

Con un nivel de significancia de 0.05, pruebe las hipótesis de que

a) las temperaturas diferentes no tienen efecto sobre la vida del componente;

b) los hornos distintos no tienen efecto en la vida del componente;

c) el tipo de horno y la temperatura no interactúan.

14.2 El Departamento de Nutrición Humana y Alimentos, del Instituto Politécnico y Universidad Estatal de Virginia, realizó un estudio sobre la estabilidad de la vitamina C en el concentrado de jugo de naranja congelado reconstituido, que se almacena en un refrigerador durante un periodo de hasta una semana; su título era *Vitamin C Retention in Reconstituted Frozen Orange Juice*. Se probaron 3 tipos de concentrado de jugo de naranja congelado con 3 periodos distintos, los cuales se refieren al número de días desde que se mezcló el jugo hasta que se probó. Se registraron los resultados, en miligramos de ácido ascórbico por litro. Utilice un nivel de significancia de 0.05 para probar las hipótesis de que

a) no hay diferencia en el contenido de ácido ascórbico entre las diferentes marcas de concentrado de jugo de naranja;

b) no existe diferencia en el contenido de ácido ascórbico para distintos periodos;

c) se combinaron las marcas de concentrado de jugo de naranja y el número de días hasta que se probó que no interactuaban.

Marca	Tiempo (días) 0		3		7	
Richfood	52.6	54.2	49.4	49.2	42.7	48.8
	49.8	46.5	42.8	53.2	40.4	47.6
Sealed-Sweet	56.0	48.0	48.8	44.0	49.2	44.0
	49.6	48.4	44.0	42.4	42.0	43.2
Minute Maid	52.5	52.0	48.0	47.0	48.5	43.3
	51.8	53.6	48.2	49.6	45.2	47.6

14.3 Se estudiaron 3 variedades de ratas en 2 ambientes distintos, para analizar su desempeño en una prueba de laberinto. Los siguientes son los registros del error de las 48 ratas:

Ambiente	Variedad Brillante		Mezclada		Lenta	
Libre	28	12	33	83	101	94
	22	23	36	14	33	56
	25	10	41	76	122	83
	36	86	22	58	35	23
Restringido	72	32	60	89	136	120
	48	93	35	126	38	153
	25	31	83	110	64	128
	91	19	99	118	87	140

Utilice un nivel de significancia de 0.01 para probar la hipótesis de que

a) no hay diferencia en los registros del error para ambientes diferentes;

b) no existe diferencia en los registros del error para variedades distintas;

c) los ambientes y las variedades de las ratas no interactúan.

14.4 La fatiga por corrosión de los metales se define como la acción simultánea de tensión cíclica y ataque

químico sobre una estructura metálica. Una técnica muy utilizada para minimizar el daño de la fatiga por corrosión en el aluminio, requiere la aplicación de un recubrimiento protector. En un estudio efectuado por el Departamento de Ingeniería Mecánica del Instituto Politécnico y Universidad Estatal de Virginia, se utilizaron distintos niveles de humedad relativa:

Bajo: 20 a 25%.

Medio: 55 a 60%.

Alto: 86 a 91%.

También se emplearon 3 tipos de recubrimientos:

No revestido: Sin recubrimiento.

Anodizado: Recubrimiento de óxido anódico por ácido sulfúrico.

Conversión: Recubrimiento por conversión química de cromato.

Los datos de fatiga por corrosión, expresados en miles de ciclos hasta que se presenta la falla, se registraron como sigue:

	Humedad relativa					
Recubrimiento	**Baja**		**Media**		**Alta**	
No revestido	361	469	314	522	1344	1216
	466	937	244	739	1027	1097
	1069	1357	261	134	1011	1011
Anodizado	114	1032	322	471	78	466
	1236	92	306	130	387	107
	533	211	68	398	130	327
Conversión	130	1482	252	874	586	524
	841	529	105	755	402	751
	1595	754	847	573	846	529

a) Lleve a cabo un análisis de varianza con $\alpha = 0.05$ para probar la significancia de los efectos y de la interacción principales.

b) Utilice la prueba de Duncan de rango múltiple con un nivel de significancia de 0.05, para determinar cuáles niveles de humedad relativa dan como resultado daños distintos de fatiga por corrosión.

14.5 Para determinar cuáles músculos necesitan sujetarse a un programa de acondicionamiento para mejorar el rendimiento individual en el servicio tendido que se usa en el tenis, el Departamento de Salud, Educación Física y Recreación del Instituto Politécnico y Universidad Estatal de Virginia realizó un estudio de cinco músculos diferentes:

1: deltoides anterior
2: pectoral mayor
3: deltoides posterior
4: deltoides medio
5: tríceps.

los cuales se probaron en 3 sujetos, y el experimento se efectuó 3 veces para cada combinación de tratamiento. Los datos electromiográficos se registraron durante el servicio, y se presentan a continuación. Use un nivel de 0.01 de significancia para probar las hipótesis de que

a) diferentes sujetos tienen mediciones iguales del electromiograma;

b) músculos diferentes no tienen efecto en las mediciones del electromiograma;

c) los sujetos y los tipos de músculo no interactúan.

	Músculo				
Sujeto	**1**	**2**	**3**	**4**	**5**
1	32	5	58	10	19
	59	1.5	61	10	20
	38	2	66	14	23
2	63	10	64	45	43
	60	9	78	61	61
	50	7	78	71	42
3	43	41	26	63	61
	54	43	29	46	85
	47	42	23	55	95

14.6 Se realizó un experimento para incrementar la adherencia de los productos de caucho. Se elaboraron 16 productos con el aditivo nuevo, y otros 16 sin éste. La adherencia que se registró es la siguiente.

	Temperatura (°C)			
	50	**60**	**70**	**80**
Con el aditivo	2.3	3.4	3.8	3.9
	2.9	3.7	3.9	3.2
	3.1	3.6	4.1	3.0
	3.2	3.2	3.8	2.7
Sin el aditivo	4.3	3.8	3.9	3.5
	3.9	3.8	4.0	3.6
	3.9	3.9	3.7	3.8
	4.2	3.5	3.6	3.9

Haga un análisis de varianza para probar la significancia de los efectos y de la interacción principales.

14.7 Se sabe que la tasa de extracción de cierto polímero depende de la temperatura de reacción y de la cantidad de catalizador empleada. Se hizo un experimento en cuatro niveles de temperatura y cinco niveles de catalizador, y se registró la tasa de extracción en la tabla que sigue:

	Cantidad de catalizador				
	0.5%	**0.6%**	**0.7%**	**0.8%**	**0.9%**
50 °C	38	45	57	59	57
	41	47	59	61	58
60 °C	44	56	70	73	61
	43	57	69	72	58
70 °C	44	56	70	73	61
	47	60	67	61	59
80 °C	49	62	70	62	53
	47	65	55	69	58

Realice un análisis de varianza. Pruebe la significancia de los efectos y de la interacción principales.

14.8 En Myers y Montgomery (2002) se estudia un escenario donde se describe un proceso de laminado por prensado. La respuesta es el espesor del material. Los factores que podrían influir en éste incluyen la cantidad de níquel (A) y el pH (B). Se diseñó un experimento con dos factores. El plan consiste en hacer un diseño completamente aleatorio, en el que las prensas individuales se asignen al azar a las combinaciones de factores. En el experimento se utilizan tres niveles de pH y dos de contenido de níquel. Los espesores en cm $\times 10^{-3}$ son los siguientes:

Contenido de níquel	pH		
(gramos)	**5**	**5.5**	**6**
18	250	211	221
	195	172	150
	188	165	170
10	115	88	69
	165	112	101
	142	108	72

a) Haga la tabla del análisis de varianza con pruebas tanto para los efectos como para la interacción principales. Muestre los valores *P*.

b) Saque conclusiones de ingeniería. ¿Qué fue lo que aprendió usted a partir del análisis de estos datos?

c) Elabore una gráfica que ilustre la presencia o ausencia de interacción.

14.9 Un ingeniero está interesado en el efecto de la velocidad de corte y la geometría de la herramienta sobre las horas de vida de una máquina-herramienta. Se utilizan dos velocidades de corte y dos geometrías distintas. Se llevan a cabo tres pruebas experimentales con cada una de las cuatro combinaciones. Los datos son los siguientes:

Geometría de	Velocidad de corte					
la herramienta	Baja			Alta		
1	22	28	20	34	37	29
2	18	15	16	11	10	10

a) Realice la tabla del análisis de varianza con pruebas sobre los efectos de la interacción y principales.

b) Haga comentarios sobre el efecto que tiene la interacción sobre la prueba de la velocidad de corte.

c) Efectúe pruebas secundarias que permitan al ingeniero aprender el impacto verdadero de la velocidad de corte.

d) Construya una gráfica que ilustre el efecto de la interacción.

14.10 En un experimento se estudiaron dos factores de un proceso de manufactura de un circuito integrado. El propósito del experimento es conocer el efecto sobre la resistividad de las obleas de silicio. Los factores son la dosis del implante (2 niveles) y la posición de la caldera (3 niveles). El experimento es costoso, por lo que sólo se hizo una corrida con cada combinación. Los datos son los siguientes.

Dosis	Posición		
1	15.5	14.8	21.3
2	27.2	24.9	26.1

Se supone que no hay interacción entre esos dos factores.

a) Escriba el modelo y explique sus términos.

b) Muestre la tabla de análisis de varianza.

c) Explique los 2 grados de libertad del "error".

d) Use una prueba de Tukey para hacer pruebas de comparaciones múltiples sobre la posición de la caldera. Explique qué es lo que muestran los resultados.

14.11 Se realizó un estudio para determinar la influencia de dos factores, el método de análisis y el laboratorio que hace el análisis, sobre el nivel de contenido de azufre del carbón. Se asignaron al azar 28 especímenes de carbón a 28 combinaciones de factores, la estructura de las unidades experimentales representada por las combinaciones de siete laboratorios y dos métodos de análisis con dos especímenes por combinación de factores. Los datos se muestran a continuación: la respuesta se expresa en porcentaje de azufre.

	Método			
Laboratorio	**1**		**2**	
1	0.109	0.105	0.105	0.108
2	0.129	0.122	0.127	0.124
3	0.115	0.112	0.109	0.111
4	0.108	0.108	0.117	0.118
5	0.097	0.096	0.110	0.097
6	0.114	0.119	0.116	0.122
7	0.155	0.145	0.164	0.160

Los datos se tomaron de Taguchi, G. "Signal to Noise Ratio and Its Applications to Testing Material", *Reports of Statistical Application Research*, Union of Japanese Scientists and Engineers, vol. 18, núm. 4, 1971.

a) Haga un análisis de varianza y muestre los resultados en la tabla correspondiente.

b) ¿Es significativa la interacción? Si lo es, analice lo que significa para el científico. En sus conclusiones utilice un valor *P*.

c) ¿Son estadísticamente significativos los efectos principales individuales, el laboratorio y el método de análisis? Analice lo que se haya aprendido y base su respuesta en el contexto de cualquier interacción significativa.

d) Haga una gráfica que ilustre el efecto de la interacción.

e) Efectúe una prueba para comparar los métodos 1 y 2 en el laboratorio 1, y haga lo mismo para el laboratorio 7. Comente lo que ilustran tales resultados.

14.12 En un experimento efectuado en el departamento de ingeniería civil del Tecnológico de Virginia, se observó el crecimiento que cierto tipo de alga tenía en el agua, como función del tiempo y la dosis de cobre

que se agregaba al líquido. Los datos se presentan a continuación. La respuesta se expresa en unidades de algas.

	Tiempo en días		
Cobre	**5**	**12**	**18**
1	0.30	0.37	0.25
	0.34	0.36	0.23
	0.32	0.35	0.24
2	0.24	0.30	0.27
	0.23	0.32	0.25
	0.22	0.31	0.25
3	0.20	0.30	0.27
	0.28	0.31	0.29
	0.24	0.30	0.25

a) Haga un análisis de varianza y muestre la tabla correspondiente.

b) Comente acerca de si los datos son suficientes para mostrar un efecto del tiempo sobre la concentración de algas.

c) Haga lo mismo para el contenido de cobre. ¿El contenido de cobre tiene un efecto sobre la concentración de algas?

d) Comente los resultados de la prueba para la interacción. ¿Cómo se ve influido el efecto del contenido de cobre por el tiempo?

14.13 En Myers, *Classical and Modern Regression with Applications*, Duxbury Classic Series, 2a. ed., 1990, se describe un experimento en el cual la Agencia de Protección Ambiental busca determinar el efecto que tienen dos métodos de tratamiento del agua sobre la absorción del magnesio. Se mide los niveles de magnesio, en gramos por centímetro cúbico (cc), y se incorpora dos niveles diferentes de tiempo al experimento. Los datos son los siguientes:

	Tratamiento					
Tiempo (horas)	**1**			**2**		
1	2.19	2.15	2.16	2.03	2.01	2.04
2	2.01	2.03	2.04	1.88	1.86	1.91

a) Haga una gráfica de la interacción. ¿Cuál es su impresión?

b) Efectúe un análisis de varianza y pruebas para los efectos y la interacción principales.

c) Mencione los descubrimientos científicos acerca de cómo influyen el tiempo y el tratamiento en la absorción del magnesio.

d) Ajuste el modelo adecuado de regresión con el tratamiento como variable categórica. En el modelo incluya la interacción.

e) ¿La interacción es significativa en el modelo de regresión?

14.14 Considere los datos del ejercicio 14.12 y responda las siguientes preguntas.

a) Tanto el factor de cobre como el tiempo son de naturaleza cuantitativa. Como resultado, podría ser de interés un modelo de regresión. Describa cuál podría ser un modelo adecuado con x_1 = contenido de cobre y x_2 = tiempo. Ajuste el modelo a los datos, mostrando los coeficientes de regresión y haga una prueba t sobre cada uno.

b) Ajuste el modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon,$$

y compárelo con el que eligió en el inciso *a*). ¿Cuál es más apropiado? Como criterio utilice R^2_{aju}.

14.4 Experimentos con tres factores

En esta sección consideramos un experimento con tres factores, A, B y C, en los niveles a, b y c, respectivamente, en un diseño completamente aleatorio. Suponga de nuevo que se tienen n observaciones para cada una de las abc combinaciones de tratamientos. Debemos proceder a realizar las pruebas de significancia para los tres efectos principales y la interacción implicada. Se espera que el lector sea capaz de usar la descripción dada para generalizar el análisis de $k > 3$ factores.

Modelo para el experimento con tres factores

El modelo para el experimento con tres factores es

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

$$i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b;\ k = 1, 2, \ldots, c; \quad \text{y} \quad l = 1, 2, \ldots, n,$$

donde α_i, β_j y γ_k son los efectos principales; $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ y $(\beta\gamma)_{jk}$ son los efectos de la interacción de dos factores, que tienen la misma interpretación que en el experimento con dos factores. El término $(\alpha\beta\gamma)_{ijk}$ se denomina el **efecto de la interacción de**

tres factores, y representa la no aditividad de las $(\alpha\beta)_{ij}$ sobre los diferentes niveles del factor C. Igual que antes, la suma de todos los efectos principales es igual a cero, y la suma sobre cualesquiera de los subíndices de los efectos de la interacción entre dos y tres factores es igual a cero. En muchas situaciones experimentales tales interacciones de orden superior son insignificantes, y sus medias cuadráticas tan sólo reflejan variación al azar; pero se debe hacer el análisis en su detalle más general.

Otra vez, con la finalidad de que se realicen pruebas válidas de significancia, debe suponerse que los errores son valores de variables aleatorias independientes con distribución normal, cada una con media igual a cero y varianza común σ^2.

La filosofía general con respecto al análisis es la misma que la que se estudió para los experimentos de uno y dos factores. Se hace la partición de la suma de los cuadrados en ocho términos, donde cada uno representa una fuente de variación de los que se obtiene estimadores independientes de σ^2 cuanto todos los efectos principales y de la interacción son iguales a cero. Si los efectos de cualquier factor dado o interacción no son iguales a cero, entonces la media cuadrática estimará la varianza del error más un componente debido al efecto sistemático en cuestión.

Suma de cuadrados para un experimentos de tres factores

$$SSA = bcn\sum_{i=1}^{a}(\bar{y}_{i\ldots} - \bar{y}_{\ldots})^2 \qquad SS(AB) = cn\sum_{i}\sum_{j}(\bar{y}_{ij.} - \bar{y}_{i\ldots} - \bar{y}_{.j.} + \bar{y}_{\ldots})^2$$

$$SSB = acn\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{\ldots})^2 \qquad SS(AC) = bn\sum_{i}\sum_{k}(\bar{y}_{i.k} - \bar{y}_{i\ldots} - \bar{y}_{..k} + \bar{y}_{\ldots})^2$$

$$SSC = abn\sum_{k=1}^{c}(\bar{y}_{..k} - \bar{y}_{\ldots})^2 \qquad SS(BC) = an\sum_{j}\sum_{k}(\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{\ldots})^2$$

$$SS(ABC) = n\sum_{i}\sum_{j}\sum_{k}(\bar{y}_{ijk} - \bar{y}_{ij.} - \bar{y}_{i.k} - \bar{y}_{.jk} + \bar{y}_{i\ldots} + \bar{y}_{.j.} + \bar{y}_{..k} - \bar{y}_{\ldots})^2$$

$$SST = \sum_{i}\sum_{j}\sum_{k}\sum_{l}(y_{ijkl} - \bar{y}_{\ldots})^2 \qquad SSE = \sum_{i}\sum_{j}\sum_{k}\sum_{l}(y_{ijkl} - \bar{y}_{ijk.})^2$$

Aunque se hace énfasis en la interpretación de la salida por computadora con comentarios de esta sección, en vez de enfrascarnos con cálculos laboriosos de sumas de cuadrados, se ofrece lo siguiente como la suma de cuadrados para los tres efectos principales y de las interacciones. Observe la evidente ampliación del problema de dos factores a uno de tres.

Los promedios en las fórmulas se definen como sigue:

$\bar{y}_{\ldots}$ = promedio de todas las $abcn$ observaciones,
$\bar{y}_{i\ldots}$ = promedio de las observaciones para el i-ésimo nivel del factor A,
$\bar{y}_{.j.}$ = promedio de las observaciones para el j-ésimo nivel del factor B,
$\bar{y}_{..k}$ = promedio de las observaciones para el k-ésimo nivel del factor C,
$\bar{y}_{ij.}$ = promedio de las observaciones para el i-ésimo nivel de A y el j-ésimo nivel de B,
$\bar{y}_{i.k}$ = promedio de las observaciones para el i-ésimo nivel de A y el k-ésimo nivel de C,
$\bar{y}_{.jk}$ = promedio de las observaciones para el j-ésimo nivel de B y el k-ésimo nivel de C,
$\bar{y}_{ijk}$ = promedio de las observaciones para la (ijk)-ésima combinación de tratamientos.

Los cálculos en la tabla del análisis de varianza para un problema de tres factores con n corridas de repetición en cada combinación de factores, se resumen en la tabla 14.7.

Tabla 14.7: ANOVA para el experimento de tres factores con n repeticiones

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado de la media cuadrática**	***f* calculada**
Efecto principal:				
A	SSA	$a-1$	s_1^2	$f_1 = \frac{s_1^2}{s^2}$
B	SSB	$b-1$	s_2^2	$f_2 = \frac{s_2^2}{s^2}$
C	SSC	$c-1$	s_3^2	$f_3 = \frac{s_3^2}{s^2}$
Interacción de dos factores:				
AB	$SS(AB)$	$(a-1)(b-1)$	s_4^2	$f_4 = \frac{s_4^2}{s^2}$
AC	$SS(AC)$	$(a-1)(c-1)$	s_5^2	$f_5 = \frac{s_5^2}{s^2}$
BC	$SS(BC)$	$(b-1)(c-1)$	s_6^2	$f_6 = \frac{s_6^2}{s^2}$
Interacción de tres factores:				
ABC	$SS(ABC)$	$(a-1)(b-1)(c-1)$	s_7^2	$f_7 = \frac{s_7^2}{s^2}$
Error	SSE	$abc\,(n-1)$	s^2	
Total	SST	$abcn-1$		

Para el experimento de tres factores con una sola corrida experimental por combinación, debe usarse el análisis de la tabla 14.7 con $n = 1$ y el empleo de la suma de cuadrados de la interacción ABC para SSE. En este caso, suponemos que los efectos de la interacción $(\alpha\beta\gamma)_{ijk}$ son todos iguales a cero, de modo que

$$E\left[\frac{SS(ABC)}{(a-1)(b-1)(c-1)}\right] = \sigma^2 + \frac{n}{(a-1)(b-1)(c-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\alpha\beta\gamma)_{ijk}^2 = \sigma^2.$$

Es decir, $SS(ABC)$ representa la variación de que sólo se debe al error experimental. Su media cuadrática proporciona así un estimador insesgado de la varianza del error. Con $n = 1$ y $SSE = SS(ABC)$, la suma de los errores al cuadrado se obtiene al restar a la suma total de cuadrados, la suma de cuadrados de los efectos principales y las interacciones de dos factores.

Ejemplo 14.4: En la producción de un material en particular hay tres variables de interés: A, el efecto del operador (tres operadores): B, el catalizador utilizado en el experimento (tres catalizadores); y C, el tiempo de lavado del producto después del proceso de enfriamiento (15 minutos y 20 minutos). Con cada combinación de factores se hicieron tres corridas. Se pensaba que deberían estudiarse todas las interacciones entre los factores. En la tabla 14.8 se presentan las salidas codificadas. Ejecute un análisis de varianza para probar los efectos que son significativos.

Solución: La tabla 14.9 muestra el análisis de varianza de los datos. Ninguna de las interacciones muestra un efecto significativo en el nivel $\alpha = 0.05$. Sin embargo, el valor P para BC es 0.0610; por ello, debe ignorarse. Los efectos de los operadores y el catalizador son significativos, en tanto que el del tiempo de lavado no lo es.

Tabla 14.8: Datos para el ejemplo 14.4

	Tiempo de lavado, C					
	15 minutos			**20 minutos**		
	B (catalizador)			**B (catalizador)**		
A (operador)	**1**	**2**	**3**	**1**	**2**	**3**
1	10.7	10.3	11.2	10.9	10.5	12.2
	10.8	10.2	11.6	12.1	11.1	11.7
	11.3	10.5	12.0	11.5	10.3	11.0
2	11.4	10.2	10.7	9.8	12.6	10.8
	11.8	10.9	10.5	11.3	7.5	10.2
	11.5	10.5	10.2	10.9	9.9	11.5
3	13.6	12.0	11.1	10.7	10.2	11.9
	14.1	11.6	11.0	11.7	11.5	11.6
	14.5	11.5	11.5	12.7	10.9	12.2

Tabla 14.9: ANOVA para un experimento de tres factores en un diseño aleatorio por completo

Fuente	**df**	**Suma de cuadrados**	**Media cuadrática**	**Valor F**	**Valor P**
A	2	13.98	6.99	11.64	0.0001
B	2	10.18	5.09	8.48	0.0010
AB	4	4.77	1.19	1.99	0.1172
C	1	1.19	1.19	1.97	0.1686
AC	2	2.91	1.46	2.43	0.1027
BC	2	3.63	1.82	3.03	0.0610
ABC	4	4.91	1.23	2.04	0.1089
Error	36	21.61	0.60		
Total	53	63.19			

Efecto de la interacción BC

Deben hacerse más análisis respecto del ejemplo 14.4, en particular acerca del efecto que la interacción entre el catalizador y el tiempo de lavado tienen sobre la prueba del efecto principal del tiempo de lavado (factor C). Recuerde nuestro análisis de la sección 14.2. Se dieron ilustraciones de la manera en que la presencia de la interacción podría cambiar la interpretación que se hizo de los efectos principales. En el ejemplo 14.4, la interacción BC es significativa al nivel de 0.06, aproximadamente. No obstante, suponga que se observa una tabla de medias de dos factores, como la 14.10.

Queda claro por qué se encontró que el tiempo de lavado no era significativo. Un analista poco cuidadoso se quedaría con la impresión de que el tiempo de lavado podría eliminarse de cualquier estudio futuro donde se midiera la salida. Sin embargo, es notorio cómo cambia el efecto del tiempo de lavado de uno negativo para el primer catalizador, a lo que parece ser otro positivo para el tercer catalizador. Si tan

Tabla 14.10: Tabla de medias de dos factores para el ejemplo 14.4

	Tiempo de lavado C	
Catalizador, B	**15 min**	**20 min**
1	12.19	11.29
2	10.86	10.50
3	11.09	11.46
Medias	11.38	11.08

sólo nos centráramos en los datos para el catalizador 1, una comparación simple entre las medias de los dos tiempos de lavado produciría un estadístico t sencillo:

$$t = \frac{12.19 - 11.29}{\sqrt{0.6(2/9)}} = 2.5,$$

que es significativo a un nivel menor que 0.02. Así, bien puede ignorarse un importante efecto negativo del tiempo de lavado para el catalizador 1, si el analista hace la interpretación amplia incorrecta de la razón F insignificante del tiempo de lavado.

Agrupamiento en modelos multifactoriales

Se ha descrito el modelo de tres factores y su análisis, en la forma más general, con la inclusión en el modelo de todas las interacciones posibles. Por supuesto, hay muchas situaciones en las que *a priori* se conoce que el modelo no debería contener ciertas interacciones. Puede sacarse ventaja de este conocimiento al combinar o agrupar las sumas de cuadrados correspondientes a interacciones despreciables con la suma de los errores al cuadrado, para formar un nuevo estimador de σ^2 con un número más grande de grados de libertad. Por ejemplo, en un experimento de metalurgia diseñado para estudiar el efecto sobre el espesor de la película de tres variables importantes del proceso, suponga que se conoce que el factor A, la concentración de ácido, no interactúa con los factores B y C. Las sumas de cuadrados SS_A, SS_B, SS_C y $SS(BC)$ se calculan usando los métodos descritos en un apartado anterior de esta sección. Todas las medias cuadráticas para los efectos restantes ahora estimarán la varianza del error σ^2. Por lo tanto, el nuevo **error cuadrático medio se forma agrupando** $SS(AB)$, $SS(AC)$, $SS(ABC)$ y SS_E, junto con los grados de libertad correspondientes. El denominador que resulta para las pruebas de significancia es, entonces, el error cuadrático medio dado por

$$s^2 = \frac{SS(AB) + SS(AC) + SS(ABC) + SS_E}{(a-1)(b-1) + (a-1)(c-1) + (a-1)(b-1)(c-1) + abc(n-1)}.$$

Por supuesto, al calcular se obtiene por sustracción la suma agrupada de cuadrados y los grados de libertad agrupados, una vez que se calculan SS_T y las sumas de cuadrados para los efectos existentes. La tabla del análisis de varianza adoptaría así la forma de la tabla 14.11.

Experimentos factoriales en bloques

En este capítulo se ha supuesto que el diseño experimental utilizado es uno aleatorio por completo. Al interpretar los niveles del factor A en la tabla 14.11, **como**

Tabla 14.11: ANOVA sin interacción del factor A

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	f calculada
Efecto principal:				
A	SSA	$a-1$	s_1^2	$f_1 = \frac{s_1^2}{s^2}$
B	SSB	$b-1$	s_2^2	$f_2 = \frac{s_2^2}{s^2}$
C	SSC	$c-1$	s_3^2	$f_3 = \frac{s_3^2}{s^2}$
Interacción de dos factores:				
BC	$SS(BC)$	$(b-1)(c-1)$	s_4^2	$f_4 = \frac{s_4^2}{s^2}$
Error	SSE	Resta	s^2	
Total	SST	$abcn-1$		

bloques diferentes, entonces se tiene el procedimiento del análisis de varianza para un experimento de dos factores en un diseño de bloques aleatorios. Por ejemplo, si se interpretan los operadores del ejemplo 14.4 como bloques, y se supone que no hay interacción entre éstos y los otros dos factores, el análisis de varianza adopta la forma de la tabla 14.12, en vez de la tabla 14.9. El lector puede verificar que el error cuadrático medio también es

$$s^2 = \frac{4.77 + 2.91 + 4.91 + 21.61}{4 + 2 + 4 + 36} = 0.74,$$

lo cual demuestra el agrupamiento de las sumas de los cuadrados para los efectos de la interacción inexistente. Observe que el factor B, el catalizador, tiene un efecto significativo sobre el producto.

Tabla 14.12: ANOVA para un experimento de dos factores en un diseño de bloques aleatorios

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	f calculada	Valor P
Bloques	13.98	2	6.99		
Efecto principal:					
B	10.18	2	5.09	6.88	0.0024
C	1.18	1	1.18	1.59	0.2130
Interacción de dos factores:					
BC	3.64	2	1.82	2.46	0.0966
Error	31.21	46	0.74		
Total	63.19	53			

Ejemplo 14.5: Se realizó un experimento para determinar el efecto de la temperatura, la presión y la intensidad de agitación sobre la tasa de filtración del producto. Esto se hizo en una planta piloto. El experimento se corrió en dos niveles de cada factor. Además, se decidió que debían utilizarse dos lotes de materia prima, los cuales fueron tratados

Tabla 14.13: Datos para el ejemplo 14.5

Lote 1					
	Tasa de agitación baja			**Tasa de agitación alta**	
Temp.	**Presión L**	**Presión H**	**Temp.**	**Presión L**	**Presión H**
L	43	49	**L**	44	47
H	64	68	**H**	97	102
Lote 2					
	Tasa de agitación baja			**Tasa de agitación alta**	
Temp.	**Presión L**	**Presión H**	**Temp.**	**Presión L**	**Presión H**
L	49	57	**L**	51	55
H	70	76	**H**	103	106

como bloques. Se hicieron ocho corridas experimentales en orden aleatorio para cada lote de materia prima. Se piensa que todas las interacciones de los dos factores son de interés. No se supone que haya interacciones con los lotes. Los datos aparecen en la tabla 14.13. Las letras "L" y "H" significan niveles bajo y alto, respectivamente. La tasa de filtración se expresa en galones por hora.

a) Muestre la tabla ANOVA completa. Agrupe en el error todas las "interacciones" con los bloques.

b) ¿Cuáles interacciones parecen ser significativas?

c) Construya gráficas que revelen las interacciones significativas e interprételas. Explique lo que significa la gráfica para el ingeniero.

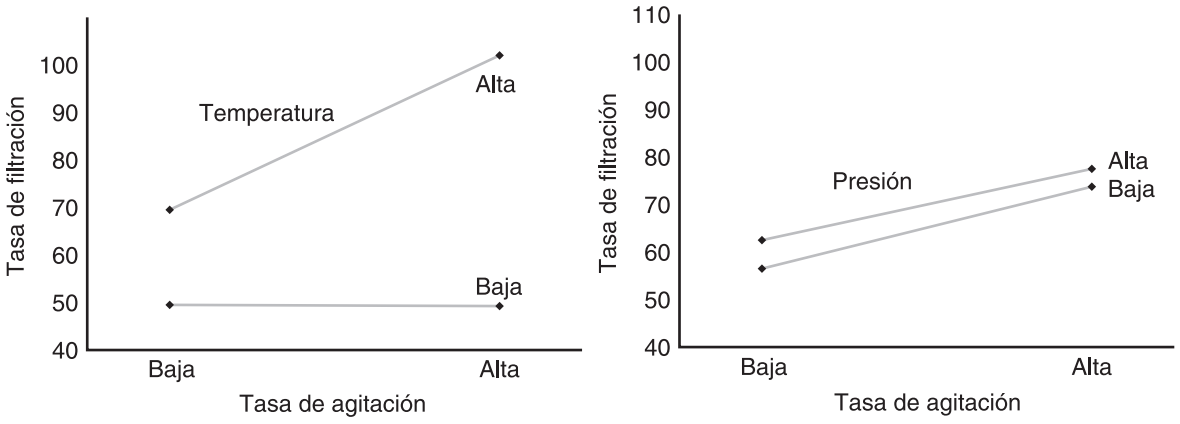
Solución: *a*) En la figura 14.7 se presenta la salida de *SAS*.

b) Como se aprecia en la figura 14.7, la interacción de la temperatura con la tasa de agitación (*strate*) parece ser muy significativa. Asimismo, la interacción de la presión con la tasa de agitación también parece ser significativa. A propósito, si se fueran a hacer más agrupamientos al combinar las interacciones insignificantes con el error, las conclusiones serían las mismas, y el valor P para la interacción de la presión con la tasa de agitación se volvería mayor: sería de 0.0517.

c) Los efectos principales tanto para la tasa de agitación y la temperatura son muy significativos, como se aprecia en la figura 14.7. El estudio de la gráfica de interacción de la figura 14.8*a*) muestra que el efecto de la tasa de agitación depende del nivel de la temperatura. Con el nivel bajo, el efecto de la tasa de agitación es despreciable; mientras que con el nivel alto la tasa de agitación tiene un efecto positivo fuerte sobre la tasa media de filtración. Para la figura 14.8*b*), la interacción entre la presión y la tasa de agitación, aunque no sea tan pronunciada como la de la figura 14.8*a*), aún muestra una inconsistencia ligera del efecto de la tasa de agitación a través de la presión.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
batch	1	175.562500	175.562500	177.14	<.0001
pressure	1	95.062500	95.062500	95.92	<.0001
temp	1	5292.562500	5292.562500	5340.24	<.0001
pressure*temp	1	0.562500	0.562500	0.57	0.4758
strate	1	1040.062500	1040.062500	1049.43	<.0001
pressure*strate	1	5.062500	5.062500	5.11	0.0583
temp*strate	1	1072.562500	1072.562500	1082.23	<.0001
pressure*temp*strate	1	1.562500	1.562500	1.58	0.2495
Error	7	6.937500	0.991071		
Corrected Total	15	7689.937500			

Figura 14.7: ANOVA para el ejemplo 14.5, interacción del lote agrupado con el error.



a) Temperatura contra tasa de agitación.

b) Presión contra tasa de agitación.

Figura 14.8: Gráficas de interacción para el ejemplo 14.5.

Ejercicios

14.15 Los datos siguientes se tomaron de un estudio sobre mediciones. Se hizo un experimento usando 3 factores, A, B y C, todos efectos fijos.

	$\mathbf{C_1}$			$\mathbf{C_2}$			$\mathbf{C_3}$		
	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
A_1	15.0	14.8	15.9	16.8	14.2	13.2	15.8	15.5	19.2
	18.5	13.6	14.8	15.4	12.9	11.6	14.3	13.7	13.5
	22.1	12.2	13.6	14.3	13.0	10.1	13.0	12.6	11.1
A_2	11.3	17.2	16.1	18.9	15.4	12.4	12.7	17.3	7.8
	14.6	15.5	14.7	17.3	17.0	13.6	14.2	15.8	11.5
	18.2	14.2	13.4	16.1	18.6	15.2	15.9	14.6	12.2

a) Haga pruebas de significancia sobre todas las interacciones con el nivel de $\alpha = 0.05$.

b) Realice pruebas de significancia sobre los efectos principales con el nivel de $\alpha = 0.05$.

c) Dé una explicación de la forma en que una interacción significativa enmascara el efecto del factor C.

14.16 Considere una situación experimental que implique los factores A, B y C, en los que se adopta un modelo de tres factores de efectos fijos, de la forma

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + \epsilon_{ijkl}.$$

Se considera que todas las demás interacciones no existen o son despreciables. Los datos se presentan enseguida.

a) Haga una prueba de significancia sobre la interacción BC con el nivel de $\alpha = 0.05$.

b) Desarrolle pruebas de significancia sobre los efectos principales A, B y C, usando un error cuadrático medio agrupado, con un nivel $\alpha = 0.05$.

	B_1			B_2		
	C_1	C_2	C_3	C_1	C_2	C_3
A_1:	4.0	3.4	3.9	4.4	3.1	3.1
	4.9	4.1	4.3	3.4	3.5	3.7
A_2:	3.6	2.8	3.1	2.7	2.9	3.7
	3.9	3.2	3.5	3.0	3.2	4.2
A_3:	4.8	3.3	3.6	3.6	2.9	2.9
	3.7	3.8	4.2	3.8	3.3	3.5
A_4:	3.6	3.2	3.2	2.2	2.9	3.6
	3.9	2.8	3.4	3.5	3.2	4.3

14.17 La fatiga por corrosión de los metales se ha definido como la acción simultánea de tensión cíclica y ataque químico sobre una estructura metálica. En el estudio *Effect of Humidity and Several Surface Coatings on the Fatigue Life of 2024-T351 Aluminum Alloy*, realizado por el Departamento de Ingeniería Mecánica del Instituto Politécnico y Universidad Estatal de Virginia, se utilizó una técnica que requería la aplicación de un recubrimiento protector de cromato, con la finalidad de minimizar el daño de la fatiga por corrosión en el aluminio. En la investigación se emplearon 3 factores con 5 repeticiones para cada combinación de tratamientos: recubrimiento, en 2 niveles, y humedad y esfuerzo cortante, ambos con 3 niveles. Los datos de fatiga, expresados en miles de ciclos antes del fallo, se presentan a continuación.

a) Realice un análisis de varianza con $\alpha = 0.01$ para probar la significancia de los efectos principales y de interacción.

b) Haga una recomendación para las combinaciones de los 3 factores que harían que sea bajo el daño por fatiga.

Recubrimiento	Humedad	Esfuerzo constante (psi) 13000	17000	20000
Sin recubrimiento	Bajo: (20-25% RH)	4580	5252	361
		10126	897	466
		1341	1465	1069
		6414	2694	469
		3549	1017	937
	Medio: (50-60% RH)	2858	799	314
		8829	3471	244
		10914	685	261
		4067	810	522
		2595	3409	739
	Alto: (86-91% RH)	6489	1862	1344
		5248	2710	1027
		6816	2632	663
		5860	2131	1216
		5901	2470	1097
Cromado	Bajo: (20-25% RH)	5395	4035	130
		2768	2022	841
		1821	914	1595
		3604	2036	1482
		4106	3524	529
	Medio: (50-60% RH)	4833	1847	252
		7414	1684	105
		10022	3042	847
		7463	4482	874
		21906	996	755
	Alto: (86-91% RH)	3287	1319	586
		5200	929	402
		5493	1263	846
		4145	2236	524
		3336	1392	751

14.18 El método de fluorescencia por rayos X es una herramienta analítica importante para determinar la concentración de material en los propulsores sólidos para misiles. En el artículo *An X-ray Fluorescence Method for Analyzing Polybutadiene Acrylic Acid* (PBAA) *Propellants, Quarterly Report*, RK-TR-62-1, Army Ordinance Missile Command (1962), se afirma que el proceso de mezcla del propulsor y el tiempo de análisis influyen en la homogeneidad del material y, por ello, en la exactitud de las mediciones de la intensidad de los rayos X. Se hizo un experimento con 3 factores: A, las condiciones de mezcla (cuatro niveles); B, el tiempo de análisis (dos niveles); y C, el método de carga del propulsor en los recipientes para tomar muestras (temperatura elevada y ambiental). Se obtuvieron los datos siguientes, que representan el análisis en peso porcentual del perclorato de amoniaco en un propulsor dado.

	Método de carga, C			
	Caliente		Temp. ambiente	
	B		B	
A	1	2	1	2
1	38.62	38.45	39.82	39.82
	37.20	38.64	39.15	40.26
	38.02	38.75	39.78	39.72
2	37.67	37.81	39.53	39.56
	37.57	37.75	39.76	39.25
	37.85	37.91	39.90	39.04
3	37.51	37.21	39.34	39.74
	37.74	37.42	39.60	39.49
	37.58	37.79	39.62	39.45
4	37.52	37.60	40.09	39.36
	37.15	37.55	39.63	39.38
	37.51	37.91	39.67	39.00

a) Lleve a cabo un análisis de varianza con $\alpha = 0.01$ con la finalidad de probar la significancia de los efectos principales y de interacción.

b) Analice la influencia de los tres factores sobre el peso porcentual del perclorato de amoniaco.

Introduzca en su análisis el papel de cualquier interacción significativa.

14.19 Las copiadoras electrónicas funcionan adhiriendo tinta negra al papel mediante electricidad estática. La etapa final del proceso de copiado comprende el calentamiento y adhesión de la tinta sobre el papel. La potencia de la adhesión durante este proceso final determina la calidad de la copia. Es posible que la temperatura, estado superficial de la adhesión en el rodillo y la dureza del rodillo de la prensa, influyan en la potencia de la adhesión de la copiadora. Se hizo un experimento con tratamientos, que consistían en una combinación de estos tres factores en cada uno de tres niveles. Los datos siguientes muestran la potencia de la adhesión para cada combinación de tratamientos. Lleve a cabo un análisis de varianza con $\alpha = 0.05$ para probar la significancia de los efectos principales y de la interacción.

	Estado superficial de la adhesión en el rodillo	**Dureza del rodillo de la prensa**					
		20		**40**		**60**	
Temp. baja	**Suave:**	0.52	0.44	0.54	0.52	0.60	0.55
		0.57	0.53	0.65	0.56	0.78	0.68
	Media:	0.64	0.59	0.79	0.73	0.49	0.48
		0.58	0.64	0.79	0.78	0.74	0.50
	Dura:	0.67	0.77	0.58	0.68	0.55	0.65
		0.74	0.65	0.57	0.59	0.57	0.58
Temp. media	**Suave:**	0.46	0.40	0.31	0.49	0.56	0.42
		0.58	0.37	0.48	0.66	0.49	0.49
	Media:	0.60	0.43	0.66	0.57	0.64	0.54
		0.62	0.61	0.72	0.56	0.74	0.56
	Dura:	0.53	0.65	0.53	0.45	0.56	0.66
		0.66	0.56	0.59	0.47	0.71	0.67
Temp. alta	**Suave:**	0.52	0.44	0.54	0.52	0.65	0.49
		0.57	0.53	0.65	0.56	0.65	0.52
	Media:	0.53	0.65	0.53	0.45	0.49	0.48
		0.66	0.56	0.59	0.47	0.74	0.50
	Dura:	0.43	0.43	0.48	0.31	0.55	0.65
		0.47	0.44	0.43	0.27	0.57	0.58

14.20 Para un estudio de la dureza de los empastes dentales de oro, se eligieron cinco dentistas al azar y se les asignó a combinaciones de tres métodos de condensación y dos tipos de oro. Se midió la dureza. [véase Hoaglin, Mosteller y Tukey (1991).] Los datos se presentan a continuación. Haga que los dentistas jueguen el papel de bloques.

a) Proponga el modelo adecuado con las suposiciones.

b) ¿Hay interacción significativa entre el método de condensación y el tipo de material de empaste de oro?

c) ¿Hay un método de condensación que parezca mejor? Explique su respuesta.

Dentista bloque	**Método**	**Tipo**	
		Lámina dorada	**Goldent**
1	1	792	824
	2	772	772
	3	782	803
2	1	803	803
	2	752	772
	3	715	707
3	1	715	724
	2	792	715
	3	762	606
4	1	673	946
	2	657	743
	3	690	245
5	1	634	715
	2	649	724
	3	724	627

14.21 Considere combinaciones de tres factores en el retiro de la suciedad de cargas estándar de lavandería. El primer factor es la marca del detergente: X, Y o Z. El segundo factor es el tipo de detergente: líquido o en polvo. El tercer factor es la temperatura del agua, caliente o tibia. El experimento se repitió tres veces. La respuesta está expresada en la remoción porcentual de la suciedad. Los datos son los siguientes:

Marca	**Tipo**	**Temperatura**			
X	En polvo	Caliente	85	88	80
		Tibia	82	83	85
	Líquido	Caliente	78	75	72
		Tibia	75	75	73
Y	En polvo	Caliente	90	92	92
		Tibia	88	86	88
	Líquido	Caliente	78	76	70
		Tibia	76	77	76
Z	En polvo	Caliente	85	87	88
		Tibia	76	74	78
	Líquido	Caliente	60	70	68
		Tibia	55	57	54

a) ¿Existen efectos significativos de la interacción, con el nivel de $\alpha = 0.05$?

b) ¿Hay diferencias significativas entre las tres marcas del detergente?

c) ¿Cuál combinación de factores preferiría utilizar?

14.22 Un científico recaba datos experimentales sobre el radio de un grano de combustible propulsor, y, como función de la temperatura del polvo, la tasa de extrusión y la temperatura del molde. Los tres factores del experimento son los siguientes:

	Temp. del polvo			
	Temp. del molde		Temp. del molde	
	150		190	
Tasa	220	250	220	250
12	82	124	88	129
24	114	157	121	164

No se dispone de recursos para hacer experimentos repetidos con las ocho combinaciones de factores. Se cree que la tasa de extrusión no interactúa con la temperatura del molde, y que la interacción entre los tres factores es despreciable. Así, esas dos interacciones pueden agruparse para producir un término de "error" de dos grados de libertad.

a) Haga un análisis de varianza que incluya los tres efectos principales e interacciones de dos factores. Determine cuáles efectos influyen en el radio del grano de combustible.

b) Construya gráficas de interacción para la temperatura del polvo usando la temperatura del molde y del polvo, mediante las interacciones de la tasa de extrusión.

c) Comente acerca de la consistencia entre la apariencia de las gráficas de interacción y las pruebas sobre las dos interacciones en el ANOVA.

14.23 En el libro *Design of Experiments for Quality Improvement*, publicado por la Asociación Japonesa de Estándares (1989), se reporta un estudio sobre la extracción de polietileno por medio de un solvente, y la manera en que la cantidad de gel (proporción) se ve influida por tres factores: el tipo de solvente, la temperatura de extracción y el tiempo de extracción. Se diseñó un experimento factorial y se obtuvieron los datos siguientes, expresados en proporción de gel.

		Tiempo					
Solvente	**Temp.**	**4**		**8**		**16**	
Etanol	**120**	94.0,	94.0	93.8,	94.2	91.1,	90.5
	80	95.3,	95.1	94.9,	95.3	92.5,	92.4
Tolueno	**120**	94.6,	94.5	93.6,	94.1	91.1,	91.0
	80	95.4,	95.4	95.6,	96.0	92.1,	92.1

a) Haga un análisis de varianza y determine cuáles factores e interacciones influyen en la proporción de gel.

b) Construya una gráfica de la interacción entre dos factores cualesquiera que sea significativa. Además, explique qué conclusión podría obtenerse de la presencia de la interacción.

c) Haga una gráfica de probabilidad normal de los residuos y coméntela.

14.24 Considere el conjunto de datos del ejercicio 14.19.

a) Construya una gráfica de la interacción de dos factores cualesquiera que sea significativa.

b) Haga una gráfica de probabilidad normal de residuos y coméntela.

14.5 Experimentos factoriales de modelos II y III

En un experimento de dos factores con efectos aleatorios, se tiene el modelo II:

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk},$$

para $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; y $k = 1, 2, \ldots, n$, donde A_i, B_j (AB_{ij}) y ϵ_{ijk} son variables aleatorias independientes con medias igual a cero y varianzas σ_α^2, σ_β^2, $\sigma_{\alpha\beta}^2$, y σ^2, respectivamente. La suma de cuadrados para experimentos del modelo II se calculan exactamente de la misma forma que para los del modelo I. Ahora se tiene interés en probar hipótesis de la forma

$$H_0^{'}\colon\ \sigma_\alpha^2 = 0, \qquad H_0^{''}\colon\ \sigma_\beta^2 = 0, \qquad H_0^{'''}\colon\ \sigma_{\alpha\beta}^2 = 0,$$
$$H_1^{'}\colon\ \sigma_\alpha^2 \neq 0, \qquad H_1^{''}\colon\ \sigma_\beta^2 \neq 0, \qquad H_1^{'''}\colon\ \sigma_{\alpha\beta}^2 \neq 0,$$

donde el denominador en la razón f no es necesariamente el error cuadrático medio. El denominador apropiado se determina al examinar los valores esperados de las distintas medias cuadráticas. Éstas se muestran en la tabla 14.14.

De la tabla 14.14, se observa que $H_0^{'}$ y $H_0^{''}$ se prueban usando s_3^2 en el denominador de la razón f; mientras que $H_0^{'''}$ se prueba con s^2 en el denominador. Los estimadores insesgados de los componentes de la varianza son

$$\hat{\sigma}^2 = s^2, \qquad \hat{\sigma}_{\alpha\beta}^2 = \frac{s_3^2 - s^2}{n}, \qquad \hat{\sigma}_\alpha^2 = \frac{s_1^2 - s_3^2}{bn}, \qquad \hat{\sigma}_\beta^2 = \frac{s_2^2 - s_3^2}{an}.$$

Tabla 14.14: Medias cuadráticas esperadas para un experimento de dos factores del modelo II

Fuente de variación	**Grados de libertad**	**Media cuadrática**	**Media cuadrática esperada**
A	$a-1$	s_1^2	$\sigma^2 + n\sigma^2_{\alpha\beta} + bn\sigma^2_{\alpha}$
B	$b-1$	s_2^2	$\sigma^2 + n\sigma^2_{\alpha\beta} + an\sigma^2_{\beta}$
AB	$(a-1)(b-1)$	s_3^2	$\sigma^2 + n\sigma^2_{\alpha\beta}$
Error	$ab(n-1)$	s^2	σ^2
Total	$abn-1$		

Tabla 14.15: Medias cuadráticas esperadas para un experimento con tres factores del modelo II

Fuente de variación	**Grados de libertad**	**Media cuadrática**	**Media cuadrática esperada**
A	$a-1$	s_1^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + cn\sigma^2_{\alpha\beta} + bn\sigma^2_{\alpha\gamma} + bcn\sigma^2_{\alpha}$
B	$b-1$	s_2^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + cn\sigma^2_{\alpha\beta} + an\sigma^2_{\beta\gamma} + acn\sigma^2_{\beta}$
C	$c-1$	s_3^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + bn\sigma^2_{\alpha\gamma} + an\sigma^2_{\beta\gamma} + abn\sigma^2_{\gamma}$
AB	$(a-1)(b-1)$	s_4^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + cn\sigma^2_{\alpha\beta}$
AC	$(a-1)(c-1)$	s_5^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + bn\sigma^2_{\alpha\gamma}$
BC	$(b-1)(c-1)$	s_6^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + an\sigma^2_{\beta\gamma}$
ABC	$(a-1)(b-1)(c-1)$	s_7^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma}$
Error	$abc(n-1)$	s^2	σ^2
Total	$abcn-1$		

En la tabla 14.15 se presentan las medias cuadráticas esperadas para el experimento de tres factores con efectos aleatorios en un diseño aleatorio por completo. A partir de las medias cuadráticas esperadas de la tabla 14.15, es evidente que se pueden formar razones f adecuadas para probar todos los componentes de la varianza de la interacción de dos y tres factores. Sin embargo, para probar una hipótesis de la forma

$$H_0\colon\ \sigma_\alpha^2 = 0,$$
$$H_1\colon\ \sigma_\alpha^2 \neq 0,$$

pareciera que no hay razón f apropiada, a menos que se encontrara que no es significativa una o más de los componentes de la varianza de la interacción de dos factores. Por ejemplo, suponga que se hubiera comparado s_5^2 (media cuadrática AC) con s_7^2 (media cuadrática ABC) y se encontrara que $\sigma^2_{\alpha\gamma}$ es despreciable. Podría argumentarse que el término $\sigma^2_{\alpha\gamma}$ debería eliminarse de todas las medias cuadráticas esperadas de la tabla 14.15; entonces, la razón s_1^2/s_4^2 ofrece una prueba de la significancia del componente σ_α^2 de la varianza. Por lo tanto, si se prueba la hipótesis concerniente a los componentes de la varianza de los efectos principales, en primer lugar es necesario investigar la significancia de los componentes de la interacción de dos factores. Una prueba aproximada derivada por Satterthwaite (véase la bibliografía), se utiliza cuando se encuentra que son significativos ciertos componentes de la varianza de la interacción de dos factores, por lo que deben permanecer como parte de la media cuadrática esperada.

Ejemplo 14.6: En un estudio para determinar cuáles son las fuentes importantes de la variación en un proceso industrial, se toman tres mediciones de la respuesta para 3 operadores elegidos al azar, y se toman en forma aleatoria 4 lotes de materia prima. Se decidió que debe hacerse una prueba significativa con un nivel de significancia de 0.05 para determinar si son significativos los componentes de la varianza debidos a los lotes, los operadores y la interacción. Además, tienen que calcularse los estimadores de los componentes de la varianza. En la tabla 14.16 se presentan los datos con la respuesta expresada en porcentaje de peso:

Tabla 14.16: Datos para el ejemplo 14.6

	Lote			
Operador	**1**	**2**	**3**	**4**
1	66.9	68.3	69.0	69.3
	68.1	67.4	69.8	70.9
	67.2	67.7	67.5	71.4
2	66.3	68.1	69.7	69.4
	65.4	66.9	68.8	69.6
	65.8	67.6	69.2	70.0
3	65.6	66.0	67.1	67.9
	66.3	66.9	66.2	68.4
	65.2	67.3	67.4	68.7

Solución: Las sumas de los cuadrados se encuentran en la forma habitual, con los resultados siguientes:

$$\begin{aligned} &SST \text{ (total)} && = 84.5564, && SSE \text{ (error)} = 10.6733,\\ &SSA \text{ (operadores)} && = 18.2106, && SSB \text{ (lotes)} = 50.1564,\\ &SS(AB) \text{ (interacción)} && = 5.5161. \end{aligned}$$

Se obtuvieron todos los demás cálculos y se presentan en la tabla 14.17. Como

$$f_{0.05}(2,6) = 5.14, \qquad f_{0.05}(3,6) = 4.76, \qquad \text{y} \qquad f_{0.05}(6,24) = 2.51,$$

se encuentra que son significativos los componentes de la varianza de los operadores y el lote. Aunque la varianza de la interacción no es significativa con un nivel $\alpha = 0.05$, el valor P es de 0.095. Los estimadores de los componentes de la varianza del efecto principal son

$$\hat{\sigma}_\alpha^2 = \frac{9.1053 - 0.9194}{12} = 0.68, \qquad \hat{\sigma}_\beta^2 = \frac{16.7188 - 0.9144}{9} = 1.76.$$

Experimento del modelo III (modelo mixto)

Hay situaciones en que el experimento dicta la suposición de un **modelo mixto** (es decir, una mezcla de efectos aleatorios y fijos). Por ejemplo, para el caso de dos factores se tiene que

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk},$$

Tabla 14.17: Análisis de la varianza para el ejemplo 14.6

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	f calculada
Operadores	18.2106	24	9.1053	9.90
Lotes	50.1564	3	16.7188	18.18
Interacción	5.5161	6	0.9194	2.07
Error	10.6733	24	0.4447	
Total	84.5564	35		

para $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; $k = 1, 2, \ldots, n$. Las A_i pueden ser variables aleatorias independientes de ϵ_{ijk}, y las B_j son de efectos fijos. La naturaleza mixta del modelo requiere que los términos de la interacción sean variables aleatorias. Como resultado, las hipótesis relevantes son de la forma

$$H_0': \ \sigma_\alpha^2 = 0, \quad H_0'': \ B_1 = B_2 = \cdots = B_b = 0 \qquad H_0''': \ \sigma_{\alpha\beta}^2 = 0,$$
$$H_1': \ \sigma_\alpha^2 \neq 0, \quad H_1'': \ \text{Al menos una de las } B_j \text{ no es igual a cero} \qquad H_1''': \ \sigma_{\alpha\beta}^2 \neq 0.$$

Otra vez, los cálculos de la suma de cuadrados son idénticos a los de las situaciones fija y modelo II, y las pruebas f las dictan las medias cuadráticas esperadas. La tabla 14.18 proporciona las medias cuadráticas esperadas para el problema de dos factores del modelo III.

Tabla 14.18. Medias cuadráticas esperadas para el experimento con dos factores del modelo III

Factor	Media cuadrática esperada
A (aleatorios)	$\sigma^2 + bn\sigma_\alpha^2$
B (fijos)	$\sigma^2 + n\sigma_{\alpha\beta}^2 + \frac{an}{b-1}\sum_j B_j^2$
AB (aleatorios)	$\sigma^2 + n\sigma_{\alpha\beta}^2$
Error	σ^2

Debido a la naturaleza de las medias cuadráticas esperadas, queda claro que la **prueba sobre el efecto aleatorio emplea el error cuadrático medio s^2** como denominador; mientras que la **prueba sobre el efecto fijo** utiliza la interacción de la media cuadrática. Suponga que ahora se consideran tres factores. En este caso, por supuesto, debe tomarse en cuenta la situación tanto de que un factor es fijo como de que dos factores son fijos. La tabla 14.19 cubre ambas situaciones.

Observe que en el caso de la A aleatoria, todos los efectos tienen pruebas f apropiadas. Pero en el caso de A y B aleatorias, el efecto principal C debe probarse con el uso de un procedimiento tipo Satterthwaitte similar al experimento del modelo II.

14.6 Elección del tamaño de la muestra

A lo largo de este capítulo, nuestro estudio de experimentos factoriales se ha restringido al uso de un diseño aleatorio por completo, con la excepción de la sección 14.4,

Tabla 14.19: Medias cuadráticas esperadas para experimentos con tres factores del modelo III

	***A* aleatoria**	***A* aleatoria, *B* aleatoria**
A	$\sigma^2 + bcn\sigma^2_\alpha$	$\sigma^2 + cn\sigma^2_{\alpha\beta} + bcn\sigma^2_\alpha$
B	$\sigma^2 + cn\sigma^2_{\alpha\beta} + acn\sum_{j=1}^{b}\frac{B_j^2}{b-1}$	$\sigma^2 + cn\sigma^2_{\alpha\beta} + acn\sigma^2_\beta$
C	$\sigma^2 + bn\sigma^2_{\alpha\gamma} + abn\sum_{k=1}^{c}\frac{C_k^2}{c-1}$	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + an\sigma^2_{\beta\gamma} + bn\sigma^2_{\alpha\gamma} + abn\sum_{k=1}^{c}\frac{C_k^2}{c-1}$
AB	$\sigma^2 + cn\sigma^2_{\alpha\beta}$	$\sigma^2 + cn\sigma^2_{\alpha\beta}$
AC	$\sigma^2 + bn\sigma^2_{\alpha\gamma}$	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + bn\sigma^2_{\alpha\gamma}$
BC	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + an\sum_j\sum_k\frac{(BC)_{jk}^2}{(b-1)(c-1)}$	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + an\sigma^2_{\beta\gamma}$
ABC	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma}$	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma}$
Error	σ^2	σ^2

donde se demostró el análisis de un experimento con dos factores en un diseño de bloques aleatorios. El diseño completamente aleatorio es fácil de plantear, y el análisis, sencillo de ejecutar; sin embargo, debe utilizarse sólo cuando el número de combinaciones de tratamientos sea pequeño y el material experimental sea homogéneo. Aunque el diseño de bloques aleatorio es ideal para dividir un grupo grande de unidades heterogéneas en subgrupos de unidades homogéneas, por lo general, es difícil obtener bloques uniformes con unidades suficientes a las cuales asignar un número grande de combinaciones de tratamientos. Esta desventaja se elimina con la elección de un diseño del catálogo de **diseños de bloques incompletos**, los cuales permiten investigar las diferencias entre los t tratamientos arreglados en b bloques, cada uno de los cuales contiene k unidades experimentales, donde $k < t$. El lector puede consultar a Box, Hunter y Hunter, para conocer más detalles.

Una vez que se ha seleccionado un diseño aleatorio por completo, se debe decidir si el número de repeticiones es suficiente para producir pruebas de potencia alta en el análisis de varianza. Si no es así, deben agregarse repeticiones, lo que a su vez hace necesario un diseño con bloques aleatorios por completo. Una vez que se haya comenzado con un diseño con bloques aleatorios, será necesario determinar si el número de éstos es suficiente para efectuar pruebas poderosas. Entonces, básicamente se regresa a la pregunta del tamaño de la muestra.

La potencia de una prueba de efectos fijos para un tamaño de muestra dado se encuentra en la tabla A.16, con el cálculo del parámetro λ de no centralidad, y la función ϕ^2 que se analizó en la sección 13.14. En la tabla 14.20 se dan las expresiones para λ y ϕ^2 para experimentos de efectos fijos de dos y tres factores.

Los resultados de la sección 13.14 para el modelo de efectos aleatorios pueden extrapolarse con facilidad a modelos de dos y tres factores. De nuevo, el procedimiento general se basa en los valores de las medias cuadráticas esperadas. Por ejemplo, si se está probando $\sigma^2_\alpha = 0$ en un experimento de dos factores, al calcular la razón s_1^2/s_3^2 (media cuadrática A/media cuadrática AB), entonces

$$f = \frac{s_1^2/(\sigma^2 + n\sigma^2_{\alpha\beta} + bn\sigma^2_\alpha)}{s_3^2/(\sigma^2 + n\sigma^2_{\alpha\beta})}$$

es un valor de la variable aleatoria F que tiene distribución F con $a - 1$ y $(a - 1)$

Tabla 14.20: Parámetros λ y ϕ^2 para modelos de dos y tres factores

	Experimentos con dos factores		**Experimentos con tres factores**		
	A	***B***	***A***	***B***	***C***
λ	$\frac{bn}{2\sigma^2}\sum_{i=1}^{a}\alpha_i^2$	$\frac{an}{2\sigma^2}\sum_{j=1}^{b}\beta_j^2$	$\frac{bcn}{2\sigma^2}\sum_{i=1}^{a}\alpha_i^2$	$\frac{acn}{2\sigma^2}\sum_{j=1}^{b}\beta_j^2$	$\frac{abn}{2\sigma^2}\sum_{k=1}^{c}\gamma_k^2$
ϕ^2	$\frac{bn}{a\sigma^2}\sum_{i=1}^{a}\alpha_i^2$	$\frac{an}{b\sigma^2}\sum_{j=1}^{b}\beta_j^2$	$\frac{bcn}{a\sigma^2}\sum_{i=1}^{a}\alpha_i^2$	$\frac{acn}{b\sigma^2}\sum_{j=1}^{b}\beta_j^2$	$\frac{abn}{c\sigma^2}\sum_{k=1}^{c}\gamma_k^2$

$(b-1)$ grados de libertad, y la potencia de la prueba es

$$1-\beta = P\left\{\frac{S_1^2}{S_3^2} > f_\alpha[(a-1),(a-1)(b-1)] \text{ cuando } \sigma_\alpha^2 \neq 0\right\}$$
$$= P\left\{F > \frac{f_\alpha[(a-1),(a-1)(b-1)](\sigma^2+n\sigma_{\alpha\beta}^2)}{s^2+n\sigma_{\alpha\beta}^2+bn\sigma_\alpha^2}\right\}.$$

Ejercicios

14.25 Para estimar los distintos componentes de la variabilidad en un proceso de filtración, el porcentaje de material que se pierde en el licor madre se mide en 12 condiciones experimentales, con 3 corridas en cada repetición. Se seleccionan al azar 3 filtros y 4 operadores para usarlos en el experimento, lo que da como resultado las siguientes mediciones:

	Operador			
Filtro	**1**	**2**	**3**	**4**
1	16.2	15.9	15.6	14.9
	16.8	15.1	15.9	15.2
	17.1	14.5	16.1	14.9
2	16.6	16.0	16.1	15.4
	16.9	16.3	16.0	14.6
	16.8	16.5	17.2	15.9
3	16.7	16.5	16.4	16.1
	16.9	16.9	17.4	15.4
	17.1	16.8	16.9	15.6

a) Pruebe la hipótesis de que no hay interacción entre los componentes de la varianza para los filtros y los operadores, con un nivel de significancia $\alpha = 0.05$.

b) Pruebe la hipótesis de que los operadores y los filtros no tienen ningún efecto sobre la variabilidad del proceso de filtración, con el nivel de significancia $\alpha = 0.05$.

c) Estime los componentes de la varianza que se debe a los filtros, operadores y error experimental.

14.26 Si se supone un experimento del modelo II para el ejercicio 14.2 de la página 587, estime los componentes de la varianza para las marcas de concentrado de jugo de naranja, para el número de días a partir del que el jugo se mezcló hasta que se hizo la prueba, y para el error experimental.

14.27 Considere el análisis de varianza siguiente para un experimento del modelo II:

Fuente de variación	**Grados de libertad**	**Media cuadrática**
A	3	140
B	1	480
C	2	325
AB	3	15
AC	6	24
BC	2	18
ABC	6	2
Error	24	5
Total	47	

Pruebe los componentes significativos de la varianza entre todos los efectos principales, y los efectos de la interacción, para un nivel de significancia de 0.01

a) usando un estimador agrupado del error cuando esto sea apropiado;

b) sin agrupar las sumas de los cuadrados de los efectos insignificantes.

14.28 En el ejercicio 14.16 de la página 597, ¿son suficientes dos observaciones para cada combinación de tratamientos, si la potencia de nuestra prueba para detectar diferencias entre los niveles del factor C con un nivel de significancia de 0.05 deben ser de al menos 0.8, cuando $\gamma_1 = -0.2$, $\gamma_2 = -0.4$, y $\gamma_3 = -0.2$? Hágalo

con el mismo estimador agrupado de σ^2 que se utilizó en el análisis de varianza.

14.29 Con el empleo de los estimadores de los componentes de la varianza del ejercicio 14.25, evalúe la potencia cuando se prueba que el componente de la varianza debido a los filtros es igual a cero.

14.30 Un contratista de la defensa está interesado en estudiar un proceso de inspección para detectar fallas y fatiga de partes de recambio. Se utilizan tres niveles de inspección que ejecutan tres inspectores elegidos al azar. Se emplean cinco lotes por cada combinación en el estudio. Los niveles de los factores están en los datos. La respuesta se expresa en fallas por cada 1000 piezas.

	Nivel de inspección		
Inspector	**Inspección militar completa**	**Inspección militar reducida**	**Comercial**
A	7.50 7.42	7.08 6.17	6.15 5.52
	5.85 5.89	5.65 5.30	5.48 5.48
	5.35	5.02	5.98
B	7.58 6.52	7.68 5.86	6.17 6.20
	6.54 5.64	5.28 5.38	5.44 5.75
	5.12	4.87	5.68
C	7.70 6.82	7.19 6.19	6.21 5.66
	6.42 5.39	5.85 5.35	5.36 5.90
	5.35	5.01	6.12

a) Escriba un modelo adecuado, con suposiciones.

b) Utilice análisis de varianza para probar las hipótesis apropiadas para los inspectores, el nivel de inspección y la interacción.

14.31 Un fabricante de pintura látex para interiores (marca *A*) quisiera demostrar que su pintura es más robusta para el material donde aplica, que la de sus dos competidores más cercanos. La respuesta es el tiempo, en años, hasta que comienza a picarse. El estudio implica las tres marcas de pintura y tres materiales seleccionados al azar. Para cada combinación se utilizan dos piezas.

	Marca de pintura		
Material	***A***	***B***	***C***
A	5.50 5.15	4.75 4.60	5.10 5.20
B	5.60 5.55	5.50 5.60	5.40 5.50
C	5.40 5.48	5.05 4.95	4.50 4.55

a) ¿Cuál es el tipo de modelo que se prefiere?

b) Analice los datos, con el empleo del modelo apropiado.

c) ¿Los datos apoyan la afirmación del fabricante de la marca *A*?

14.32 A un gerente de una planta le gustaría demostrar que la producción de una fábrica de lana de su planta no depende del operador de la máquina ni de la hora del día, y que es consistentemente elevada Para hacer el estudio se eligen al azar cuatro operadores y tres horas del día también al azar. Se mide el producto en yardas por minuto. Se toman muestras aleatorias en tres días elegidos al azar. Los datos son los siguientes:

	Operador			
Hora	**1**	**2**	**3**	**4**
1	9.5	9.8	9.8	10.0
	9.8	10.1	10.3	9.7
	10.0	9.6	9.7	10.2
2	10.2	10.1	10.2	10.3
	9.9	9.8	9.8	10.1
	9.5	9.7	9.7	9.9
3	10.5	10.4	9.9	10.0
	10.2	10.2	10.3	10.1
	9.3	9.8	10.2	9.7

a) Escriba el modelo apropiado.

b) Evalúe los componentes de la varianza para el operador y la hora.

c) Saque sus conclusiones.

14.33 Un ingeniero de procesos desea determinar si la energía que se alimenta a las máquinas que llenan ciertos tipos de cajas de cereal tienen un efecto significativo sobre el peso real del producto. El estudio consiste en tomar al azar 3 tipos de cereal elaborados por la compañía, y 3 flujos fijos de energía. Para cada combinación se mide el peso de cuatro cajas de cereal diferentes seleccionadas al azar. El peso que se desea es de 400 gramos. A continuación se presentan los datos.

Flujo de energía	**Tipo del cereal**		
	1	**2**	**3**
Bajo	395 390	392 392	402 405
	401 400	394 401	399 399
Actual	396 399	390 392	404 403
	400 402	395 502	400 399
Alto	410 408	404 406	415 412
	408 407	401 400	413 415

a) Dé el modelo adecuado y liste las suposiciones que se hacen.

b) ¿Hay un efecto significativo debido al flujo de energía?

c) ¿Existe un componente significativo de la varianza debido al tipo de cereal?

Ejercicios de repaso

14.34 El Centro de Consulta en Estadística, del Instituto Politécnico y Universidad Estatal de Virginia, participa en el análisis de un conjunto de datos tomados por el personal del Departamento de Nutrición Humana y Alimentos, en el cual hay interés por estudiar los efectos del tipo de harina y el porcentaje de edulcorante, sobre ciertos atributos físicos de un tipo de pastel. Se usaron harinas para todo uso y para pasteles, y el porcentaje de edulcorante varió en cuatro niveles. Los datos siguientes muestran información acerca de la gravedad específica de las muestras de pastel. Con cada una de las ocho combinaciones de factores se prepararon tres pasteles.

Concentración de edulcorante	Todo uso			Pastel		
0	0.90	0.87	0.90	0.91	0.90	0.80
50	0.86	0.89	0.91	0.88	0.82	0.83
75	0.93	0.88	0.87	0.86	0.85	0.80
100	0.79	0.82	0.80	0.86	0.85	0.85

a) Haga el análisis de varianza con dos factores. Pruebe las diferencias entre el tipo de harina. Pruebe las diferencias entre la concentración de edulcorante.

b) Analice el efecto de la interacción, si lo hubiera. Para todas las pruebas, señale los valores *P*.

14.35 Se hizo un experimento en el Departamento de Ciencia de Alimentos, del Instituto Politécnico y Universidad Estatal de Virginia. Tenía interés caracterizar la textura de ciertos tipos de peces de la familia de los arenques. También se estudió el efecto de los tipos de salsa empleada para preparar el pescado. La respuesta en el experimento era un "valor de textura", medido con una máquina que rebanaba el producto de los peces. Los siguientes datos son los valores de textura:

Tipo de salsa	Sábalo sin curar		Sábalo curado		Arenque	
Crema ácida	27.6	57.4	64.0	66.9	107.0	83.9
	47.8	71.1	66.5	66.8	110.4	93.4
	53.8		53.8		83.1	
Salsa envinada	49.8	31.0	48.3	62.2	88.0	95.2
	11.8	35.1	54.6	43.6	108.2	86.7
	16.1		41.8		105.2	

a) Haga un análisis de varianza. Determine si hay o no interacción entre el tipo de salsa y el tipo de pez.

b) Con base en los resultados del inciso *a*) y en pruebas *F* de los efectos principales, determine si hay diferencia en la textura debido a los tipos de salsa, y determine si existe una diferencia significativa entre los tipos de peces.

14.36 Se hizo un estudio para determinar si las condiciones de humedad tienen efecto sobre la fuerza que se requiere para separar piezas de plástico pegadas. Se probaron tres tipos de plástico con cuatro niveles de humedad. Los resultados, en kilogramos, son los siguientes:

	Humedad			
Tipo de plástico	**30%**	**50%**	**70%**	**90%**
A	39.0	33.1	33.8	33.0
	42.8	37.8	30.7	32.9
B	36.9	27.2	29.7	28.5
	41.0	26.8	29.1	27.9
C	27.4	29.2	26.7	30.9
	30.3	29.9	32.0	31.5

a) Si se supone un experimento de modelo I, lleve a cabo un análisis de varianza y pruebe la hipótesis de que no hay interacción entre la humedad y el tipo de plástico, con un nivel de significancia de 0.05.

b) Usando sólo plásticos *A* y *B* y el valor de s^2 del inciso *a*), pruebe otra vez la presencia de la interacción con un nivel de significancia de 0.05.

c) Utilice una comparación con un grado de libertad y el valor de s^2 del inciso *a*), para comparar, con un nivel de significancia de 0.05, la fuerza que se requiere con 30% de humedad contra 50, 70 y 90%.

d) Repita el inciso *c*), usando sólo el plástico *C* y el valor de s^2 del inciso *a*).

14.37 Personal del Departamento de Ingeniería de Materiales del Instituto Politécnico y Universidad Estatal de Virginia llevó a cabo un experimento para estudiar los efectos de los factores ambientales sobre la estabilidad de cierto tipo de aleación cobre-níquel. La respuesta básica fue la vida de fatiga del material. Los factores son *nivel* de esfuerzo y *ambiente*. Los datos son los siguientes.

	Nivel de esfuerzo		
Ambiente	**Bajo**	**Medio**	**Alto**
Hidrógeno seco	11.08	13.12	14.18
	10.98	13.04	14.90
	11.24	13.37	15.10
Humedad elevada (95%)	10.75	12.73	14.15
	10.52	12.87	14.42
	10.43	12.95	14.25

a) Haga un análisis de varianza para probar la interacción entre los factores. Use $\alpha = 0.05$.

b) Con base en el inciso *a*), efectúe un análisis sobre los dos efectos principales y saque sus conclusiones. Utilice el enfoque del valor *P* para obtener las conclusiones.

14.38 En el experimento del ejercicio de repaso 14.34, también se utilizó como respuesta el volumen del pastel. Las unidades en que se expresa son pulgadas cúbicas. Pruebe la interacción entre los factores y analice

los efectos principales. Suponga que los dos factores son efectos fijos.

Concentración de edulcorante	Harina					
	Todo uso			Pastel		
0	4.48	3.98	4.42	4.12	4.92	5.10
50	3.68	5.04	3.72	5.00	4.26	4.34
75	3.92	3.82	4.06	4.82	4.34	4.40
100	3.26	3.80	3.40	4.32	4.18	4.30

14.39 Una válvula de control necesita ser muy sensible al voltaje de entrada, para así generar un voltaje de salida adecuado. Un ingeniero gira las perillas de control para cambiar el voltaje de entrada. En el libro *SN-Ratio for the Quality Evaluation*, publicado por la Japanese Standards Association (1988), se describe un estudio sobre la forma en que esos tres factores (posición relativa de las perillas de control, rango de control de las perillas y voltaje de entrada) influyen en la sensibilidad de una válvula de control. A continuación se presentan los factores y sus niveles. Los datos muestran la sensibilidad de una válvula de control.

Factor A: posición relativa de las perillas de control: centro -0.5, centro y centro $+0.5$.
Factor B: rango de control de las perillas: 2, 4.5 y 7 (mm).
Factor C: voltaje de entrada: 100, 120 y 150 (V).

		C					
A	B	C_1		C_2		C_3	
A_1	B_1	151	135	151	135	151	138
A_1	B_2	178	171	180	173	181	174
A_1	B_3	204	190	205	190	206	192
A_2	B_1	156	148	158	149	158	150
A_2	B_2	183	168	183	170	183	172
A_2	B_3	210	204	211	203	213	204
A_3	B_1	161	145	162	148	163	148
A_3	B_2	189	182	191	184	192	183
A_3	B_3	215	202	216	203	217	205

Realice un análisis de varianza con $\alpha = 0.05$ para probar la significancia de los efectos principal y de interacción. Saque sus conclusiones.

14.40 En el ejercicio 14.23 de la página 600 se describe un experimento que implica la extracción de polietileno a través de un solvente.

Solvente	Temp.	Tiempo					
		4		8		16	
Etanol	**120**	94.0,	94.0	93.8,	94.2	91.1,	90.5
	80	95.3,	95.1	94.9,	95.3	92.5,	92.4
Tolueno	**120**	94.6,	94.5	93.6,	94.1	91.1,	91.0
	80	95.4,	95.4	95.6,	96.0	92.1,	92.1

a) Haga una clase diferente de análisis de los datos. Ajuste un modelo adecuado de regresión con una variable categórica del solvente, un término de temperatura, otro del tiempo, y uno para la interacción de la temperatura con el tiempo, una interacción del solvente con la temperatura, y otra del solvente con el tiempo. Realice pruebas t sobre todos los coeficientes y describa sus descubrimientos.

b) ¿Sus resultados sugieren que modelos diferentes son apropiados para el etanol y el tolueno, o tienen separación equivalente de las intersecciones? Explique su respuesta.

c) ¿Obtuvo conclusiones que contradigan las de la solución del ejercicio 14.23? Explique.

14.41 En el libro *SN-Ratio for the Quality Evaluation*, publicado por la Japanese Standards Association (1988), se describe un estudio que se realizó acerca de cómo la presión del aire de las llantas afecta la maniobrabilidad de un automóvil. Se compararon tres presiones distintas del aire en aquellas, sobre tres superficies diferentes de manejo. Las tres presiones del aire fueron de 6 kgf/cm^2 para las llantas tanto del lado izquierdo como del derecho infladas a 6 kgf/cm^2, las llantas del lado izquierdo infladas a 3kgf/cm^2 y las llantas de ambos lados infladas a 3 kgf/cm^2. Las tres superficies de manejo fueron asfalto, asfalto seco y cemento seco. Se observó dos veces el radio de curvatura de un vehículo de prueba para cada nivel de presión de las llantas sobre cada una de las tres superficies de manejo.

	Presión del aire de las llantas					
	1		2		3	
Asfalto	44.0	25.5	34.2	37.2	27.4	42.8
Asfalto seco	31.9	33.7	31.8	27.6	43.7	38.2
Cemento seco	27.3	39.5	46.6	28.1	35.5	34.6

Realice un análisis de varianza con los datos anteriores. Haga comentarios acerca de la interpretación de los efectos principal y de interacción.

14.42 El fabricante de cierta marca de café secado por congelación espera reducir el tiempo del proceso sin arriesgar la integridad del producto. Desea usar tres temperaturas para la cámara de secado y cuatro tiempos para secar. El tiempo de secado actual es de 3 horas con una temperatura de -15 °C. La respuesta del sabor es un promedio de las calificaciones de cuatro jueces profesionales. La calificación está en una escala de 1 a 10, donde 10 es la mejor. En la tabla que sigue se presentan los datos.

Tiempo	Temperatura					
	-20 °C		-15 °C		-10 °C	
1 h	9.60	9.63	9.55	9.50	9.40	9.43
1.5 h	9.75	9.73	9.60	9.61	9.55	9.48
2 h	9.82	9.93	9.81	9.78	9.50	9.52
3 h	9.78	9.81	9.80	9.75	9.55	9.58

a) ¿Qué tipo de modelo debe utilizarse? Plantee las suposiciones.

b) Analice los datos en forma apropiada.

c) Escriba un reporte breve al vicepresidente encargado y hágale una recomendación para la elaboración futura de este producto.

14.43 Para garantizar el número de cajeros necesarios durante las horas pico de la operación, se recabaron datos en un banco citadino. Se estudiaron cuatro cajeros durante tres horarios "ocupados", **1.** entre semana, de 10:00 a 11:00 A.M., **2.** por las tardes entre semana, entre las 2:00 y las 3:00 P.M., y **3.** las mañanas de los sábados, entre las 11:00 y las 12:00. Un analista eligió al azar cuatro horarios dentro de cada uno de los tres periodos, para cada una de las cuatro posiciones de los cajeros durante un periodo de varios meses y se observó el número de clientes atendidos. Los datos son los siguientes:

	Periodo		
Cajero	**1**	**2**	**3**
1	18, 24, 17, 22	25, 29, 23, 32	29, 30, 21, 34
2	16, 11, 19, 14	23, 32, 25, 17	27, 29, 18, 16
3	12, 19, 11, 22	27, 33, 27, 24	25, 20, 29, 15
4	11, 9, 13, 8	10, 7, 19, 8	11, 9, 17, 9

Se supone que el número de clientes atendidos es una variable aleatoria de Poisson.

a) Comente sobre el riesgo de llevar a cabo un análisis de varianza estándar con los datos anteriores. ¿Qué suposiciones, si las hubiera, se trasgredirían?

b) Realice una tabla de ANOVA estándar que incluya pruebas F de los efectos y las interacciones principales. Si las interacciones y los efectos principales son significativos, establezca las conclusiones científicas. ¿Qué aprendimos? Asegúrese de interpretar cualquier interacción significativa. Utilice su propio juicio con respecto a los valores P.

c) Realice un análisis completo de nuevo usando una transformación apropiada en la respuesta. ¿Encontró alguna diferencia en los resultados? Haga comentarios al respecto.

14.7 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

Uno de los temas más susceptibles de confusión en el análisis de experimentos factoriales, radica en la interpretación de los efectos principales ante la presencia de interacción. La existencia de un valor P relativamente grande para un efecto principal, cuando es clara la presencia de interacciones, podría tentar al analista a concluir que "no existe efecto principal significativo". Sin embargo, debe entenderse que si un efecto principal está implicado en una interacción significativa, entonces el efecto principal **está influyendo en la respuesta**. La naturaleza del efecto es inconsistente a través de los niveles de otros efectos. La naturaleza del papel del efecto principal se deduce en las **gráficas de interacción**.

A la luz de lo que se afirma en el párrafo anterior, hay peligro sustancial de usar de manera equivocada la estadística cuando se emplea en un prueba de comparación múltiple sobre los efectos principales ante la presencia clara de la interacción entre los factores.

Debe tenerse precaución en el análisis de un experimento factorial cuando se hace la suposición de un diseño aleatorio por completo y en realidad éste no ocurre. Por ejemplo, es común que se encuentren factores que son **muy difíciles de cambiar**. Como resultado, podría ser necesario mantener sin cambio, durante largos periodos, los niveles de factores a través del experimento. El ejemplo de la temperatura es común. Subirla o bajarla en un esquema aleatorio es un plan costoso y la mayoría de los experimentadores evitarán hacerlo. Los diseños experimentales con *restricciones en la aleatorización* son muy comunes y reciben el nombre de **diseños de gráficas separadas**. Están más allá del alcance de este libro, pero en Montgomery, 2001, se encuentra su presentación.

Capítulo 15

Experimentos factoriales 2^k y fracciones

15.1 Introducción

Ya se han expuesto ciertos conceptos de diseños experimentales. El plan de muestreo para la prueba t simple sobre la media de una población normal, y también el análisis de varianza que implica la asignación, a las unidades experimentales, los tratamientos seleccionados previamente al azar. El diseño por bloques completamente aleatorios, donde los tratamientos se asignan a las unidades dentro de bloques relativamente homogéneos, implica aleatorización restringida.

En este capítulo se presta atención especial a diseños experimentales, en los cuales el plan experimental requiere el estudio del efecto sobre una respuesta de k factores, cada uno de los cuales en dos niveles. Éstos, por lo general, se conocen como **experimentos factoriales 2^k**. Es frecuente que los niveles se denoten como "alto" y "bajo", aun cuando esa notación sea arbitraria en el caso de variables cualitativas. El diseño factorial completo requiere que cada nivel de cada factor ocurra con cada nivel de cada uno de los demás factores, lo que da un total de **2^k combinaciones de tratamientos**.

Exploración de factores y experimentación secuencial

Es frecuente que cuando se lleva a cabo experimentación, ya sea en un nivel de investigación o desarrollo, un diseño experimental bien planeado es una **etapa** de lo que en realidad es el **plan secuencial** de la experimentación. Es más bien frecuente que los científicos e ingenieros, al finalizar el estudio, no tengan cuidado de qué factores son importantes ni de cuáles son los rangos apropiados en los factores potenciales por donde debería conducirse la experimentación. Por ejemplo, en el libro *Response Surface Methodology*, Myers y Montgomery (2002) dan un ejemplo de investigación de un experimento en una planta piloto, en el que se varían cuatro factores —temperatura, presión, concentración de formaldehído y tasa de enfriamiento—, con la finalidad de establecer su influencia sobre la respuesta, la tasa de filtración de cierto producto químico. Aun al nivel de planta piloto, los científicos no están seguros de si deben intervenir en todos los factores en el modelo. Además, el objetivo final

consiste en determinar los niveles adecuados en que contribuyen los factores para maximizar la tasa de filtración. Así, existe la necesidad de determinar la **región apropiada de experimentación**. Las preguntas pueden responderse sólo si el plan experimental total se realiza en forma secuencial. Muchos intentos son planes que implican *aprendizaje iterativo*, el tipo de aprendizaje consistente con el método científico, en el que la palabra *iterativo* implica experimentación con sabiduría en cada etapa.

Por lo general, la etapa inicial del plan secuencial ideal es variable, o de **exploración factorial**, que es un procedimiento que implica un diseño experimental de bajo costo para buscar los **factores candidatos**. Esto tiene importancia especial cuando el plan requiere un sistema complejo, como un proceso de manufactura. La información obtenida a partir de los resultados de un *diseño exploratorio* se emplea para diseñar uno o más experimentos posteriores, en los cuales se realizan ajustes en los factores importantes. Se trata de ajustes con los que se obtendrán mejorías en el sistema o proceso.

Los experimentos factoriales 2^k y fracciones de 2^k son poderosas herramientas que son diseños exploratorios ideales. Son sencillos y prácticos, y tiene una atracción intuitiva. Muchos de los conceptos generales que se estudian en el capítulo 14 siguen siendo válidos. Sin embargo, hay métodos gráficos que brindan anticipaciones útiles en el análisis de los dos niveles de diseño.

Diseños exploratorios para números grandes de factores

Cuando k es pequeña, digamos $k = 2$ o incluso $k = 3$, queda clara la utilidad del factorial 2^k para la exploración de factores. Tanto el análisis de varianza como el de regresión, según se estudiaron e ilustraron en los capítulos 12, 13 y 14, continúan siendo herramientas útiles. Además, los enfoques gráficos serán evidentes.

Si k es grande, por ejemplo 6, 7 u 8, el número de combinaciones de factores y, por lo tanto, de corridas experimentales, con frecuencia resultará ser prohibitivo. Por ejemplo, suponga que hay interés en ejecutar un diseño exploratorio que involucre $k = 8$ factores. Podría desearse obtener información acerca de tolos los $k = 8$ efectos principales, así como de las $\frac{k(k-1)}{2} = 28$ interacciones de dos factores. Sin embargo, $2^8 = 256$ corridas parecería un número demasiado grande y excesivo para estudiar $28 + 8 = 36$ efectos. Pero, como se verá en secciones futuras, cuando k es grande, es posible obtener información considerable de manera eficaz usando sólo una fracción del experimento factorial 2^k completo. Esta clase de diseños es la clase de *diseños factoriales fraccionarios*. La meta es recuperar información de alta calidad acerca de los efectos principales y las interacciones interesantes, aun cuando el tamaño del diseño se reduzca en forma considerable.

15.2 El factorial 2^k: Cálculo de los efectos y análisis de varianza

Considere inicialmente un factorial 2^2 con factores A y B y n *observaciones experimentales por combinación de factores*. Es útil emplear los símbolos (1), a, b y ab para denotar los puntos del diseño, donde la presencia de una letra minúscula implique que el factor (A o B) está en el *nivel alto*. Así, la ausencia de la minúscula implica que el factor está en el *nivel bajo*. Por lo que ab es el punto de diseño (+,+), a es (+,−), b es (−,+) y (1) es (−,−). Más adelante habrá situaciones en que la notación también se aplique para los datos de respuesta en el punto de diseño en cuestión. Como introducción al cálculo de **efectos** importantes que ayuden a la determinación de la influencia

de los factores y **sumas de cuadrados** que están incorporados en los cálculos del análisis de varianza, se presenta la tabla 15.1.

Tabla 15.1: Un experimento factorial 2^2

		A	**Media**
B	b	ab	$\frac{b+ab}{2n}$
	(1)	a	$\frac{(1)+a}{2n}$
Media	$\frac{(1)+b}{2n}$	$\frac{a+ab}{2n}$	

En esta tabla, (1), a, b y ab representan totales de los n valores de la respuesta en los puntos de diseño individuales. La simplicidad del factorial 2^2 se define por el hecho de que aparte del error experimental, la información importante se le da al analista en componentes de un solo grado de libertad, uno para los dos efectos principales A y B, y un grado de libertad para la interacción AB. La información que se recupera sobre todos estos aspectos adopta la forma de tres **contrastes**. Se definirán los siguientes contrastes entre los totales de los tratamientos:

$$\text{contraste de } A = ab + a - b - (1),$$

$$\text{contraste de } B = ab - a + b - (1),$$

$$\text{contraste de } AB = ab - a - b + (1).$$

Los tres **efectos** del experimento implican estos contrastes y requieren de sentido común e intuición. Los dos efectos principales calculados tienen la forma

$$\text{efecto} = \bar{y}_H - \bar{y}_L,$$

donde $\bar{y}_H$ y $\bar{y}_L$ son las respuestas promedio en el nivel alto o "+", y en el nivel bajo o "−", respectivamente. Como resultado, queda

Cálculo de los efectos principales

$$A = \frac{ab + a - b - (1)}{2n} = \frac{A \text{ contraste}}{2n},$$

y

$$B = \frac{ab - a + b - (1)}{2n} = \frac{B \text{ contraste}}{2n}.$$

La cantidad A es vista como la *diferencia entre la respuesta media en los niveles alto y bajo del factor A*. De hecho, A se denomina el **efecto principal** de A. En forma similar, B es el efecto principal del factor B. En la tabla 15.1, al inspeccionar la diferencia entre $ab - b$ y $a - (1)$, o entre $ab - a$ y $b - (1)$, se observa interacción aparente en los datos. Por ejemplo, si

$$ab - a \approx b - (1) \quad \text{o bien} \quad ab - a - b + (1) \approx 0,$$

una recta que conectara las respuestas para cada nivel del factor A en el nivel alto del factor B, sería aproximadamente paralela a una recta que uniera la respuesta para cada nivel del factor A en el nivel bajo del factor B. Las rectas no paralelas

de la figura 15.1 sugieren la presencia de interacción. Para probar si tal interacción aparente es significativa, se construye un tercer contraste en los totales del tratamiento ortogonal a los contrastes del efecto principal, el cual se denomina **efecto de la interacción**, lo que se hace con la evaluación de

Efecto de la interacción

$$AB = \frac{ab - a - b + (1)}{2n} = \frac{AB \text{ contraste}}{2n}.$$

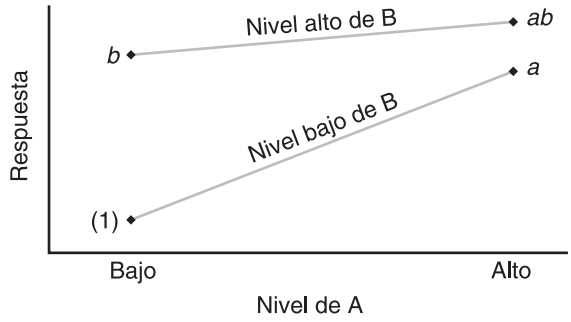


Figura 15.1: Respuesta que sugiere interacción aparente.

Ejemplo 15.1: Considere los datos de las tablas 15.2 y 15.3, con $n = 1$ para un experimento factorial 2^2.

Tabla 15.2: Factorial 2^2 sin interacción

	B	
A	$-$	$+$
$+$	50	70
$-$	80	100

Tabla 15.3: Factorial 2^2 con interacción

	B	
A	$-$	$+$
$+$	50	70
$-$	80	40

Los números en las celdas de las tablas 15.2 y 15.3 ilustran con claridad la manera en que los contrastes y el cálculo resultante de los dos efectos principales y conclusiones que surgen pueden recibir una gran influencia de la presencia de interacción. En la tabla 15.2, el efecto de A es -30 en los niveles tanto bajo como alto de B, y el efecto de B es 20 en los dos niveles del factor A, bajo y alto. Esta "consistencia del efecto" (no hay interacción) es una información muy importante para el analista. Los efectos principales son

$$A = \frac{70 + 50}{2} - \frac{100 + 80}{2} = 60 - 90 = -30,$$
$$B = \frac{100 + 70}{2} - \frac{80 + 50}{2} = 85 - 65 = 20,$$

mientras que el efecto de la interacción es

$$AB = \frac{100 + 50}{2} - \frac{80 + 70}{2} = 75 - 75 = 0.$$

Por otro lado, en la tabla 15.3, el efecto A es de nuevo -30 al nivel bajo de B, pero $+30$ al nivel alto de B. Esta "inconsistencia del efecto" (interacción) también está presente para B a través de los niveles de A. En estos casos, los efectos principales pueden carecer de significado y, en efecto, confundir mucho. Por ejemplo el efecto de A es

$$A = \frac{50 + 70}{2} - \frac{80 + 40}{2} = 0,$$

ya que hay un "enmascaramiento" completo del efecto conforme se promedia sobre los niveles de B. La interacción fuerte se ilustra con el efecto calculado

$$AB = \frac{70 + 80}{2} - \frac{50 + 40}{2} = 30.$$

Aquí, es conveniente ilustrar los escenarios de las tablas 15.2 y 15.3 con las gráficas de interacción. Observe el paralelismo en la gráfica de la figura 15.2 y la interacción visible en la figura 15.3.

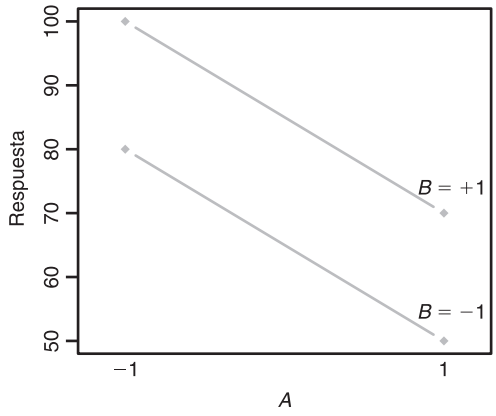


Figura 15.2: Gráfica de la interacción para los datos de la tabla 15.2.

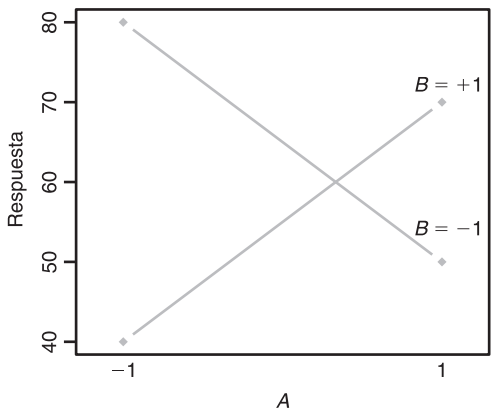


Figura 15.3: Gráfica de la interacción para los datos de la tabla 15.3.

Cálculo de las sumas de los cuadrados

Se aprovecha el hecho de que en el factorial 2^2, o para el caso general del experimento factorial 2^k, cada efecto principal y efecto de la interacción tiene asociado **un solo grado de libertad**. Por lo tanto, es posible escribir contrastes ortogonales $2^k - 1$ de un solo grado de libertad en las combinaciones de tratamientos, y cada una es responsable de la variación debida a cierto efecto principal o interacción. Así, con las suposiciones usuales de independencia y normalidad en el modelo del experimento, se hacen pruebas para determinar si el contraste refleja variación sistemática o sólo variaciones probabilísticas o aleatorias. Las sumas de los cuadrados para cada contraste se encuentran al seguir los procedimientos que se dieron en la sección 13.5. Queda

$$Y_{1..} = b + (1), \qquad Y_{2..} = ab + a, \qquad c_1 = -1, \qquad \text{y} \qquad c_2 = 1,$$

donde $Y_{1..}$ y $Y_{2..}$ son el total de $2n$ observaciones, y se tiene

$$SSA = SSw_A = \frac{\left(\sum_{i=1}^{2} c_i Y_{i..}\right)^2}{2n\sum_{i=1}^{2} c_i^2} = \frac{[ab + a - b - (1)]^2}{2^2 n} = \frac{(A \text{ contraste})^2}{2^2 n},$$

con 1 grado de libertad. Asimismo, se encuentra que

$$SSB = \frac{[ab + b - a - (1)]^2}{2^2 n} = \frac{(B \text{ contraste})^2}{2^2 n},$$

y

$$SS(AB) = \frac{[ab + (1) - a - b]^2}{2^2 n} = \frac{(AB \text{ contraste})^2}{2^2 n}.$$

Cada contraste tienen 1 grado de libertad, mientras que las sumas de los cuadrados de los errores, con $2^2(n-1)$ grados de libertad, se obtienen por sustracción, de la fórmula

$$SSE = SST - SSA - SSB - SS(AB).$$

Al calcular las sumas de los cuadrados para los efectos principales A y B y el efecto de la interacción AB, es conveniente presentar las salidas totales de las combinaciones de tratamiento, junto con los signos algebraicos apropiados para cada contraste, como se ve en la tabla 15.4. Los efectos principales se obtienen como comparaciones simples entre los niveles alto y bajo. Por lo tanto, se establece un signo positivo para la combinación de tratamiento que esté en el nivel alto de un factor dado, y uno negativo a la del nivel bajo. Los signos positivo y negativo para el efecto de la interacción se obtienen al multiplicar los signos correspondientes a los contrastes de los factores de la interacción.

Tabla 15.4: Signos de los contrastes en un experimento factorial 2^2

Combinación de tratamientos	**Efecto factorial**		
	A	***B***	***AB***
(1)	−	−	+
a	+	−	−
b	−	+	−
ab	+	+	+

El factorial 2^3

Ahora consideremos un experimento con el uso de tres factores, A, B y C, cada uno con niveles -1 y $+1$. Éste es un experimento factorial 2^3 que da ocho combinaciones de tratamiento (1), a, b, c, ab, ac, bc y abc. En la tabla 15.5 se presentan las combinaciones de tratamiento y los signos algebraicos apropiados para cada contraste, que se usan para calcular las sumas de los cuadrados para los efectos principales y los de la interacción.

Tabla 15.5: Signos de los contrastes en un experimento factorial 2^3

Combinación de tratamiento	Efecto factorial (simbólico)						
	A	B	C	AB	AC	BC	ABC
(1)	−	−	−	+	+	+	−
a	+	−	−	−	−	+	+
b	−	+	−	−	+	−	+
c	−	−	+	+	−	−	+
ab	+	+	−	+	−	−	−
ac	+	−	+	−	+	−	−
bc	−	+	+	−	−	+	−
abc	+	+	+	+	+	+	+

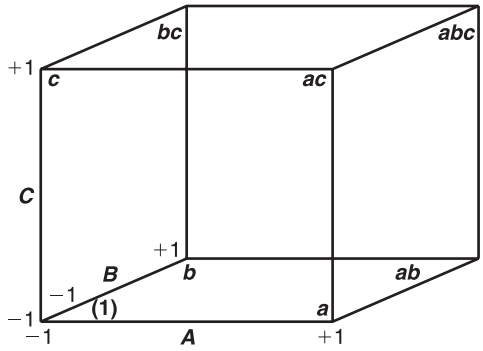


Figura 15.4: Vista geométrica de 2^3.

Es de ayuda analizar e ilustrar la geometría del factorial 2^3 del mismo modo que se hizo para el 2^2 en la figura 15.1. Para el 2^3, los **ocho puntos de diseño** representan los vértices de un cubo que se muestra en la figura 15.4.

Las columnas de la tabla 15.5 representan los signos que se utilizan para los contrastes y los cálculos de siete efectos y las sumas de los cuadrados correspondientes. Estas columnas son análogas a las que se dan en la tabla 15.4 para el caso de 2^2. Como los puntos de diseño son ocho, hay siete efectos disponibles. Por ejemplo,

$$A = \frac{a + ab + ac + abc - (1) - b - c - bc}{4n},$$
$$AB = \frac{(1) + c + ab + abc - a - b - ac - bc}{4n},$$

y así sucesivamente. Las sumas de cuadrados están dadas por

$$SS(\text{efecto}) = \frac{(\text{contraste})^2}{2^3 n}.$$

El estudio de la tabla 15.5 revela que para el experimento 2^3 todos los contrastes entre los siete son mutuamente ortogonales y por ello los siete efectos se evalúan en forma independiente.

Efectos y suma de cuadrados para el 2^k

Para un experimento factorial 2^k, las sumas de cuadrados de un solo grado de libertad para los efectos principales, y los efectos de la interacción se obtienen al elevar al cuadrado los contrastes apropiados en los totales del tratamiento y dividiendo entre 2^3n, donde n es el número de repeticiones de las combinaciones del tratamiento.

Como antes, un efecto siempre se calcula al restar la respuesta promedio al nivel "bajo", de la respuesta promedio al nivel "alto". Para los efectos principales, alto y bajo están muy en claro. El alto y bajo simbólicos para las interacciones son evidentes a partir de la información de la tabla 15.5.

La propiedad de ortogonalidad tiene la misma importancia que tenía en el estudio de las comparaciones que se hizo en el capítulo 13. La ortogonalidad de los contrastes implica que los efectos estimados y, por lo tanto, las sumas de los cuadrados, son independientes. Esta independencia se ilustra con claridad en el experimento factorial 2^3 si la salida, con el factor A en su nivel alto, se incrementa en una cantidad x, en la tabla 15.5. Sólo el contraste A conduce a una suma de cuadrados más grande porque el efecto x se cancela en la formación de los seis contrastes remanentes, como resultado de los dos signos positivos y dos negativos, asociados con las combinaciones de tratamientos en los que A se halla en el nivel alto.

Hay ventajas adicionales producidas por la ortogonalidad. Éstas se verán cuando se estudie el experimento factorial 2^k en situaciones de regresión.

15.3 Experimento factorial 2^k no replicado

El factorial completo 2^k con frecuencia requiere experimentación considerable, en particular cuando k es grande. Como resultado, no es raro que no se permita la replicación de cada combinación de factores. Si en el modelo del experimento se incluyen todos los efectos, inclusive todas las interacciones, no se permite ningún grado de libertad para el error. Frecuentemente, cuando k es grande, el analista de los datos *agrupará* las sumas de los cuadrados y los grados de libertad correspondientes para las interacciones de orden superior que se sabe, o se supone, son despreciables. Esto producirá pruebas F para muchos efectos e interacciones de orden inferior.

Graficación de diagnóstico con experimentos factoriales 2^k no replicados

Las gráficas de probabilidad normal son un método muy útil para determinar la importancia relativa de los efectos, en un experimento factorizado de dos niveles razonablemente grande, cuando no hay replicación. En especial este tipo de gráfica de diagnóstico es de mucha ayuda cuando el analista duda en agrupar interacciones de orden superior, por temor de que algunos de los efectos agrupados en el "error" en verdad sean efectos reales y no sólo aleatorios. El lector debe recordar que todos los efectos que no son reales (es decir, son *estimadores de cero* independientes) siguen una distribución normal con media cercana a cero y varianza constante. Por ejemplo, en un experimento factorial 2^4, se debe recordar que todos los efectos (hay que tener en cuanta que $n = 1$) son de la forma

$$AB = \frac{\text{contraste}}{8} = \bar{y}_H - \bar{y}_L,$$

donde $\bar{y}_H$ es el promedio de ocho corridas experimentales independientes en el nivel alto o "+", y $\bar{y}_L$ es el promedio de ocho corridas independientes en el nivel bajo o

"−". Así, la varianza de cada contraste es $Var(\bar{y}_H - \bar{y}_L) = \sigma^2/4$. Para cualesquiera esfectos reales $E(\bar{y}_H - \bar{y}_L) \neq 0$. Así, la gráfica de probabilidad normal debería revelar efectos "significativos" como aquellos que caen fuera de la línea recta que describe realizaciones de variables aleatorias con distribución normal idéntica e independientes.

La gráfica de probabilidad puede adoptar una de muchas formas. Se recomienda al lector que consulte el capítulo 8, donde se presentan dichas gráficas. Puede usarse la gráfica cuantil-cuantil, normal y empírica. También es posible utilizar el procedimiento de graficación con el papel de probabilidad normal. Además, hay otros tipos de gráficas de probabilidad normal para el diagnóstico. En resumen, las gráficas de efectos para el diagnóstico son como sigue.

Gráfica de efectos de probabilidad para experimentos factoriales 2^3 no replicados

1. Calcule los efectos como

$$\text{efecto} = \frac{\text{contraste}}{2^{k-1}}.$$

2. Construya una gráfica de probabilidad normal de todos los efectos.

3. Los efectos que caigan fuera de la línea recta deben considerarse reales.

A continuación se hacen más comentarios respecto de las gráficas de probabilidad normal de los efectos. En primer lugar, el analista quizá se sintiera frustrado si las utilizara con un experimento pequeño. Es probable que la graficación dé resultados satisfactorios cuando haya *escasez del efecto*: muchos efectos que en verdad no son reales. Esta escasez será evidente en experimentos grandes en los cuales es probable que no sean reales las interacciones de orden superior.

15.4 Estudio de caso del moldeo por inyección

Ejemplo 15.2: Muchas compañías manufactureras de Estados Unidos y el extranjero utilizan partes moldeadas como componentes de un proceso. Es frecuente que el rebasamiento sea un problema grande. A menudo, un molde troquelado de una parte se construye con un tamaño más grande que el nominal para permitir el derrame. En la situación experimental siguiente se produce un molde nuevo, y es importante encontrar las especificaciones adecuadas del proceso, con la finalidad de minimizar el derrame. En el experimento siguiente, los valores de la respuesta son desviaciones del nominal (derrames). Los factores y niveles son los siguientes:

	Niveles de código	
	−1	**+1**
***A*. Velocidad de inyección (pies/seg)**	1.0	2.0
***B*. Temperatura del molde (°C)**	100	150
***C*. Presión del molde (psi)**	500	1000
***D*. Presión posterior (psi)**	75	120

El propósito del experimento es determinar cuáles efectos (principales y de interacción) influyen en el derrame. El experimento se consideró un sondeo preliminar, a partir del cual se determinarán los factores para un análisis más completo. Asimismo, se espera obtener alguna perspectiva de cómo podrían determinarse los factores

importantes que influyen en el derrame. En la tabla 15.6 se presentan los datos de un experimento factorial 2^4 no replicado.

Tabla 15.6: Datos para el ejemplo 15.2

Combinación de factores	**Respuesta (cm $\times$ 10^4)**	**Combinación de factores**	**Respuesta (cm $\times$ 10^4)**
(1)	72.68	d	73.52
a	71.74	ad	75.97
b	76.09	bd	74.28
ab	93.19	abd	92.87
c	71.25	cd	79.34
ac	70.59	acd	75.12
bc	70.92	bcd	79.67
abc	104.96	$abcd$	97.80

Inicialmente, se calcularon los efectos y se plasmaron en una gráfica de probabilidad normal. Los efectos calculados son los siguientes:

$$\begin{aligned} A &= 10.5613, & BD &= -2.2787, & B &= 12.4463, \\ C &= 2.4138, & D &= 2.1438, & AB &= 11.4038, \\ AC &= 1.2613, & AD &= -1.8238, & BC &= 1.8163, \\ CD &= 1.4088, & ABC &= 2.8588, & ABD &= -1.7813, \\ ACD &= -3.0438, & BCD &= -0.4788, & ABCD &= -1.3063. \end{aligned}$$

La gráfica de probabilidad normal se muestra en la figura 15.5. La gráfica parece implicar que los efectos A, B y AB son importantes. Los signos de los efectos importantes indican que las conclusiones preliminares son las siguientes:

1. Un incremento en la velocidad de inyección de 1.0 a 2.0 aumenta el derrame.
2. Un aumento en la temperatura del molde, de 100 °C a 150 °C incrementa el derrame.
3. Hay una interacción entre la velocidad de inyección y la temperatura del molde; aunque ambos efectos principales son importantes, es crucial que se entienda el efecto de la interacción de los dos factores.

Análisis con error cuadrático medio agrupado: Salida anotada de computadora

Puede ser de interés observar un análisis de varianza de los datos del moldeo por inyección con interacciones de orden superior agrupadas para formar un error cuadrático medio. Las interacciones de orden tres y cuatro están agrupadas. En la figura 15.6 se muestra una salida de *SAS* PROC GLM. El análisis de varianza revela en esencia la misma conclusión que la gráfica de probabilidad normal.

Las pruebas y los valores P que se observan en la figura 15.6 requieren una interpretación. Un valor significativo P sugiere que el efecto difiere de cero en forma significativa. Las pruebas sobre los efectos principales (que en presencia de las interacciones pueden considerarse como los efectos promediados sobre el nivel de los demás factores) indican la significancia para los efectos A y B. Los signos de los

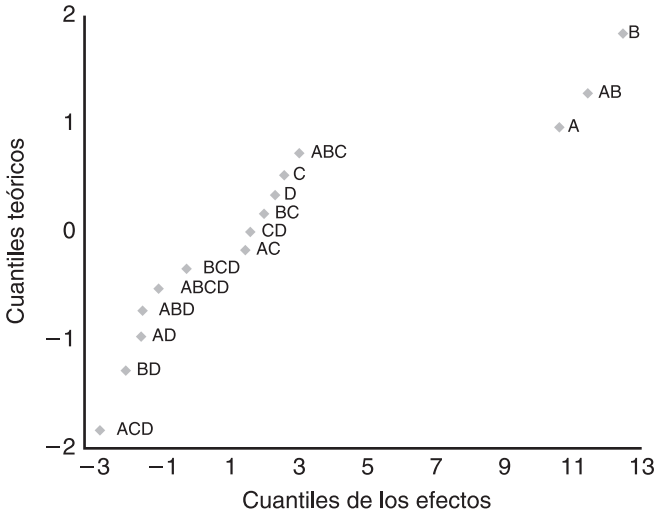


Figura 15.5: Gráfica cuantil-cuantil normal de los efectos para el estudio de caso del ejemplo 15.2.

efectos también son importantes. Un aumento en los niveles de bajo a alto en A, la velocidad de inyección, ocasiona un incremento del derrame. Sucede lo mismo para B. Sin embargo, debido a la interacción significativa AB, las interpretaciones del efecto principal se ven como tendencias a través de los niveles de los demás factores. El impacto de la interacción AB significativa se entiende mejor si se emplea una tabla de medias de dos factores.

Interpretación de la interacción de dos factores

Como se esperaría, una tabla de medias de dos factores debería facilitar la interpretación de la interacción AB. Considere la situación de dos factores de la tabla 15.7.

Tabla 15.7: Ilustración de la interacción de dos factores

	***B* (temperatura)**	
***A* (velocidad)**	**100**	**150**
2	73.355	97.205
1	74.1975	75.240

Note que la media muestral grande a velocidad y temperatura elevadas creó la interacción significativa. El **derrame se incrementa en forma no aditiva**. La temperatura del molde parece tener un efecto positivo a pesar del nivel de velocidad. Sin embargo, el efecto es el mayor a velocidad elevada. El efecto de la velocidad es muy ligero a temperaturas bajas; pero es claramente positivo para una temperatura alta del molde. Para controlar el derrame a bajo nivel, *debería evitarse el uso simultáneo de una velocidad alta de inyección y una temperatura del molde elevada.* Todos estos resultados se ilustran en forma gráfica en la figura 15.7.

```
                        The GLM Procedure
Dependent Variable: y
                              Sum of
Source                DF      Squares  Mean Square  F Value  Pr > F
Model                 10  1689.237462   168.923746     9.37  0.0117
Error                  5    90.180831    18.036166
Corrected Total       15  1779.418294
R-Square     Coeff Var      Root MSE        y Mean
0.949320      5.308667      4.246901      79.99938
Source      DF   Type III SS   Mean Square  F Value    Pr > F
A            1   446.1600062   446.1600062    24.74    0.0042
B            1   619.6365563   619.6365563    34.36    0.0020
C            1    23.3047563    23.3047563     1.29    0.3072
D            1    18.3826563    18.3826563     1.02    0.3590
A*B          1   520.1820562   520.1820562    28.84    0.0030
A*C          1     6.3630063     6.3630063     0.35    0.5784
A*D          1    13.3042562    13.3042562     0.74    0.4297
B*C          1    13.1950562    13.1950562     0.73    0.4314
B*D          1    20.7708062    20.7708062     1.15    0.3322
C*D          1     7.9383063     7.9383063     0.44    0.5364
                              Standard
Parameter      Estimate          Error  t Value    Pr > |t|
Intercept   79.99937500    1.06172520    75.35      <.0001
A            5.28062500    1.06172520     4.97      0.0042
B            6.22312500    1.06172520     5.86      0.0020
C            1.20687500    1.06172520     1.14      0.3072
D            1.07187500    1.06172520     1.01      0.3590
A*B          5.70187500    1.06172520     5.37      0.0030
A*C          0.63062500    1.06172520     0.59      0.5784
A*D         -0.91187500    1.06172520    -0.86      0.4297
B*C          0.90812500    1.06172520     0.86      0.4314
B*D         -1.13937500    1.06172520    -1.07      0.3322
C*D          0.70437500    1.06172520     0.66      0.5364
```

Figura 15.6: Salida del *SAS* para los datos del estudio de caso del ejemplo 15.2.

Ejercicios

15.1 Los siguientes datos se obtuvieron de un experimento factorial 2^3 que se repitió tres veces. Con el método del contraste, evalúe las sumas de los cuadrados para todos los efectos factoriales. Saque sus conclusiones.

Combinación de tratamientos	Rep 1	Rep 2	Rep 3
(1)	12	19	10
a	15	20	16
b	24	16	17
ab	23	17	27
c	17	25	21
ac	16	19	19
bc	24	23	29
abc	28	25	20

15.2 En un experimento efectuado por el Departamento de Ingeniería de Minas del Instituto Politécnico y Universidad Estatal de Virginia, para estudiar un sistema

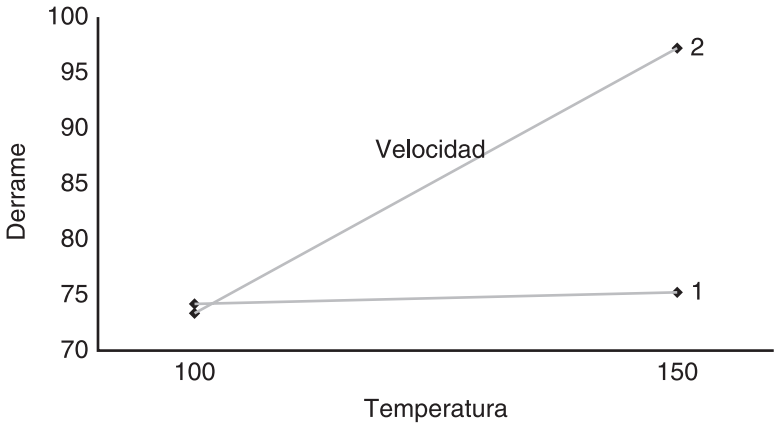


Figura 15.7: Gráfica de la interacción para el ejemplo 15.2.

de filtrado particular para carbón, se agregó un coagulante a la solución en un tanque que contenía carbón y sedimentos, que luego se puso en un sistema de recirculación para que el carbón se lavara. Tres factores variaron en el proceso experimental:

Factor A:	Porcentaje de sólidos circulados inicialmente en el flujo hacia adelante.
Factor B:	Tasa de flujo del polímero.
Factor C:	pH del tanque.

La cantidad de sólidos en el flujo inferior del sistema de purificación determina qué tan limpio ha quedado el carbón. Se emplearon dos niveles de cada factor y se hicieron dos corridas experimentales para cada una de las $2^3 = 8$ combinaciones. Las respuestas, sólidos porcentuales por peso, en el flujo inferior del sistema de circulación, se especifican en la tabla siguiente:

Combinación de tratamientos	**Respuesta**	
	Replicación 1	**Replicación 2**
(1)	4.65	5.81
a	21.42	21.35
b	12.66	12.56
ab	18.27	16.62
c	7.93	7.88
ac	13.18	12.87
bc	6.51	6.26
abc	18.23	17.83

Si se supone que todas las interacciones son potencialmente importantes, haga un análisis completo de los datos. Use valores P en la conclusión.

15.3 En un experimento metalúrgico se desea probar el efecto de cuatro factores y sus interacciones sobre la concentración (peso en porcentaje) de cierto compuesto particular de fósforo en el material de fundición. Las variables son A, porcentaje de fósforo en la refinación; B, porcentaje del material vuelto a fundir; C, tiempo de flujo; y D, tiempo de espera. Se varían los cuatro factores en un experimento factorial 2^4 con dos fundiciones tomadas a cada combinación de factores. Las 32 fundiciones se hicieron en orden aleatorio. La tabla siguiente muestra los datos y la tabla ANOVA se da en la figura 15.8 de la página 626. Analice los efectos de los factores y sus interacciones sobre la concentración del compuesto de fósforo.

Combinación de tratamientos	**Peso Porcentaje de compuesto de fósforo**		
	Rep 1	**Rep 2**	**Total**
(1)	30.3	28.6	58.9
a	28.5	31.4	59.9
b	24.5	25.6	50.1
ab	25.9	27.2	53.1
c	24.8	23.4	48.2
ac	26.9	23.8	50.7
bc	24.8	27.8	52.6
abc	22.2	24.9	47.1
d	31.7	33.5	65.2
ad	24.6	26.2	50.8
bd	27.6	30.6	58.2
abd	26.3	27.8	54.1
cd	29.9	27.7	57.6
acd	26.8	24.2	51.0
bcd	26.4	24.9	51.3
abcd	26.9	29.3	56.2
Total	428.1	436.9	865.0

15.4 Se realizó un experimento preliminar para estudiar los efectos de cuatro factores y sus interacciones sobre la salida de cierta operación de maquinado. Se hi-

cieron dos corridas en cada una de las combinaciones de tratamientos para obtener una medida del error experimental puro. Se emplearon dos niveles de cada factor, y se obtuvieron los datos que se muestran en seguida. Efectúe pruebas sobre todos los efectos principales y las interacciones al nivel de significancia de 0.05. Saque sus conclusiones.

Combinación de tratamiento	Réplica 1	Réplica 2
(1)	7.9	9.6
a	9.1	10.2
b	8.6	5.8
c	10.4	12.0
d	7.1	8.3
ab	11.1	12.3
ac	16.4	15.5
ad	7.1	8.7
bc	12.6	15.2
bd	4.7	5.8
cd	7.4	10.9
abc	21.9	21.9
abd	9.8	7.8
acd	13.8	11.2
bcd	10.2	11.1
abcd	12.8	14.3

15.5 En el estudio *An X-Ray Fluorescence Method for Analyzing Polybutadiene-Acrylic Acid (PBAA) Propellants*, Quarterly Reports, RK-TR-62-1, Army Ordnance Missile Command, se realizó un experimento para determinar si hay o no una diferencia significativa en la cantidad de aluminio alcanzado en el análisis entre ciertos niveles de algunas variables de procesamiento. En la tabla que sigue se presentan los datos registrados.

Obs.	Estado físico	Tiempo de mezclado	Velocidad de las aspas	Condición de nitrogeno	Aluminio
1	1	1	2	2	16.3
2	1	2	2	2	16.0
3	1	1	1	1	16.2
4	1	2	1	2	16.1
5	1	1	1	2	16.0
6	1	2	1	1	16.0
7	1	2	2	1	15.5
8	1	1	2	1	15.9
9	2	1	2	2	16.7
10	2	2	2	2	16.1
11	2	1	1	1	16.3
12	2	2	1	2	15.8
13	2	1	1	2	15.9
14	2	2	1	1	15.9
15	2	2	2	1	15.6
16	2	1	2	1	15.8

A continuación se dan las variables.

A: Tiempo de mezclado
nivel 1-2 horas
nivel 2-4 horas

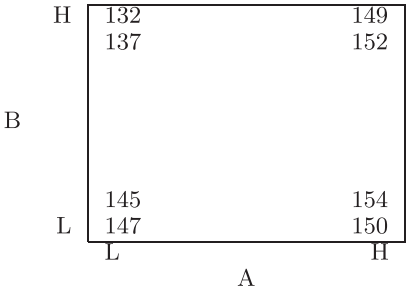
B: Velocidad de las aspas
nivel 1-36 rpm
nivel 2-78 rpm

C: Condición de nitrógeno pasado sobre el combustible
nivel 1-seco
nivel 2-72% de humedad relativa

D: Estado físico del combustible
nivel 1-no refinado
2-refinado

Haga el análisis de los datos si se supone que todas las interacciones de tres y cuatro factores son despreciables. Utilice un nivel de significancia de 0.05. Escriba un breve informe que resuma sus descubrimientos.

15.6 Es importante estudiar el efecto de la concentración del reactivo y la tasa de alimentación de la viscosidad del producto de cierto proceso químico. La concentración del reactivo será el factor *A* a los niveles 15 y 25%. La tasa de alimentación será el factor *B* con niveles de 20 lb/h y 30 lb/h. El experimento implica 2 corridas experimentales en cada una de las cuatro combinaciones (L = bajo y H = alto). Las lecturas de la viscosidad son las siguientes.



a) Suponga un modelo que contenga dos efectos principales y una interacción; calcule los tres efectos. ¿Tiene usted alguna interpretación en este momento?

b) Realice un análisis de varianza y haga pruebas para la interacción. Dé sus conclusiones.

c) Realice pruebas para los efectos principales y dé sus conclusiones finales acerca de la importancia de todos estos efectos.

15.7 Considere de nuevo el ejercicio 15.3. Para el investigador es de interes saber que no sólo las interacciones *AD*, *BC* y quizá *AB* son importantes. Aunque también es de interés lo que significan de manera científica. Muestre la gráfica de interacción bidemensional para las tres y dé su interpretación.

15.8 Considere de nuevo el ejercicio 15.3. Es frecuente que las interacciones de tres factores no sean significativas y, aun si lo fueran, serían difíciles de interpretar. La interacción *ABD* parece ser importante. Para obtener algún sentido de interpretación, haga dos gráficas de la interacción *AD*, una para $B = -1$ y otra para $B = +1$. A partir de la apariencia de éstas, dé una interpretación de la interacción *ABD*.

15.9 Considere el ejercicio 15.6. Utilice una escala de "+1" y "−1", para "alto" y "bajo", respectivamente, y realice una regresión lineal múltiple con el modelo

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

con x_{1i}, = concentración del reactivo (−1, +1) y x_{2i} = tasa de alimentación (−1, +1).

a) Calcule los coeficientes de regresión.

b) ¿Cómo se relacionan los coeficientes b_1, b_2 y b_{12} con los efectos que el lector encontró en el ejercicio 15.6*a*)?

c) En su análisis de regresión haga pruebas t sobre b_1, b_2 y b_{12}. ¿Cómo se relacionan estos resultados de la prueba con aquellos del ejercicio 15.6*b*) y *c*)?

15.10 Considere el ejercicio 15.5. Calcule los 15 efectos y haga gráficas de probabilidad normal de los efectos.

a) ¿Parece válida la suposición de que son despreciables las interacciones de tres y cuatro factores?

b) ¿Los resultados de las gráficas del efecto son consistentes con lo que el lector concluyó sobre la importancia de los efectos principales y las interacciones de dos factores, en su informe de resumen?

15.11 En Myers y Montgomery (2002), se analiza un conjunto de datos para el que un ingeniero empleó un factorial 2^3 para estudiar los efectos de la velocidad de corte (A), la geometría de la herramienta (B) y el ángulo de corte (C), sobre la vida (en horas) de una máquina herramienta. Se eligen dos niveles de cada factor, y se corrieron duplicados en cada punto del diseño en un orden aleatorio. A continuación se presentan los datos.

a) Calcule los siete efectos. ¿Con base en su magnitud, cuál parece ser importante?

b) Haga un análisis de varianza y observe los valores P.

c) ¿Concuerdan los resultados de los incisos *a*) y *b*)?

d) El ingeniero tiene confianza en que deben interactuar la velocidad y el ángulo de corte en que se hacen. Si esta interacción es significativa, haga una gráfica de la interacción y analice el significado de la interacción desde el punto de vista de la ingeniería.

	A	B	C	Vida
(1)	−	−	−	22, 31
a	+	−	−	32, 43
b	−	+	−	35, 34
ab	+	+	−	35, 47
c	−	−	+	44, 45
ac	+	−	+	40, 37
bc	−	+	+	60, 50
abc	+	+	+	39, 41

15.12 Considere el ejercicio 15.11. Suponga que hubo cierta dificultad experimental para hacer las corridas. En realidad, todo el experimento tuvo que ser detenido después de solo cuatro corridas. Como resultado, el experimento abreviado está dado por

	Vida
a	43
b	35
c	44
abc	39

Con sólo estas corridas, los signos para los contrastes están dados por

	A	B	C	AB	AC	BC	ABC
a	+	−	−	−	−	+	+
b	−	+	−	−	+	−	+
c	−	−	+	+	−	−	+
abc	+	+	+	+	+	+	+

Haga comentarios. Como parte de ellos, determine si los contrastes son ortogonales o no. ¿Cuáles lo son y cuáles no? ¿Los efectos principales son ortogonales entre sí? En ese experimento abreviado (titulado *factorial fraccionario*), ¿se puede estudiar las interacciones en forma independiente de los efectos principales? ¿Un experimento es útil si se está convencido de que las interacciones son despreciables? Explique su respuesta.

15.5 Experimentos factoriales en la preparación de la regresión

En gran parte de este capítulo 15, nuestro análisis de los datos para un factorial 2^k se ha limitado hasta este momento al método del análisis de varianza. La única referencia a un análisis alternativo se hizo en el ejercicio 15.9 de esta página. En realidad, ese ejercicio introduce mucho de lo que es la motivación de la presente sección. Hay situaciones donde el ajuste del modelo es importante **y pueden controlarse** los factores que se estudian. Por ejemplo, un biólogo podría querer estudiar el crecimiento de cierto tipo de alga en el agua, y en ese caso sería muy útil un modelo que relacionara las unidades de algas como función de la *cantidad de cierto contaminante*, y, digamos, el *tiempo*. Así, el estudio requiere un experimento factorial en una preparación de laboratorio, en el que los factores son la concentración del contaminante y el tiempo. Como se verá más adelante en esta sección, puede ajustarse un

Fuente de variación	**Efectos**	**Suma de cuadrados**	**Grados de cuadrado**	**Libertad media**	***f* calculada**	**Valor *P***
Efecto principal :						
A	−1.2000	11.52	1	11.52	4.68	0.0459
B	−1.2250	12.01	1	12.01	4.88	0.0421
C	−2.2250	39.61	1	39.61	16.10	0.0010
D	1.4875	17.70	1	17.70	7.20	0.0163
Interacción de dos factores :						
AB	0.9875	7.80	1	7.80	3.17	0.0939
AC	−0.6125	3.00	1	3.00	1.22	0.2857
AD	−1.3250	14.05	1	14.05	5.71	0.0295
BC	1.1875	11.28	1	11.28	4.59	0.0480
BD	0.6250	3.13	1	3.13	1.27	0.2763
CD	0.7000	3.92	1	3.92	1.59	0.2249
Interacción de tres factores :						
ABC	−0.5500	2.42	1	2.42	0.98	0.3360
ABD	1.7375	24.15	1	24.15	9.82	0.0064
ACD	1.4875	17.70	1	17.70	7.20	0.0163
BCD	−0.8625	5.95	1	5.95	2.42	0.1394
Interacción de cuatro factores :						
$ABCD$	0.7000	3.92	1	3.92	1.59	0.2249
Error		39.36	16	2.46		
Total		217.51	31			

Figura 15.8: Tabla ANOVA para el ejercicio 15.3.

modelo más preciso si los factores están controlados en un arreglo factorial, para el que con frecuencia es útil la selección de un factorial 2^k. En muchos procesos biológicos y químicos, los niveles de las variables regresoras pueden y deberían controlarse.

Hay que recordar que el modelo de regresión empleado en el capítulo 12 puede escribirse con notación matricial como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

La matriz $\mathbf{X}$ se denomina **matriz del modelo**. Por ejemplo, suponga que se emplea un experimento factorial 2^3 con las variables

Temperatura:	150 °C	200 °C
Humedad	15%	20%
Presión (psi):	1000	1500

Los niveles familiares +1 y −1 se generan a través de centrar y dar escala a las siguientes *unidades de diseño*:

$$x_1 = \frac{\text{temperatura} - 175}{25}, \quad x_2 = \frac{\text{humedad} - 17.5}{2.5}, \quad x_3 = \frac{\text{presión} - 1250}{250}.$$

Como resultado, la matriz $\mathbf{X}$ resulta ser la siguiente

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} & x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \begin{array}{c} \text{Identificación para el diseño} \\ (1) \\ a \\ b \\ c \\ ab \\ ac \\ bc \\ abc \end{array}$$

Ahora se observa que los contrastes ilustrados y analizados en la sección 15.2 están relacionados directamente con los coeficientes de regresión. Observe que todas las columnas de la matriz $\mathbf{X}$ en el ejemplo 2^3, son *ortogonales*. Como resultado, el cálculo de los coeficientes de regresión que se describió en la sección 12.3 se convierte en

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(\frac{1}{8}\mathbf{I}\right)\mathbf{X}'\mathbf{y}$$
$$= \frac{1}{8}\begin{bmatrix} a + ab + ac + abc + (1) + b + c + bc \\ a + ab + ac + abc - (1) - b - c - bc \\ b + ab + bc + abc - (1) - a - c - ac \\ c + ac + bc + abc - (1) - a - b - ab \end{bmatrix},$$

donde a, ab, etcétera, son mediciones de las respuestas.

Se observa que la noción de *principales efectos calculados*, de los que se ha hecho énfasis en todo este capítulo, con 2^k factoriales, se relaciona con los coeficientes de un modelo de ajuste por regresión cuando los factores son cuantitativos. En realidad, para un 2^k con, digamos, n corridas experimentales por punto de diseño, las relaciones entre los efectos y los coeficientes de regresión son como sigue:

$$\text{Efecto} = \frac{\text{contraste}}{2^{k-1}(n)}$$
$$\text{Coeficiente de regresión} = \frac{\text{contraste}}{2^k(n)} = \frac{\text{efecto}}{2}.$$

Esta relación debería tener sentido para el lector, ya que un coeficiente de regresión b_j es una tasa de cambio promedio en respuesta *por unidad de cambio* de x_j. Por supuesto, cuando se va de -1 a $+1$ en x_j (bajo a alto), la variable de diseño ha cambiado en 2 unidades.

Ejemplo 15.3: Considere un experimento donde un ingeniero desea ajustar una regresión lineal del producto y contra el tiempo de espera x_1 y el tiempo de doblado x_2 en cierto sistema químico. Todos los demás factores se mantienen fijos. Los datos en las unidades naturales se dan en la tabla 15.8. Estime el modelo de regresión lineal múltiple.

Solución: Como resultado, el modelo de regresión ajustada es

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

Tabla 15.8: Datos para el ejemplo 15.3

Tiempo de espera (h)	**Tiempo de doblado (h)**	**Producto (%)**
0.5	0.10	28
0.8	0.10	39
0.5	0.20	32
0.8	0.20	46

Las unidades de diseño son

$$x_1 = \frac{\text{tiempo de espera} - 0.65}{0.15}, \qquad x_2 = \frac{\text{tiempo de doblado} - 0.15}{0.05}$$

y la matriz $\mathbf{X}$ es

$$\begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} & x_1 & x_2 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

con los coeficientes de regresión

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{(1)+a+b+ab}{4} \\ \frac{a+ab-(1)-b}{4} \\ \frac{b+ab-(1)-a}{4} \end{bmatrix} = \begin{bmatrix} 36.25 \\ 6.25 \\ 2.75 \end{bmatrix}.$$

Así, la ecuación de regresión de mínimos cuadrados es

$$\hat{y} = 36.25 + 6.25x_1 + 2.75x_2.$$

Este ejemplo proporciona una ilustración del uso del experimento factorial de dos niveles en un planteamiento de regresión. Las cuatro corridas experimentales en el diseño 2^2 se usaron para obtener una ecuación de regresión, con la interpretación evidente de los coeficientes de regresión. El valor $b_1 = 6.25$ representa el incremento estimado de la respuesta (salida porcentual) por *unidad de diseño* que cambia (0.15 horas) el tiempo de espera. El valor $b_2 = 2.75$ representa una tasa de cambio similar para el tiempo de doblado.

Interacción en el modelo de regresión

Los contrastes de las interacciones que se estudiaron en la sección 15.2, tienen interpretaciones definidas en el contexto de la regresión. En realidad, las interacciones son tomadas en cuenta en los modelos de regresión mediante los términos que son productos. Esto se ilustra con el ejemplo 15.3, en el cual el modelo con interacción es

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2$$

con b_0, b_1 y b_2, como antes, y

$$b_{12} = \frac{ab + (1) - a - b}{4} = \frac{46 + 28 - 39 - 32}{4} = 0.75.$$

Así, la ecuación de regresión que expresa dos *efectos principales lineales* e interacción, es

$$\hat{y} = 36.25 + 6.25x_1, +2.75x_2 + 0.75x_1x_2.$$

El contexto de la regresión proporciona un marco de referencia con el cual el lector debería entener mejor la ventaja de la ortogonalidad de que se disfruta con el factorial 2^k. En la sección 15.2, los méritos de la ortogonalidad se analizan desde el punto de vista del *análisis de la varianza* de los datos en un experimento factorial 2^k. Se aclaró que la ortogonalidad entre los efectos conduce a la independencia entre las sumas de los cuadrados. Por supuesto, la presencia de variables de regresión no rige el uso del análisis de varianza. De hecho, las pruebas F se llevan a cabo tal como se describió en la sección 15.2. Salta a la vista que debe hacerse una distinción. En el caso del ANOVA, las hipótesis evolucionan a partir de medias poblacionales; en tanto que en el caso de la regresión las hipótesis implican coeficientes de regresión.

Por ejemplo, considere el diseño experimental del ejercicio 15.2 de la página 622. Cada factor es continuo y suponga que los niveles son los siguientes

A (x_1):	20%	40%
B (x_2):	5 lb/seg	10 lb/seg
C (x_3):	5	5.5

y que se tiene, para los niveles de diseño,

$$x_1 = \frac{(\text{sólidos} - 30)}{10}, \quad x_2 = \frac{(\text{tasa de flujo} - 7.5)}{2.5}, \quad x_3 = \frac{(\text{pH} - 5.25)}{0.25}.$$

Suponga que es de interés ajustar un modelo de regresión múltiple, en el cual tengan que considerarse todos los coeficientes lineales y las interacciones disponibles. Además, al ingeniero le interesa dar alguna perspectiva acerca de qué niveles del factor *maximizarán* la limpieza (es decir, maximizar la respuesta). Este problema será un estudio de caso en el ejemplo 15.4.

Ejemplo 15.4: **Estudio de caso: Experimento de limpieza de carbón**[1] La figura 15.9 representa una salida anotada por computadora para el análisis de regresión del modelo ajustado

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3,$$

donde x_1, x_2 y x_3 son sólidos porcentuales, tasa de flujo y pH del sistema, respectivamente. El sistema de cómputo que se usó es el *SAS* PROC REG.

En la salida, observe los estimadores del parámetro, el error estándar y los valores P. Los estimadores del parámetro representan los coeficientes del modelo. Todos éstos son significativos, excepto el término x_2x_3 (interacción BC). También note que los residuos, los intervalos de confianza y los intervalos de predicción aparecen según se estudiaron en el material sobre regresión de los capítulos 11 y 12.

El lector puede usar los valores de los coeficientes del modelo y pronosticar otros valores a partir de la salida, para asegurarse de que la combinación de los factores dé como resultado la **eficiencia máxima de limpieza**. El factor A (sólidos porcentuales circulados) tiene un coeficiente positivo grande, lo cual sugiere que un valor elevado para los sólidos porcentuales. Además, se sugiere un valor bajo para el factor C (pH del tanque). Aunque el coeficiente del efecto principal B (tasa de flujo del

[1] Véase el ejercicio 15.2.

```
Dependent Variable: Y
                    Analysis of Variance
                         Sum of      Mean
Source           DF     Squares    Square  F Value  Pr > F
Model             7  490.23499  70.03357   254.43  <.0001
Error             8    2.20205   0.27526
Corrected Total  15  492.43704
Root MSE         0.52465    R-Square     0.9955
Dependent Mean  12.75188    Adj R-Sq     0.9916
Coeff Var        4.11429
                  Parameter Estimates
               Parameter  Standard
Variable   DF   Estimate     Error   t Value  Pr > |t|
Intercept   1   12.75188   0.13116     97.22    <.0001
A           1    4.71938   0.13116     35.98    <.0001
B           1    0.86563   0.13116      6.60    0.0002
C           1   -1.41563   0.13116    -10.79    <.0001
AB          1   -0.59938   0.13116     -4.57    0.0018
AC          1   -0.52813   0.13116     -4.03    0.0038
BC          1    0.00562   0.13116      0.04    0.9668
ABC         1    2.23063   0.13116     17.01    <.0001
    Dependent Predicted    Std Error
Obs  Variable     Value Mean Predict    95% CL Mean       95% CL Predict Residual
  1    4.6500    5.2300       0.3710   4.3745   6.0855   3.7483   6.7117  -0.5800
  2   21.4200   21.3850       0.3710  20.5295  22.2405  19.9033  22.8667   0.0350
  3   12.6600   12.6100       0.3710  11.7545  13.4655  11.1283  14.0917   0.0500
  4   18.2700   17.4450       0.3710  16.5895  18.3005  15.9633  18.9267   0.8250
  5    7.9300    7.9050       0.3710   7.0495   8.7605   6.4233   9.3867   0.0250
  6   13.1800   13.0250       0.3710  12.1695  13.8805  11.5433  14.5067   0.1550
  7    6.5100    6.3850       0.3710   5.5295   7.2405   4.9033   7.8667   0.1250
  8   18.2300   18.0300       0.3710  17.1745  18.8855  16.5483  19.5117   0.2000
  9    5.8100    5.2300       0.3710   4.3745   6.0855   3.7483   6.7117   0.5800
 10   21.3500   21.3850       0.3710  20.5295  22.2405  19.9033  22.8667  -0.0350
 11   12.5600   12.6100       0.3710  11.7545  13.4655  11.1283  14.0917  -0.0500
 12   16.6200   17.4450       0.3710  16.5895  18.3005  15.9633  18.9267  -0.8250
 13    7.8800    7.9050       0.3710   7.0495   8.7605   6.4233   9.3867  -0.0250
 14   12.8700   13.0250       0.3710  12.1695  13.8805  11.5433  14.5067  -0.1550
 15    6.2600    6.3850       0.3710   5.5295   7.2405   4.9033   7.8667  -0.1250
 16   17.8300   18.0300       0.3710  17.1745  18.8855  16.5483  19.5117  -0.2000
```

Figura 15.9: Salida del *SAS* para los datos del ejemplo 15.4.

polímero) es positivo, el coeficiente positivo aún más grande de $x_1x_2x_3$ (ABC) sugeriría que la tasa de flujo debería estar en el nivel bajo con la finalidad de mejorar la eficiencia. Aún más, el modelo de regresión generado en la salida del *SAS* sugiere que la combinación de factores que producen resultados óptimos, o quizá sugieran la dirección para la experimentación adicional, está dada por

A: nivel alto
B: nivel bajo
C: nivel bajo

15.6 El diseño ortogonal

En situaciones experimentales en las que es apropiado ajustar modelos que son lineales en las variables de diseño, y posiblemente impliquen interacciones o términos que son productos, existen ventajas que surgen del *diseño ortogonal* de dos niveles, o arreglo ortogonal, que quiere decir que hay ortogonalidad entre las columnas de la matriz $\mathbf{X}$. Considere la matriz $\mathbf{X}$ para el factorial 2^2 del ejemplo 15.3. Note que las tres columnas son mutuamente ortogonales. La matriz $\mathbf{X}$ del factorial 2^3 también contiene columnas ortogonales. El factorial 2^3 con interacciones produciría una matriz $\mathbf{X}$ del tipo

$$\mathbf{X} = \begin{array}{c} \begin{matrix} \quad & x_1 & x_2 & x_3 & x_1x_2 & x_1x_3 & x_2x_3 & x_1x_2x_3 \end{matrix} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

El panorama de los grados de libertad es

Fuente	g.l.	
Regresión	3	
Falta de ajuste	4	$(x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)$
Error (puro)	8	
Total	15	

Los ocho grados de libertad para el error puro se obtienen a partir de las *corridas duplicadas* en cada punto del diseño. La falta de ajuste de los grados de libertad puede verse como la diferencia entre el número de puntos de diseño distintos y el número total de términos en el modelo; en este caso, hay 8 puntos y 4 términos en el modelo.

Error estándar de los coeficientes y pruebas *t*

En las secciones anteriores vimos cómo el analista de un experimento puede aprovechar el concepto de ortogonalidad para diseñar un experimento de regresión, con coeficientes que tengan varianza mínima sobre la base del costo. Se debe ser capaz de utilizar el conocimiento de la regresión que se expuso en la sección 12.4 para calcular estimadores de las varianzas de los coeficientes y, con ello, los errores estándar. También resulta de interés resaltar la relación entre el estadístico t sobre un coeficiente y el estadístico F descrito e ilustrado en capítulos anteriores.

Recuerde el lector que en la sección 12.4 se vio que las varianzas y las covarianzas de los coeficientes aparecen en la matriz A^{-1}, o, en términos de la notación actual, la *matriz de varianza-covarianza* de coeficientes es

$$\sigma^2 A^{-1} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

En el caso del experimento factorial 2^k, las columnas de $\mathbf{X}$ son mutuamente ortogonales,

lo que impone una estructura muy especial. En general, para 2^k se tiene que

$$\begin{array}{cccccccc} & & x_1 & x_2 & \cdots & x_k & x_1x_2 & \cdots \\ \mathbf{X} = [\mathbf{1} & & \pm\mathbf{1} & \pm\mathbf{1} & \cdots & \pm\mathbf{1} & \pm\mathbf{1} & \cdots], \end{array}$$

donde cada columna contiene 2^k entradas o $2^k n$, donde n es el número de corridas repetidas en cada punto del diseño. Así, la formación de $\mathbf{X}'\mathbf{X}$ lleva a

$$\mathbf{X}'\mathbf{X} = 2^k n\mathbf{I}_p,$$

donde $\mathbf{I}$ es la matriz de identidad de la dimensión p, el número de parámetros del modelo.

Ejemplo 15.5: Considere un 2^3 con corridas por duplicado que se ajusta al modelo

$$E(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_{12}x_1x_2 + \beta_{13}x_1x_3 + \beta_{23}x_2x_3.$$

Dé expresiones para los errores estándar de los estimadores de mínimos cuadrados de b_0, b_1, b_2, b_3, b_{12}, b_{13} y b_{23}.

Solución:

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & x_1 & x_2 & x_3 & x_1x_2 & x_1x_3 & x_2x_3 \end{array} \\ \left[\begin{array}{ccccccc} 1 & -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}\right] \end{array}$$

con cada unidad vista como *repetida* (es deicr, cada observación está duplicada). Como resultado,

$$\mathbf{X}'\mathbf{X} = 16\mathbf{I}_7.$$

Así,

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{16}\mathbf{I}_7.$$

De lo anterior, queda claro que las varianzas de todos los coeficientes para un factorial 2^k con n corridas en cada punto de diseño son

$$Var(b_j) = \frac{\sigma^2}{2^k n},$$

y, por supuesto, todas las covarianzas son igual a cero. Como resultado, los errores estándar de los coeficientes se calculan como

$$s_{b_j} = s\sqrt{\frac{1}{2^k n}},$$

donde s se encuentra con la raíz cuadrada del error cuadrático medio (se tiene la esperanza de obtenerlo a partir de una repetición adecuada). Así, en nuestro caso de 2^3,

$$s_{b_j} = s\left(\frac{1}{4}\right).$$

Ejemplo 15.6: Considere el experimento de la metalurgia del ejercicio 15.3 de la página 623. Suponga que el modelo ajustado es

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 \triangleright.$$

¿Cuáles son los errores estándar de los coeficientes de regresión de mínimos cuadrados?

Solución: Los errores estándar de todos los coeficientes para el factorial 2^k son iguales, y son

$$s_{b_j} = s\sqrt{\frac{1}{2^k n}},$$

que en esta ilustración es

$$s_{b_j} = s\sqrt{\frac{1}{(16)(2)}}.$$

En este caso, el error cuadrático medio puro está dado por $s^2 = 2.46$ (16 grados de libertad). Entonces,

$$s_{b_j} = 0.28.$$

Los errores estándar de los coeficientes se usan para construir estadísticos t sobre todos los coeficientes. Estos valores t se relacionan con los estadísticos F del análisis de varianza. Ya se demostró que un estadístico F sobre un coeficiente, usando el factorial 2^k, es

$$F = \frac{(\text{contraste})^2}{(2^k)(n)s^2}.$$

Ésta es la forma del estadístico F de la página 626 para el experimento de metalurgia (véase el ejercicio 15.3). Es fácil comprobar que si se escribe

$$t = \frac{b_j}{s_{b_j}}, \qquad \text{cuando} \qquad b_j = \frac{\text{contraste}}{2^k n},$$

entonces

$$t^2 = \frac{(\text{contraste})^2}{s^2 2^k n} = F.$$

Como resultado, la relación usual se cumple entre estadísticos t sobre los coeficientes y los valores F. Como era de esperar, la única diferencia en el uso de t o F para estudiar la significancia está en el hecho de que el estadístico t indica el signo o la dirección del efecto del coeficiente.

Parecería que el plan del factorial 2^k se adapta a muchas situaciones prácticas a las cuales se ajustan modelos de regresión. Alberga términos lineales y de interacción, lo que da estimadores óptimos de todos los coeficientes (desde un punto de vista de la varianza). Sin embargo, cuando k es grande, el número de puntos del diseño requerido es muy grande. Es frecuente que puedan usarse partes del diseño total y todavía haya ortogonalidad, con todas las ventajas que ello implica. En la sección 15.8, a continuación, se estudian tales diseños.

Una mirada más amplia a la propiedad de ortogonalidad en el factorial 2^k

Ya vimos que para el caso del factorial 2^k toda la información que obtiene el analista sobre los efectos y las interacciones principales está en la forma de contrastes. Estas "2^{k-1} piezas de información" conllevan un solo grado de libertad cada una y son independientes entre sí. En un análisis de varianza se manifiestan como *efectos*; mientras que si lo que se construye es un modelo de regresión, los efectos son los coeficientes de la regresión, aparte de un factor de 2. Con cada forma de análisis, es posible hacer pruebas de significancia y las pruebas t para un efecto dado son las mismas en cuanto a los resultados numéricos, que para los coeficientes de regresión correspondientes. En el caso del ANOVA, son importantes la exposición de las variables y la interpretación científica de las interacciones; en tanto que en el caso de un análisis de regresión, se usa un modelo para predecir la respuesta y/o determinar cuáles combinaciones de nivel de factores son las óptimas (es decir, maximizan la salida o la eficiencia de la limpieza, como en el estudio de caso del ejemplo 15.4).

Resulta que la propiedad de ortogonalidad es importante sea que el análisis consista en ANOVA o regresión. La ortogonalidad entre las columnas de X, la matriz del modelo en, digamos, el ejemplo 15.5, brinda condiciones especiales que tiene una influencia importante sobre los **efectos de la varianza** o los **coeficientes de regresión**. En realidad, ya se ha hecho evidente que el diseño ortogonal da como resultado la igualdad de varianza para todos los efectos o coeficientes. Así, de ese modo, para propósitos de estimación o de prueba, la precisión es la misma para todos los coeficientes, los efectos principales o las interacciones. Además, si el modelo de regresión sólo contiene términos lineales y, por ello, sólo los efectos principales son de interés, las condiciones siguientes dan como resultado la minimización de las varianzas de todos los efectos (o, en forma correspondiente, de los coeficientes de regresión de primer orden).

Condiciones para varianzas mínimas de los coeficientes	Si el modelo de regresión contiene términos no mayores de primer orden, y si los rangos de las variables están dados por $x_j \in [-1, +1]$ para $j = 1, 2, \ldots, k$, entonces $Var(b_j)/\sigma^2$, para $j = 1, 2, \ldots, k$, se minimiza si el diseño es ortogonal y todos los niveles x_i del diseño son ± 1 para $i = 1, 2, \ldots, k$.

Así, en términos de los coeficientes del modelo o los efectos principales, la ortogonalidad en el 2^k es una propiedad muy deseable.

Otro enfoque para lograr una mejor comprensión del "balance" proporcionado por el 2^3 es el gráfico. En la figura 15.10 se aprecian cada uno de los contrastes ortogonales y que por esto son mutuamente independientes. Se presentan gráficas que muestran los planos de los cuadrados, cuyos vértices contienen las respuestas etiquetadas con "+" y se comparan con las que tienen "−". Las que se dan en el inciso *a*) muestran contrastes para efectos principales y deberían ser evidentes para el lector. Las del inciso *b*) presentan los planos determinados por los vértices "+" y "−" para los tres contrastes de interacción de dos factores. En el inciso *c*), se aprecia la representación geométrica de los contrastes para la interacción de tres factores (ABC).

Corridas centrales con 2^k diseños

En la situación en que se implanta el diseño 2^k con variables **continuas** de diseño, y se busca ajustar un modelo de regresión lineal, es muy útil el uso de corridas repetidas en el **diseño central**. En realidad, muy aparte de las ventajas que se verán

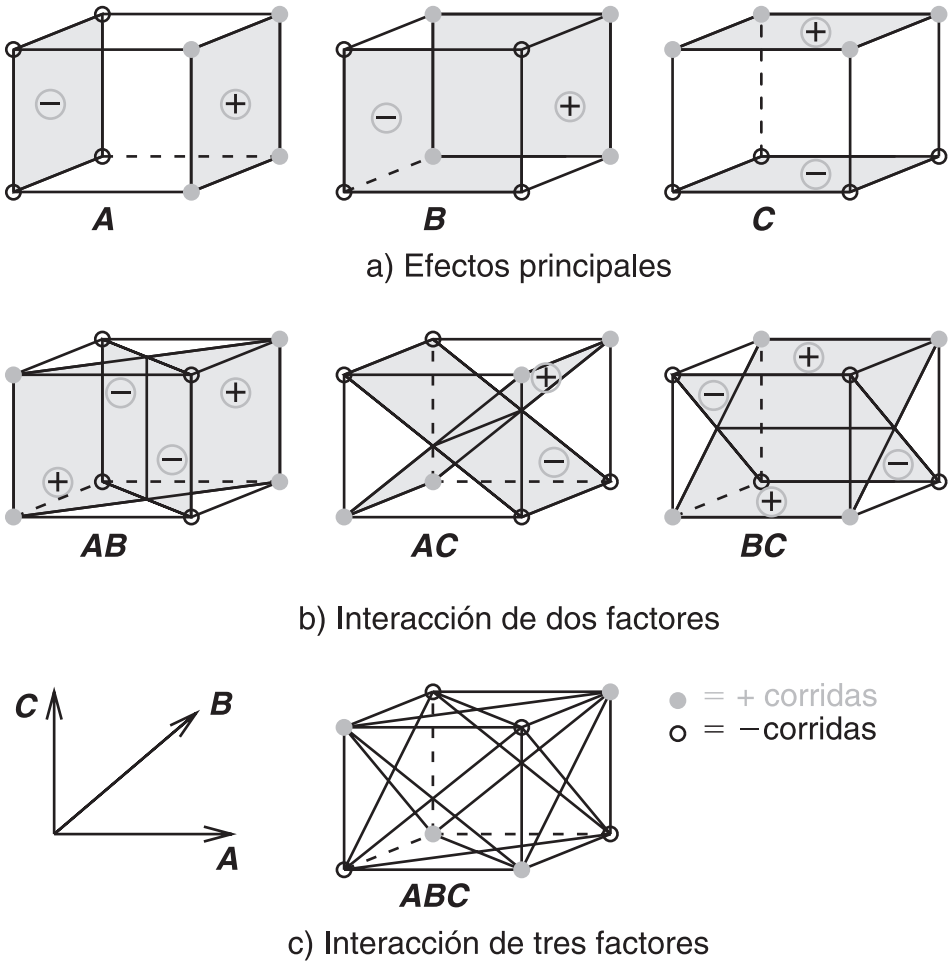


Figura 15.10: Presentación geométrica de los contrastes para el diseño factorial 2^3.

a continuación, la mayoría de los científicos e ingenieros considerarían que las corridas centrales (es decir, corridas en $x_i = 0$ para $i = 1, 2, \ldots, k$) no sólo son una práctica razonable, sino algo que tiene una atracción intuitiva. En muchas áreas de aplicación del diseño 2^k, el científico desea determinar si sería benéfico pasar a una región diferente de interés en cuanto a los factores. En muchos casos, el centro (el punto $[0, 0, \ldots, 0]$, en los factores codificados) con frecuencia representa las condiciones de operación actuales del proceso, o al menos aquellas condiciones que se consideran "óptimas en ese momento". De manera que es frecuente el caso que el científico requerirá datos en la respuesta en el centro.

Corridas centrales y falta de ajuste

Además de la atracción intuitiva del aumento del 2^k con corridas centrales, se tiene otra ventaja que se relaciona con la clase de modelo que se ajusta a los datos. Por ejemplo, considere el lector el caso con $k = 2$, según se ilustra en la figura 15.11.

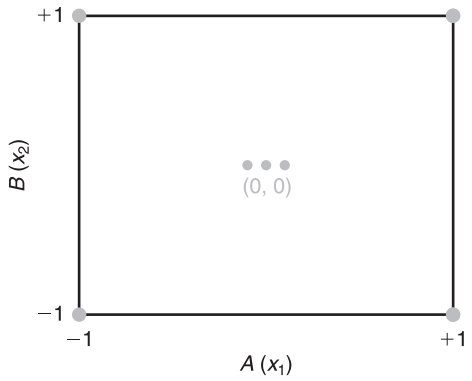


Figura 15.11: Diseño 2^2 con corridas centrales.

Queda claro que *sin las corridas centrales*, los términos del modelo son, además de la intersección, x_1, x_2, x_1x_2. Esto es importante para los cuatro grados de libertad del modelo distribuidos para los cuatro puntos del diseño, aparte de cualquier repetición. Como cada factor tiene disponible información de respuesta *sólo en dos ubicaciones* $\{-1, +1\}$, en el modelo no tienen cabida términos "puros" de curvatura de segundo orden (es decir, x_1^2 o x_2^2). Pero la información en (0, 0) produce un grado de libertad adicional del modelo. Si bien este importante grado de libertad no permite que ni x_1^2 ni x_2^2 se empleen en el modelo, sí lo permite para probar la significancia de una combinación lineal de x_1^2 y x_2^2. Para n_c corridas centrales, entonces, hay $n_c - 1$ grados de libertad disponibles para la repetición del error "puro". Esto permite un estimador de σ^2 para probar los términos del modelo y la significancia del único grado de libertad para la **falta de ajuste cuadrático**. El concepto aquí se parece mucho al material del capítulo 11, donde se estudió la falta de ajuste.

Con la finalidad de entender por completo cómo funciona la prueba de falta de ajuste, suponga que para $k = 2$ el **modelo verdadero** contiene todo el complemento de segundo orden de los términos, inclusive x_1^2 y x_2^2. En otras palabras,

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2.$$

Ahora, considere el contraste

$$\bar{y}_f - \bar{y}_0,$$

donde $\bar{y}_f$ es la respuesta promedio de las ubicaciones factoriales y $\bar{y}_0$ es la respuesta promedio en el punto central. Es fácil demostrar (véase el ejercicio de repaso 15.50) que

$$E(\bar{y}_f - \bar{y}_0) = \beta_{11} + \beta_{22},$$

y, en efecto, para el caso general con k factores,

$$E(\bar{y}_f - \bar{y}_0) = \sum_{i=1}^{k} \beta_{ii}.$$

Como resultado, la prueba de la falta de ajuste es una prueba t simple (o $F = t^2$) con

$$t_{n_c-1} = \frac{\bar{y}_f - \bar{y}_0}{s_{\bar{y}_f - \bar{y}_0}} = \frac{\bar{y}_f - \bar{y}_0}{\sqrt{MSE\,(1/n_f + 1/n_c)}},$$

donde n_c es el número de puntos factoriales y *MSE* sólo es la varianza muestral de los valores de la respuesta en $(0, 0, \ldots, 0)$.

Ejemplo 15.7: Se tomó un ejemplo de Myers y Montgomery (2002). Un ingeniero químico trata de modelar la conversión porcentual en un proceso. Hay dos variables de interés, el tiempo de reacción y la temperatura de ésta. En un intento por llegar al modelo apropiado, se realiza un experimento preliminar en un factorial 2^2 usando la región actual de interés en el tiempo de reacción y su temperatura. Se hizo una sola corrida en cada uno de los cuatro puntos factoriales y 5 en el centro del diseño, con la finalidad de que pudiera realizarse una prueba de la falta de ajuste para la curvatura. En la figura 15.12 se presenta la región del diseño y las corridas experimentales sobre el producto.

Las lecturas del tiempo y la temperatura en el centro son, por supuesto, 35 minutos y 145 °C. Los estimadores de los efectos principales y el coeficiente de interacción única se calculan mediante contrastes, igual que antes. Las corridas en el centro **no juegan ningún papel en el cálculo de b_1, b_2 y b_{12}**. La intuición del lector debería decirle que esto es razonable. Para todo el experimento, la intersección es tan sólo $\bar{y}$. Este valor es $\bar{y} = 40.4444$. Los errores estándar se encuentran usando los elementos de la diagonal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$, como ya se dijo. Para este caso,

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} & x_1 & x_2 & x_1x_2 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & -0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Después de hacer los cálculos, se tiene que

$$\begin{array}{llll} b_0 = 40.4444, & b_1 = 0.7750, & b_2 = 0.3250, & b_{12} = -0.0250, \\ s_{b_0} = 0.06231, & s_{b_1} = 0.09347, & s_{b_2} = 0.09347, & s_{b_{12}} = 0.09347, \\ t_{b_0} = 649.07 & t_{b_1} = 8.29 & t_{b_2} = 3.48 & t_{b_{12}} = 0.018, \quad (P = 0.800). \end{array}$$

El contraste $\bar{y}_f - \bar{y}_0 = 40.425 - 40.46 = -0.035$ y el estadístico t que *prueba la curvatura* está dado por

$$t = \frac{40.425 - 40.46}{\sqrt{0.0430(1/4 + 1/5)}} = 0.252, \qquad (P = 0.814).$$

Como resultado, parece como si el modelo apropiado debería contener sólo términos de primer orden (además de la intersección).

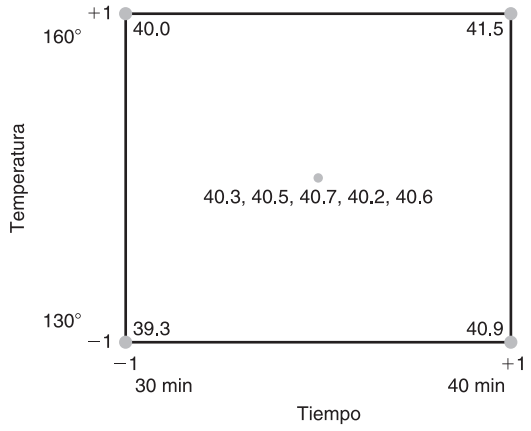


Figura 15.12: Factorial 2^2 con 5 corridas en el centro.

Una mirada intuitiva a la prueba sobre la curvatura

Si se considera el caso sencillo de una sola variable de diseño con corridas en -1 y $+1$, debe quedar claro que la respuesta promedio en -1 y $+1$ debe estar cerca de la respuesta en 0, el centro, si la naturaleza del modelo es de primer orden. Cualquier desviación sugeriría, con seguridad, curvatura. Esto es fácil de extenderse a dos variables. Considere el lector la figura 15.13.

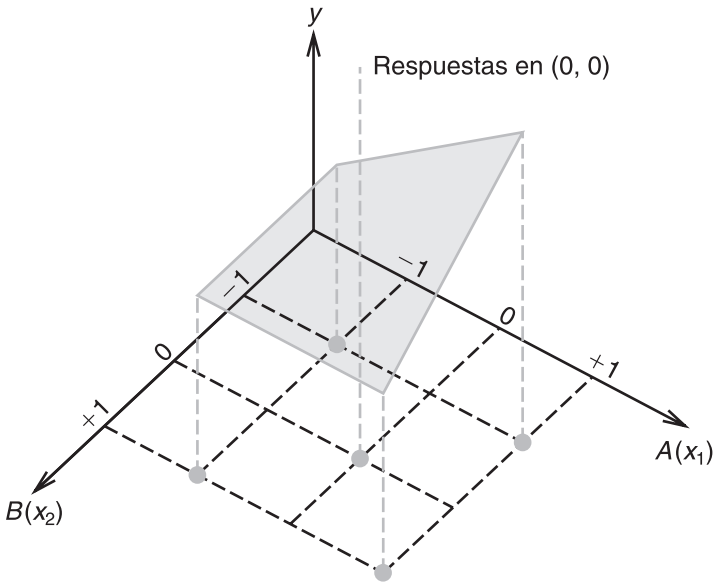


Figura 15.13: El factorial 2^2 con corridas en (0, 0).

La figura muestra el plano sobre y que pasa a través de los puntos factoriales. Éste es el plano que representaría el ajuste perfecto para el modelo que contiene x_1, x_2 y x_1x_2. Si el modelo no contiene curvatura cuadrática (es decir, $\beta_{11} = \beta_{22} = 0$, se esperaría que la respuesta en (0, 0) estuviera en el plano o cerca de éste. Si la respuesta estuviera lejos del plano, como es el caso en la figura 15.13, entonces es posible ver en forma gráfica que la curvatura cuadrática está presente.

15.7 Experimentos factoriales en bloques incompletos

El experimento factorial 2^k permite en sí mismo hacer la partición en *bloques incompletos*. Para un experimento con k factores, cuando no pueden aplicarse todas las 2^k combinaciones de tratamientos en condiciones homogéneas, con frecuencia resulta útil emplear un diseño en 2^p bloques ($p < k$). La desventaja con este planteamiento experimental es que como resultado de la formación de bloques se sacrifican por completo ciertos efectos, y la cantidad de sacrificio depende del número de bloques que se requieren. Por ejemplo, suponga que en un experimento factorial 2^3 deben correrse, en dos bloques de tamaño cuatro, las ocho combinaciones de tratamientos. Además, suponga que se desea sacrificar la interacción ABC. Note los "signos de contraste" en la tabla 15.5 de la página 617. Un arreglo razonable es el siguiente

Bloque 1	Bloque 2
(1)	a
ab	b
ac	c
bc	abc

Concepto de confusión

Si se acepta el modelo usual con el efecto aditivo de bloques, éste se cancela en la formación de los contrastes sobre todos los efectos excepto el ABC. Para ilustrarlo, se hace que x denote la contribución a la salida debida a la diferencia entre bloques. Si las salidas en el diseño se escriben como

Bloque 1	Bloque 1
(1)	$a + x$
ab	$b + x$
ac	$c + x$
bc	$abc + x$

se observa que el contraste ABC y también el contraste que compara los dos bloques, están dados por

$$\begin{aligned}\text{contraste } ABC &= (abc + x) + (c + x) + (b + x) + (a + x) - (1) - ab - ac - bc \\ &= abc + a + b + c - (1) - ab - ac - bc + 4x.\end{aligned}$$

Por lo tanto, se está midiendo el **efecto ABC más el efecto de bloques**, y no hay manera de evaluar el efecto de la interacción ABC independiente de los bloques. Entonces, se dice que la interacción ABC está **completamente confundida con**

los bloques. Por necesidad, se ha sacrificado información sobre ABC. Por otro lado, el efecto de bloque se cancela en la formación de todos los demás contrastes. Por ejemplo, el contraste A está dado por

$$\begin{aligned}\text{contraste } A &= (abc + x) + (a + x) + ab + ac - (b + x) - (c + x) - bc - (1)\\ &= abc + a + ab + ac - b - c - bc - (1),\end{aligned}$$

como en el caso de un diseño aleatorio por completo. Se dice que los efectos A, B, C, AB, AC y BC, son *ortogonales a los bloques*. Por lo general, para un experimento factorial 2^k en 2^p bloques, el número de efectos confundido con los bloques es $2^p - 1$, que es equivalente a los grados de libertad para los bloques.

Factorial 2^k en dos bloques

Cuando se van a emplear dos bloques con un experimento factorial 2^k, como **contraste definitorio** se elige un efecto, por lo general, una interacción de orden superior. Este efecto va a confundirse con los bloques. Los efectos adicionales $2^k - 2$ son ortogonales con el contraste definitorio y por ello con los bloques.

Suponga el lector que el contraste definitorio se representa como $A^{\gamma 1}B^{\gamma 2}C^{\gamma 3}\ldots$, donde γ_i toma el valor de 0 o de 1. Esto genera la expresión

$$L = \gamma_1 + \gamma_2 + \cdots \gamma_k,$$

que a la vez se evalúa para cada una de las 2^k combinaciones de tratamientos por medio de hacer γ_i igual a 0 o a 1, según si la combinación de tratamientos contiene el i-ésimo factor en sus niveles alto o bajo. Entonces, los valores L se reducen (módulo 2) ya sea a 0 o a 1, lo cual determina a cuál bloque de combinaciones de tratamientos tienen que asignarse. En otras palabras, las combinaciones de tratamientos se dividen en dos bloques, según si los valores de L dejan un residuo de 0 o 1 cuando se dividen entre 2.

Ejemplo 15.8: Determine los valores de L (módulo 2) para un experimento factorial 2^3 cuando el contraste definitorio es ABC.

Solución: Con el contraste definitorio ABC, se tiene que

$$L = \gamma_1 + \gamma_2 + \gamma_3,$$

que se aplica a cada tratamiento en la forma siguiente:

$$\begin{aligned}
(1)&: \quad L = 0 + 0 + 0 = 0 = 0 \quad (\text{módulo } 2)\\
a&: \quad L = 1 + 0 + 0 = 1 = 1 \quad (\text{módulo } 2)\\
b&: \quad L = 0 + 1 + 0 = 1 = 1 \quad (\text{módulo } 2)\\
ab&: \quad L = 1 + 1 + 0 = 2 = 0 \quad (\text{módulo } 2)\\
c&: \quad L = 0 + 0 + 1 = 1 = 1 \quad (\text{módulo } 2)\\
ac&: \quad L = 1 + 0 + 1 = 2 = 0 \quad (\text{módulo } 2)\\
bc&: \quad L = 0 + 1 + 1 = 2 = 0 \quad (\text{módulo } 2)\\
abc&: \quad L = 1 + 1 + 1 = 3 = 1 \quad (\text{módulo } 2).
\end{aligned}$$

Igual que antes, el arreglo de los bloques, en los que ABC está confundido, es

Bloque 1	Bloque 2
(1)	a
ab	b
ac	c
bc	abc

Los efectos A, B, C, AB, AC y BC, y las sumas de los cuadrados, se calculan en la forma habitual, ignorando los bloques.

Observe el lector que este arreglo es el mismo esquema de formación de bloques que resultaría al asignar las combinaciones de factores con signo "+" para el contraste ABC a uno de los bloques, y aquellas con signo "−" para el contraste ABC al otro bloque.

El bloque que contiene la combinación de tratamiento (1) en este ejemplo se denomina el **bloque principal**. Éste forma un grupo algebraico con respecto a la multiplicación cuando los exponentes se reducen al módulo de base 2. Por ejemplo, cumple la propiedad de cerradura, ya que

$$(ab)(bc) = ab^2c = ac, \qquad (ab)(ab) = a^2b^2 = (1),$$

y así sucesivamente.

Factorial 2^k en cuatro bloques

Si se requiere que el experimentador asigne las combinaciones de los tratamientos en cuatro bloques, elegirá dos contrastes definitorios. Un tercer efecto, conocido como su **interacción generalizada**, se confunde en forma automática con los bloques, estos tres efectos corresponden a los tres grados de libertad para los bloques. El procedimiento para construir el diseño se explica mejor con un ejemplo. Suponga que se decidió que para un factorial 2^4 los contrastes definitorios son AB y CD. El tercer efecto confundido, la interacción generalizada de aquellos, está formada con la multiplicación de los dos módulos 2 iniciales. Entonces, el efecto

$$(AB)(CD) = ABCD$$

también está confundido con los bloques. El diseño se construye calculando las expresiones

$$\begin{aligned} L_1 &= \gamma_1 + \gamma_2 \qquad (AB), \\ L_2 &= \gamma_3 + \gamma_4 \qquad (CD) \end{aligned}$$

módulo 2, para cada una de las 16 combinaciones de tratamiento, para generar el siguiente esquema de bloques:

Bloque 1	Bloque 2	Bloque 3	Bloque 4
(1)	a	c	ac
ab	b	abc	bc
cd	acd	d	ad
$abcd$	bcd	abd	bd
$L_1 = 0$	$L_1 = 1$	$L_1 = 0$	$L_1 = 1$
$L_2 = 0$	$L_2 = 0$	$L_2 = 1$	$L_2 = 1$

Para construir los bloques restantes después de haber generado el bloque principal, se usa un atajo del procedimiento. Se comienza por colocar cualquier combinación de tratamientos no en el bloque principal, sino en un segundo bloque, y se construye el bloque con la multiplicación (módulo 2) por las combinaciones de tratamientos en el bloque principal. En el ejemplo anterior, los bloques segundo, tercero y cuarto se generaron como sigue:

Bloque 2	Bloque 3	Bloque 4
$a(1) = a$	$c(1) = c$	$ac(1) = ac$
$a(ab) = b$	$c(ab) = abc$	$ac(ab) = bc$
$a(cd) = acd$	$c(cd) = d$	$ac(cd) = ad$
$a(abcd) = bcd$	$c(abcd) = abd$	$ac(abcd) = bd$

El análisis para el caso de cuatro bloques es muy sencillo. Se calculan en la forma habitual todos los efectos que son ortogonales a los bloques (aquellos que definen los contrastes).

Factorial 2^k en 2^p bloques

El esquema general para el experimento factorial 2^k en 2^p no es difícil. Se selecciona p definiendo contrastes tales que ninguno sea la interacción generalizada de cualesquiera dos en el grupo. Como hay $2^p - 1$ grados de libertad para los bloques, se tienen $2^p - 1 - p$ efectos adicionales confundidos con los bloques. Por ejemplo, en un experimento factorial 2^6 en ocho bloques, se eligen ACF, $BCDE$ y $ABDF$ como los contrastes definitorios. Entonces,

$$\begin{aligned}(ACF)(BCDE) &= ABDEF,\\ (ACF)(ABDF) &= BCD,\\ (BCDE)(ABDF) &= ACEF,\\ (ACF)(BCDE)(ABDF) &= E\end{aligned}$$

son los cuatro efectos adicionales confundidos con los bloques. Éste no es un esquema deseable para la formación de bloques, ya que uno de los efectos confundidos es el E principal. El diseño se construye con la evaluación de

$$\begin{aligned}L_1 &= \gamma_1 + \gamma_3 + \gamma_6,\\ L_2 &= \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5,\\ L_3 &= \gamma_1 + \gamma_2 + \gamma_4 + \gamma_6\end{aligned}$$

y la asignación de los tratamientos en combinaciones a los bloques, de acuerdo con el esquema siguiente:

$$\begin{aligned}&\text{Bloque 1:} \quad L_1 = 0, \quad L_2 = 0, \quad L_3 = 0\\ &\text{Bloque 2:} \quad L_1 = 0, \quad L_2 = 0, \quad L_3 = 1\\ &\text{Bloque 3:} \quad L_1 = 0, \quad L_2 = 1, \quad L_3 = 0\\ &\text{Bloque 4:} \quad L_1 = 0, \quad L_2 = 1, \quad L_3 = 1\\ &\text{Bloque 5:} \quad L_1 = 1, \quad L_2 = 0, \quad L_3 = 0\\ &\text{Bloque 6:} \quad L_1 = 1, \quad L_2 = 0, \quad L_3 = 1\end{aligned}$$

Bloque 7: $L_1 = 1, \quad L_2 = 1, \quad L_3 = 0$
Bloque 8: $L_1 = 1, \quad L_2 = 1, \quad L_3 = 1.$

El atajo del procedimiento que se ilustró para el caso de cuatro bloques, también se aplica aquí. Por lo tanto, los siete bloques restantes se construyen a partir del bloque principal.

Ejemplo 15.9: Es de interés estudiar el efecto de los cinco factores sobre alguna respuesta con la suposición de que son despreciables las interacciones que implican tres, cuatro y cinco de los factores. Deben dividirse las 32 combinaciones de tratamiento en cuatro bloques, usando contrastes definitorios $BCDE$ y $ABCD$. Así,

$$(BCDE)(ABCD) = AE$$

también se confunde con los bloques. En la tabla 15.9 se da el diseño experimental y las observaciones.

Tabla 15.9: Datos para un experimento 2^5 en cuatro bloques

Bloque 1	Bloque 2	Bloque 3	Bloque 4
$(1) = 30.6$	$a = 32.4$	$b = 32.6$	$e = 30.7$
$bc = 31.5$	$abc = 32.4$	$c = 31.9$	$bce = 31.7$
$bd = 32.4$	$abd = 32.1$	$d = 33.3$	$bde = 32.2$
$cd = 31.5$	$acd = 35.3$	$bcd = 33.0$	$cde = 31.8$
$abe = 32.8$	$be = 31.5$	$ae = 32.0$	$ab = 32.0$
$ace = 32.1$	$ce = 32.7$	$abce = 33.1$	$ac = 33.1$
$ade = 32.4$	$de = 33.4$	$abde = 32.9$	$ad = 32.2$
$abcde = 31.8$	$bcde = 32.9$	$acde = 35.0$	$abcd = 32.3$

La asignación de combinaciones de tratamientos a unidades experimentales en los bloques es, por supuesto, aleatoria. Al agrupar las tres, cuatro y cinco interacciones de factores no confundidas, para formar el término del error, realice el análisis de varianza para los datos de la tabla 15.9.

Solución: Se calculan las sumas de los cuadrados de cada uno de los 31 contrastes, y se encuentra que la suma de los cuadrados de los bloques es

$$SS(\text{bloques}) = SS(ABCD) + SS(BCDE) + SS(AE) = 7.538.$$

El análisis de varianza se presenta en la tabla 15.10. Ninguna de las interacciones de dos factores es significativa con un nivel $\alpha = 0.05$, cuando se comparan con $f_{0.05}(1.14) = 4.60$. Los efectos principales A y D son significativos y ambos dan efectos positivos sobre la respuesta, conforme se pasa del nivel bajo al alto.

Confusión parcial

Con los métodos descritos en la sección 15.7, es posible confundir cualquier efecto con los bloques. Suponga el lector que se considera un experimento factorial 2^3 en dos bloques con tres repeticiones completas. Si ABC está confundida con los bloques

Tabla 15.10: Análisis de varianza para los datos de la tabla 15.9

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Media cuadrática**	***f* calculada**
Efecto principal:				
A	3.251	1	3.251	6.32
B	0.320	1	0.320	0.62
C	1.361	1	1.361	2.64
D	4.061	1	4.061	7.89
E	0.005	1	0.005	0.01
Interacción de dos factores:				
AB	1.531	1	1.531	2.97
AC	1.125	1	1.125	2.18
AD	0.320	1	0.320	0.62
BC	1.201	1	1.201	2.33
BD	1.711	1	1.711	3.32
BE	0.020	1	0.020	0.04
CD	0.045	1	0.045	0.09
CE	0.001	1	0.001	0.002
DE	0.001	1	0.001	0.002
Bloques $(ABCD, BCDE, AE)$:	7.538	3	2.513	
Error	7.208	14	0.515	

en las tres réplicas, se procede igual que antes y se determinan sumas de cuadrados de un solo grado de libertad para todos los efectos principales y los efectos de interacción de dos factores. La suma de cuadrados para los bloques tiene cinco grados de libertad, lo que deja $23 - 5 - 6 = 12$ grados de libertad para el error.

Ahora, se confundirá ABC en una réplica, AC en la segunda y BC en la tercera. El plan para este tipo de experimento sería el siguiente:

Réplica 1 (ABC Confundida)		Réplica 2 (AC Confundida)		Réplica 3 (BC Confundida)	
Bloque 1	Bloque 2	Bloque 1	Bloque 2	Bloque 1	Bloque 2
abc	ab	abc	ab	abc	ab
a	ac	ac	bc	bc	ac
b	bc	b	a	a	b
c	(1)	(1)	c	(1)	c

Se dice que los efectos ABC, AC y BC están **parcialmente confundidos con los bloques**. Estos tres factores se estiman a partir de dos de las tres repeticiones. La razón 2/3 sirve como medida del grado de confusión. Esta razón da la cantidad de información disponible sobre el efecto parcialmente confundido con respecto de la disponible sobre el efecto no confundido.

En la tabla 15.11 se presenta el análisis de varianza. Las sumas de los cuadrados para los bloques y para los efectos no confundidos A, B, C y AB, se encuentran en la forma habitual. Las sumas de cuadrados para AC, BC y ABC se calculan a partir de las dos repeticiones en las que el efecto particular no está confundido. Cuando se

obtengan las sumas de cuadrados para los efectos parcialmente confundidos, debe tenerse cuidado en dividir entre 16 en vez de entre 24, ya que sólo se usan 16 observaciones. En la tabla 15.11 los primeros están insertados con los grados de libertad, como recordatorio de que estos efectos están confundidos parcialmente y requieren cálculos especiales.

Tabla 15.11: Análisis de varianza con confusión parcial

Fuente de variación	**Grados de libertad**
Bloques	5
A	1
B	1
C	1
AB	1
AC	$1'$
BC	$1'$
ABC	$1'$
Error	11
Total	23

Ejercicios

15.13 Presente el arreglo de bloques para un experimento factorial 2^3 con tres repeticiones y, mediante una tabla de análisis de varianza, indique los efectos a probar y sus grados de libertad, cuando la interacción AB esté confundida con los bloques.

15.14 Se realizó el siguiente experimento para estudiar los efectos principales y todas las interacciones. Se emplearon cuatro factores con cuatro niveles cada uno. El experimento se repitió y fueron necesarios dos bloques en cada réplica. Los datos se presentan a continuación.

a) ¿Cuál efecto está confundido con los bloques en la primera repetición del experimento? ¿Y en la segunda?

b) Efectúe un análisis de varianza adecuado con pruebas sobre todos los efectos principales y los de las interacciones. Utilice un nivel de significancia de 0.05.

Réplica 1		**Réplica 2**	
Bloque 1	**Bloque 2**	**Bloque 3**	**Bloque 4**
$(1) = 17.1$	$a = 15.5$	$(1) = 18.7$	$a = 17.0$
$d = 16.8$	$b = 14.8$	$ab = 18.6$	$b = 17.1$
$ab = 16.4$	$c = 16.2$	$ac = 18.5$	$c = 17.2$
$ac = 17.2$	$ad = 17.2$	$ad = 18.7$	$d = 17.6$
$bc = 16.8$	$bd = 18.3$	$bc = 18.9$	$abc = 17.5$
$abd = 18.1$	$cd = 17.3$	$bd = 17.0$	$abd = 18.3$
$acd = 19.1$	$abc = 17.7$	$cd = 18.7$	$acd = 18.4$
$bcd = 18.4$	$abcd = 19.2$	$abcd = 19.8$	$bcd = 18.3$

15.15 Divida las combinaciones de tratamientos de un experimento factorial 2^4 en cuatro bloques, confundiendo ABC y ABD. ¿Cuál es el efecto adicional que también está confundido con los bloques?

15.16 Se realiza un experimento para determinar la fuerza de frenado de cierta aleación que contiene cinco metales, A, B, C, D y E. Se emplean dos porcentajes diferentes de cada metal para formar las $2^5 = 32$ aleaciones distintas. Como sólo es posible probar ocho aleaciones en un día dado, el experimento se lleva a cabo durante un periodo de cuatro días, durante los cuales se confunden con éstos los efectos $ABDE$ y AE. Los datos experimentales se presentan enseguida.

a) Plantee el esquema de bloques para los 4 días.

b) ¿Qué efecto adicional se confunde con los días?

c) Obtenga las sumas de los cuadrados para todos los efectos principales.

Comb. de trat.	**Fuerza de frenado**	**Comb. de trat.**	**Fuerza de frenado**
(1)	21.4	e	29.5
a	32.5	ae	31.3
b	28.1	be	33.0
ab	25.7	abe	23.7
c	34.2	ce	26.1
ac	34.0	ace	25.9

Comb. de trat.	Fuerza de frenado	Comb. de trat.	Fuerza de frenado
bc	23.5	*bce*	35.2
abc	24.7	*abce*	30.4
d	32.6	*de*	28.5
ad	29.0	*ade*	36.2
bd	30.1	*bde*	24.7
abd	27.3	*abde*	29.0
cd	22.0	*cde*	31.3
acd	35.8	*acde*	34.7
bcd	26.8	*bcde*	26.8
abcd	36.4	*abcde*	23.7

15.17 Al confundir ABC en dos réplicas y AB en una tercera, muestre el arreglo de bloques y la tabla de análisis de varianza para un experimento factorial 2^3 con tres repeticiones. ¿Cuál es la información relativa sobre los efectos confundidos?

15.18 Los datos codificados que siguen representan la resistencia de cierto tipo de lote de envoltura para pan, que procede de 16 condiciones diferentes, las cuales representan dos niveles de cada una de cuatro variables de proceso. En el modelo se introdujo un efecto operador, ya que era necesario obtener la mitad de corridas experimentales con el operador 1, y la otra mitad con el 2. Se pensaba que los operadores tendrían un efecto en la calidad del producto.

a) Si se supone que las interacciones son despreciables, haga pruebas de significancia para los factores A, B, C y D. Use un nivel de significancia de 0.05.

b) ¿Qué interacción está confundida con los operadores?

Operador 1	Operador 2
$(1) = 18.8$	$a = 14.7$
$ab = 16.5$	$b = 15.1$
$ac = 17.8$	$c = 14.7$
$bc = 17.3$	$abc = 19.0$
$d = 13.5$	$ad = 16.9$
$abd = 17.6$	$bd = 17.5$
$acd = 18.5$	$cd = 18.2$
$bcd = 17.6$	$abcd = 20.1$

15.19 Considere un experimento 2^5 donde las corridas experimentales son sobre 4 máquinas diferentes. Use las máquinas como bloques y suponga que todos los efectos principales y las interacciones de dos factores son importantes.

a) ¿Cuáles corridas se harían sobre cada una de las 4 máquinas?

b) ¿Cuáles efectos se confunden con los bloques?

15.20 En un experimento descrito en Myers y Montgomery (2002), se plantea que se buscan las condiciones óptimas para almacenar semen de bovinos, con la finalidad de obtener la supervivencia máxima. Las variables son el porcentaje de citrato de sodio, el porcentaje de glicerol y el tiempo de equilibrio en horas. La respuesta es el porcentaje de supervivencia de los espermatozoides móviles. Los niveles naturales se encuentran en la referencia mencionada. A continuación se presentan los datos con los niveles codificados para la porción factorial del diseño y las corridas centrales.

x_1, Porcentaje de citrato de sodio	x_2, Porcentaje de glicerol	x_3, Tiempo de equilibrio	Porcentaje de supervivencia
−1	−1	−1	57
1	−1	−1	40
−1	1	1	19
1	1	1	40
−1	−1	−1	54
1	−1	−1	41
−1	1	1	21
1	1	1	43
0	0	0	63
0	0	0	61

a) Ajuste un modelo de regresión lineal a los datos y determine cuáles términos lineales y de interacción son significativos. Suponga que la interacción $x_1x_2x_3$ es despreciable.

b) Pruebe la falta de ajuste para el modelo cuadrático y comente la respuesta.

15.21 Los productores de petróleo están interesados en aleaciones de níquel de alta resistencia contra la corrosión. Se realizó un experimento en el que se compararon a la tensión especímenes de aleaciones de níquel, cargados en una solución de ácido sulfúrico saturada con disulfuro de carbón. Se combinaron dos aleaciones; una con 75% de níquel y otra con 35% de este metal. Se probaron las aleaciones con dos tiempos de carga distintos, 25 y 50 días. Se realizó un factorial 2^3 con los factores siguientes:

% de ácido sulfúrico al 4%, 6%: (x_1)
tiempo de carga, 25 días, 50 días: (x_2)
composición de níquel, 30%, 75%: (x_3)

Se preparó un espécimen para cada una de las ocho condiciones. Como los ingenieros no estaban seguros de la naturaleza del modelo (es decir, si se necesitarían o no términos cuadráticos), se incorporó un tercer nivel (medio) y se emplearon cuatro corridas centrales con el empleo de cuatro especímenes con ácido sulfúrico al 5%, 37.5 días y 52.5% de níquel. Las siguientes son las resistencias en kilogramos por pulgada cuadrada.

	Tiempo de carga			
	25 días		50 días	
Comp. de níquel	Ácido sulfúrico		Ácido sulfúrico	
	4%	6%	4%	6%
75%	52.5	56.5	47.9	47.2
30%	50.2	50.8	47.4	41.7

Las corridas centrales dan las resistencias siguientes:

51.6, 51.4, 52.4, 52.9

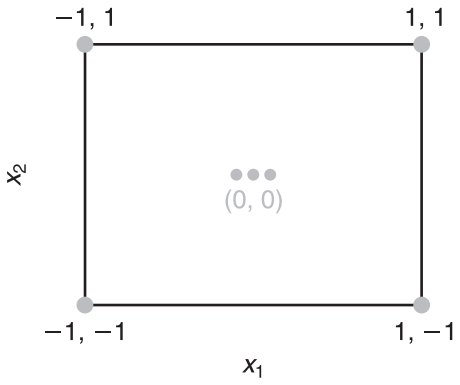


Figura 15.14: Gráfica para el ejercicio 15.23.

a) Haga pruebas para determinar cuáles efectos e interacciones deberían incluirse en el modelo ajustado.

b) Pruebe para la curvatura cuadrática.

c) Si la curvatura cuadrática es significativa, ¿cuántos puntos de diseño adicionales se necesitan para determinar cuáles términos cuadráticos deberían incluirse en el modelo?

15.22 Suponga el lector que podría efectuarse una segunda repetición del experimento del ejercicio 15.19.

a) ¿La mejor selección sería una segunda repetición del esquema de bloques del ejercicio 15.19?

b) Si la respuesta del inciso *a*) es *no*, proporcione el bosquejo para una selección mejor para la segunda réplica.

c) ¿Qué concepto utilizó en su selección del diseño?

15.23 Considere la figura 15.14, la cual representa un factorial 2^2 con 3 corridas centrales. Si la cuadratura cuadrática es significativa, ¿cuáles puntos de diseño adicionales seleccionaría el lector que permitieran estimar los términos x_1^2 y x_2^2. Explique su respuesta.

15.8 Experimentos factoriales fraccionarios

Cuando el valor de k es grande, el experimento factorial 2^k puede hacerse muy demandante, en términos del número de unidades experimentales que se requiere. Una de las ventajas reales con este plan experimental es que permite un grado de libertad para cada interacción. Sin embargo, en muchas situaciones experimentales se sabe que ciertas interacciones son despreciables, por lo que sería un desperdicio de esfuerzo experimental el usar el experimento factorial completo. En realidad, el experimentador podría tener una restricción económica que no permitiera tomar observaciones en todas las 2^k combinaciones de tratamientos. Cuando k es grande, es frecuente usar un **experimento factorial fraccionario** donde en realidad se considere la mitad, un cuarto o incluso un octavo del total de planes factoriales.

Construcción de la fracción $\frac{1}{2}$

La construcción del diseño de media repetición es idéntico a la asignación del experimento factorial 2^k en dos bloques. Se comienza por seleccionar un contraste definitorio que se vaya a sacrificar por completo. Después, se construyen los dos bloques en concordancia y se elige cualquiera de ellos como plan experimental.

Es frecuente que se haga referencia a la fracción $\frac{1}{2}$ de un factorial 2^k como diseño 2^{k-1}, el cual indica el número de puntos de diseño. La primera ilustración de un diseño 2^{k-1} es uno de $\frac{1}{2}$ o uno de 2^{3-1}. En otras palabras, el científico o el ingeniero no puede usar el complemento completo (es decir, el total de 2^3 con 8 puntos de diseño), por lo que debe apelar a un diseño con sólo 4 puntos de diseño. La pregunta es la siguiente: de los puntos de diseño (1), a, b, ab, ac, c, bc y abc, ¿cuáles son los cuatro puntos de diseño que resultarían en el diseño más útil? La respuesta, junto con los conceptos importantes relacionados, aparece en la tabla de signos + y − que muestran los contrastes para todos los 2^3. Considere la tabla 15.12.

Tabla 15.12: Contrastes para los siete efectos disponibles para un experimento factorial 2^3

	Combinación de tratamientos	**Efectos**							
		I	***A***	***B***	***C***	***AB***	***AC***	***BC***	***ABC***
2^{3-1}	a	+	+	−	−	−	−	+	+
	b	+	−	+	−	−	+	−	+
	c	+	−	−	+	+	−	−	+
	abc	+	+	+	+	+	+	+	+
2^{3-1}	ab	+	+	+	−	+	−	−	−
	ac	+	+	−	+	−	+	−	−
	bc	+	−	+	+	−	−	+	−
	(1)	+	−	−	−	+	+	+	−

Observe que las dos fracciones $\frac{1}{2}$ son $\{a, b, c, abc\}$ y $\{ab, ac, bc, (1)\}$. También note en la tabla 15.12, que en ambos diseños ABC no tiene contraste, pero todos los demás efectos sí lo tienen. En una de las fracciones se tiene que ABC contiene todos los signos + y en la otra fracción el efecto ABC todos los signos −. Como resultado, se dice que el diseño de la parte superior de la tabla está descrito por $\boldsymbol{ABC = I}$, y el de la parte inferior por $\boldsymbol{ABC = -I}$. La interacción $\boldsymbol{ABC}$ se denomina **generador del diseño,** y $ABC = I$ (o $ABC = -I$, para el segundo diseño) recibe el nombre de **relación definitoria**.

Alias en el 2^{3-1}

Si nos centramos en el diseño $ABC = I$ (el 2^{3-1} superior), es evidente que seis efectos contienen contrastes. Esto produce la apariencia inicial de que todos los *efectos* se estudian por separado de ABC. Sin embargo, es seguro que el lector recuerda que con sólo cuatro puntos de diseño, aun si se repiten, los grados de libertad disponibles (además de aquel para el error experimental) son

Términos del modelo de regresión	3
Intersección	1
	4

Un análisis más de cerca sugiere que siete efectos no son ortogonales y, en realidad, cada contraste está representado en otro efecto. De hecho, si se emplea el símbolo ≡ para denotar **contrastes idénticos**, se tiene que

$$A \equiv BC; \qquad B \equiv AC; \qquad C \equiv AB.$$

Como resultado, dentro de un par, un efecto no puede estimarse en forma independiente de su "socio" alias. De hecho, los efectos

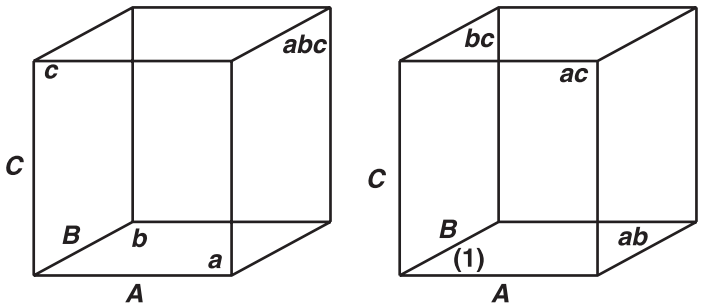
$$A = \frac{a + abc - b - c}{2}, \qquad \text{y} \qquad BC = \frac{a + abc - b - c}{2}$$

producirán el mismo resultado numérico, por lo que contienen la misma información. En realidad, es frecuente decir que **comparten un grado de libertad**. El efecto estimado, realmente estima la suma, es decir, $A + BC$. Se dice que A y BC son alias, al igual que B y AC, y que C y AB.

Para la fracción $ABC = -I$, se observa que los alias son los mismos que para la fracción $ABC = I$, signo aparte. Así, se tiene

$$A \equiv -BC; \qquad B \equiv -AC; \qquad C \equiv -AB.$$

Las dos fracciones aparecen en las esquinas del cubo de la figura 15.15a) y 15.15b).



a) $ABC =$ fracción I b) $ABC =$ fracción -1

Figura 15.15: Fracciones $\frac{1}{2}$ del factorial 2^3.

Forma en que se determinan los alias en general

En general, para un 2^{k-1}, cada efecto, además de aquel definido por el generador, tendrá un *solo socio alias*. El efecto definido por el generador no tendrá alias en otro efecto pero el suyo será la media, ya que el estimador de mínimos cuadrados será la media. Para determinar el alias de cada efecto, sólo se comienza con la relación definitoria, digamos $ABC = I$ para el 2^{3-1}. Después de hallar, digamos, el alias para el efecto A, se multiplica A por ambos lados de la ecuación $ABC = I$ y se reduce cualquier exponente por el módulo 2. Por ejemplo,

$$A \cdot ABC = A, \text{ con lo que } BC \equiv A.$$

En forma similar,

$$B \equiv B \cdot ABC \equiv AB^2C \equiv AC,$$

y, por supuesto,

$$C \equiv C \cdot ABC \equiv ABC^2 \equiv AB.$$

Ahora, para la segunda fracción (es decir, definida por la relación $ABC = -I$),

$$A \equiv -BC; \qquad B \equiv -AC; \qquad C \equiv -AB.$$

Como resultado, el valor numérico del efecto A en realidad estima a $A - BC$. Asimismo, el valor de B estima $B - AC$, y el valor de C estima a $C - AB$.

Construcción formal de 2^{k-1}

La comprensión clara del concepto de los alias hace muy sencillo entender la construcción de 2^{k-1}. Se comienza con la investigación de 2^{3-1}. Se requieren tres factores y cuatro puntos de diseño. El procedimiento comienza con un **factorial completo** en $k - 1 = 2$ factores A y B. Después se agrega un tercer factor de acuerdo con las estructuras de alias deseadas. Por ejemplo, con ABC como el generador, resulta claro que $C = \pm AB$. Así, se encuentra $C = AB$ o $-AB$ para que proporcione el factorial completo en A y B. La tabla 15.13 ilustra lo que es un procedimiento muy sencillo.

Tabla 15.13: Construcción de los dos diseños 2^{3-1}

2^2 **Básico**		2^{3-1}; $\boldsymbol{ABC = I}$			2^{3-1}; $\boldsymbol{ABC = -I}$		
$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{C = AB}$	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{C = -AB}$
−	−	−	−	+	−	−	−
+	−	+	−	−	+	−	+
−	+	−	+	−	−	+	+
+	+	+	+	+	+	+	−

Ya vimos que $ABC = I$ brinda los puntos de diseño (1), a, b, abc; en tanto que $ABC = -I$ da (1), ac, bc y ab. Desde antes era posible construir los mismos diseños usando los contrastes que se muestran en la tabla 15.12. Sin embargo, conforme el diseño se vuelve más complicado con fracciones superiores, esas tablas de contrastes se vuelven más difíciles de trabajar.

Ahora considere un 2^{4-1}, (es decir, $\frac{1}{2}$ de un diseño factorial 2^4) donde intervienen los factores A, B, C y D. Como en el caso del 2^{3-1}, la interacción de mayor orden, $ABCD$, es la que se usa como generador. Debe recordarse que $ABCD = I$, la relación definitoria sugiere que la información sobre $ABCD$ resulta sacrificada. Aquí se comienza con 2^3 completo en A, B y C, y se forma $D = \pm ABC$ para generar los dos diseños 2^{4-1}. La tabla 15.14 ilustra la construcción de ambos diseños.

Aquí, empleando las notaciones de a, b, c, etcétera, se tienen los diseños siguientes:

$$ABCD = I,\ (1),\ ad,\ bd,\ ab,\ cd,\ ac,\ bc,\ abcd$$
$$ABCD = -I,\ d,\ a,\ b,\ abc,\ c,\ acd,\ bcd,\ abc.$$

En el caso de 2^{4-1} se encuentran los alias como ya se ilustró para 2^{3-1}. Cada efecto tiene un solo socio alias que se encuentra con la multiplicación y usando la

Tabla 15.14: Construcción de los dos diseños 2^{4-1}

2^3 **Básico**			2^{4-1}; $ABCD = I$				2^{4-1}; $ABCD = -I$			
A	B	C	A	B	C	$D = ABC$	A	B	C	$D = -ABC$
−	−	−	−	−	−	−	−	−	−	+
+	−	−	+	−	−	+	+	−	−	−
−	+	−	−	+	−	+	−	+	−	−
+	+	−	+	+	−	−	+	+	−	+
−	−	+	−	−	+	+	−	−	+	−
+	−	+	+	−	+	−	+	−	+	+
−	+	+	−	+	+	−	−	+	+	+
+	+	+	+	+	+	+	+	+	+	−

relación definitoria. Por ejemplo, el alias de A para el diseño $ABCD = I$ está dado por

$$A = A \cdot ABCD = A^2BCD = BCD.$$

El alias para AB está dado por

$$AB = AB \cdot ABCD = A^2B^2CD = CD.$$

Como es fácil observar, los efectos principales tienen alias con tres interacciones de factores, y dos interacciones de factores los tienen con otras dos interacciones. La lista completa es la que sigue:

$$\begin{aligned} A &= BCD \qquad & AB &= CD \\ B &= ACD \qquad & AC &= BD \\ C &= ABD \qquad & AD &= BC \\ D &= ABC. \end{aligned}$$

Construcción de la fracción $\frac{1}{4}$

En el caso de la fracción $\frac{1}{4}$, en vez de una se seleccionan dos interacciones para ser sacrificadas, y la tercera resulta al encontrar la interacción generalizada de las dos seleccionadas. Observe que esto se parece mucho a la construcción de cuatro bloques que se estudió en la sección 15.7. La fracción que se emplea tan sólo es uno de los bloques. Un ejemplo sencillo ayuda mucho para ver la conexión con la construcción de la fracción $\frac{1}{2}$. Considere el lector la construcción de $\frac{1}{4}$ de un factorial 2^5 (es decir, 2^{5-2}), con los factores A, B, C, D y E. Un procedimiento que **evita la confusión de dos efectos principales** es la selección de ABD y ACE como las interacciones que corresponden a los dos generadores, lo que da $ABD = I$ y $ACE = I$ como las relaciones definitorias. La tercera interacción sacrificada sería $(ABD)(ACE) = A^2BCDE = BCDE$. Para la construcción del diseño se comienza con el factorial $2^{5-2} = 2^3$ en A, B y C. Se usan las interacciones ABD y ACE para suministrar los generadores, de manera que el factorial 2^3 en A, B y C es suministrado por el factor $D = \pm\, AB$ y $E = \pm AC$.

Así, una de las fracciones está dada por

$$\begin{array}{ccccc|l} A & B & C & D=AB & E=AC & \\ \hline - & - & - & + & + & de \\ + & - & - & - & - & a \\ - & + & - & - & + & be \\ + & + & - & + & - & abd \\ - & - & + & + & - & cd \\ + & - & + & - & + & ace \\ - & + & + & - & - & bc \\ + & + & + & + & + & abcde \end{array}$$

Las otras tres fracciones se encuentran utilizando el generador $\{D=-AB, E=AC\}$, $\{D=AB, E=-AC\}$ y $\{D=-AB, E=-AC\}$. Considere un análisis del diseño 2^{5-2} anterior. Contiene 8 puntos de diseño para estudiar cinco factores. Los alias para los efectos principales están dados por

$$\begin{array}{lll} A(ABD)\equiv BD; & A(ACE)\equiv CE, & A(BCDE)\equiv ABCDE \\ \quad B\equiv AD & \quad\equiv ABCE & \quad\equiv CDE \\ \quad C\equiv ABCD & \quad\equiv AE & \quad\equiv BDE \\ \quad D\equiv AB & \quad\equiv ACDE & \quad\equiv BCE \\ \quad E\equiv ABDE & \quad\equiv AC & \quad\equiv BCD \end{array}$$

Los alias para otros efectos se pueden encontrar de la misma manera. La clasificación de los grados de libertad está dada por (repetición aparte)

Efectos principales	5	
Falta de ajuste	2	$(CD=BE,\ BC=DE)$
Total	7	

Se enlistan las interacciones solo para el grado dos en la falta de ajuste.

Ahora, considere el lector el caso de 2^{6-2}, lo que permite 16 puntos de diseño para estudiar seis factores. Otra vez se eligen dos generadores de diseño. Una selección pragmática para obtener un factorial completo $2^{6-2}= 2^4$ en A, B, C y D, consiste en usar $E=\pm ABC$ y $F=\pm BCD$. La construcción se da en la tabla 15.15.

Es evidente que con más de 8 puntos de diseño que 2^{5-2}, los alias de los efectos principales no representarán un problema difícil. En realidad, observe que con las relaciones definitorias $ABCE=\pm I$, $BCDF=\pm I$, y $(ABCE)(BCDF)=\text{ADEF}=\pm I$, los efectos principales tendrán alias con las interacciones que no son más complejas que las de tercer orden. La estructura de los alias para los efectos principales es:

$$\begin{array}{ll} A\equiv BCE\equiv ABCDF\equiv DEF, & D\equiv ABCDE\equiv BCF\equiv AEF, \\ B\equiv ACE\equiv CDF\equiv ABDEF, & E\equiv ABC\equiv BCDEF\equiv ADF, \\ C\equiv ABE\equiv BDF\equiv ACDEF, & F\equiv ABCEF\equiv BCD\equiv ADE, \end{array}$$

Tabla 15.15: Diseño 2^{6-2}

A	B	C	D	$E = ABC$	$F = BCD$	Combinación de tratamientos
−	−	−	−	−	−	(1)
+	−	−	−	+	−	ae
−	+	−	−	+	+	bef
+	+	−	−	−	+	abf
−	−	+	−	+	+	cef
+	−	+	−	−	+	acf
−	+	+	−	−	−	bc
+	+	+	−	+	−	$abce$
−	−	−	+	−	+	df
+	−	−	+	+	+	$adef$
−	+	−	+	+	−	bde
+	+	−	+	−	−	abd
−	−	+	+	+	−	cde
+	−	+	+	−	−	acd
−	+	+	+	−	+	$bcdf$
+	+	+	+	+	+	$abcdef$

cada uno con un solo grado de libertad. Para las interacciones de dos factores,

$$\begin{aligned} AB &\equiv CE \equiv ACDF \equiv BDEF, \qquad AF \equiv BCEF \equiv ABCD \equiv DE, \\ AC &\equiv BE \equiv ABDF \equiv CDEF, \qquad BD \equiv ACDE \equiv CF \equiv ABEF, \\ AD &\equiv BCDE \equiv ABCF \equiv EF, \qquad BF \equiv ACEF \equiv CD \equiv ABDE, \\ AE &\equiv BC \equiv ABCDEF \equiv DF. \end{aligned}$$

Por supuesto, aquí hay algunos alias entre las interacciones de dos factores. Los dos grados de libertad restantes son tomados en cuenta por los grupos siguientes:

$$ABD \equiv CDE \equiv ACF \equiv BEF, \qquad ACD \equiv BDE \equiv ABF \equiv CEF.$$

Es evidente que siempre se debe estar alerta de que la estructura de alias es para un experimento fraccionario, antes de que se recomiende, finalmente, el plan experimental. Es importante la selección adecuada de contrastes definitorios, ya que es lo que dicta la estructura de los alias.

15.9 Análisis de los experimentos factoriales fraccionarios

La dificultad de hacer pruebas formales de significancia con datos de experimentos factoriales fraccionarios estriba en la determinación del término del error apropiado. A menos que haya datos disponibles de experimentos anteriores, el error debe venir de un conjunto de contrastes que representan efectos que se presume son despreciables.

Las sumas de cuadrados para los efectos individuales se encuentran usando en esencia los mismos procedimientos dados para el factorial completo. Es posible formar un contraste en las combinaciones de tratamientos con la construcción de la tabla de signos positivos y negativos. Por ejemplo, para la mitad de réplica de un

experimento factorial 2^3, con ABC como contraste definitorio, en la tabla 15.16 se presentan un conjunto posible de combinaciones de tratamientos, el signo algebraico apropiado para cada contraste, usado para calcular los efectos y las sumas de los cuadrados de los distintos efectos.

Tabla 15.16: Signos para los contrastes en media réplica de un experimento factorial 2^3

Combinación de tratamientos	**Efecto factorial**						
	A	B	C	AB	AC	BC	ABC
a	+	−	−	−	−	+	+
b	−	+	−	−	+	−	+
c	−	−	+	+	−	−	+
abc	+	+	+	+	+	+	+

Observe que en la tabla 15.16, los contrastes A y BC son idénticos, lo cual ilustra los alias. Asimismo, $B \equiv AC$ y $C \equiv AB$. En esta situación se tienen tres contrastes ortogonales que representan los 3 grados de libertad disponibles. Si se obtienen dos observaciones para cada una de las cuatro combinaciones de tratamientos, entonces se tendría un estimador de la varianza del error con 4 grados de libertad. Si se supone que los efectos de la interacción son despreciables, podrían hacerse pruebas para la significancia de todos los efectos principales.

Un ejemplo del efecto y la suma de cuadrados correspondientes es

$$A = \frac{a - b - c + abc}{2n}, \qquad SSA = \frac{(a - b - c + abc)^2}{2^2 n}.$$

En general, la suma de cuadrados con un grado de libertad para cualquier efecto en una fracción 2^{-p} de un experimento factorial 2^k ($p < k$), se obtiene elevando al cuadrado los contrastes en los totales de los tratamientos seleccionados, y dividiendo entre $2^{k-p}\, n$, donde n es el número de réplicas de estas combinaciones de tratamientos.

Ejemplo 15.10: Suponga que se desea emplear una media réplica para estudiar los efectos de cinco factores, cada uno en dos niveles, sobre alguna respuesta, y que se conoce que cualquiera que sea el efecto de cada factor, será constante para cada nivel de los demás factores. En otras palabras, no hay interacciones. Sea el contraste definitorio $ABCDE$ que ocasiona que los efectos principales tengan alias con interacciones de cuatro factores. El agrupamiento de contrastes que implica interacciones provee $15 - 5 = 10$ grados de libertad para el error. Ejecute un análisis de varianza sobre los datos de la tabla 15.17, con la prueba de todos los efectos principales para un nivel de significancia de 0.05.

Solución: Las sumas de cuadrados y los efectos para los efectos principales son

$$SSA = \frac{(11.3 - 15.6 - \cdots - 14.7 + 13.2)^2}{2^{5-1}} = \frac{(-17.5)^2}{16} = 19.14,$$

$$A = -\tfrac{17.5}{8} = -2.19,$$

$$SSB = \frac{(-11.3 + 15.6 - \cdots - 14.7 + 13.2)^2}{2^{5-1}} = \frac{(18.1)^2}{16} = 20.48,$$

$$B = \tfrac{18.1}{8} = 2.26,$$

$$SSC = \frac{(-11.3 - 15.6 + \cdots + 14.7 + 13.2)^2}{2^{5-1}} = \frac{(10.3)^2}{16} = 6.63,$$

Tabla 15.17: Datos para el ejemplo 15.10

Tratamiento	Respuesta	Tratamiento	Respuesta
a	11.3	bcd	14.1
b	15.6	abe	14.2
c	12.7	ace	11.7
d	10.4	ade	9.4
e	9.2	bce	16.2
abc	11.0	bde	13.9
abd	8.9	cde	14.7
acd	9.6	$abcde$	13.2

$C = \frac{10.3}{8} = 1.31,$

$$SSD = \frac{(-11.3 - 15.6 - \cdots + 14.7 + 13.2)^2}{2^{5-1}} = \frac{(-7.7)^2}{16} = 3.71,$$

$D = \frac{-7.7}{8} = -0.96,$

$$SS(E) = \frac{(-11.3 - 15.6 - \cdots + 14.7 + 13.2)^2}{2^{5-1}} = \frac{(8.9)^2}{16} = 4.95,$$

$E = \frac{8.9}{8} = 1.11.$

Todos los demás cálculos y pruebas de significancia se resumen en la tabla 15.18. Las pruebas indican que el factor A tiene un efecto negativo significativo sobre la respuesta; mientras que el factor B lo tiene positivo y significativo. Los factores C, D y E no son significativos al nivel de significancia de 0.05.

Tabla 15.18: Análisis de varianza para los datos de media réplica de un experimento factorial 2^5

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	f calculada
Efecto principal				
A	19.14	1	19.14	6.21
B	20.48	1	20.48	6.65
C	6.63	1	6.63	2.15
D	3.71	1	3.71	1.20
E	4.95	1	4.95	1.61
Error	30.83	10	3.08	
Total	85.74	15		

Ejercicios

15.24 Liste los alias de los diferentes efectos en un experimento factorial 2^5, si el contraste definitorio es *ACDE*.

15.25 *a*) Obtenga una fracción $\frac{1}{2}$ de un diseño factorial 2^4 usando *BCD* como el contraste definitorio.

b) Divida la fracción $\frac{1}{2}$ en 2 bloques de 4 unidades cada uno, confundiendo *ABC*.

c) Construya la tabla de análisis de varianza (fuentes de variación y grados de libertad) para probar todos los efectos principales confundidos, si se acepta que todas las interacciones de los efectos son despreciables.

15.26 Construya una fracción $\frac{1}{4}$ de un diseño factorial 2^6 con el uso de *ABCD* y *BDEF* como los contrastes definitorios. Diga cuáles efectos tienen alias con los seis efectos principales.

15.27 *a*) Con los contrastes definitorios *ABCE* y *ABDF*, obtenga una fracción $\frac{1}{4}$ de un diseño 2^6.

b) Muestre la tabla del análisis de varianza (fuentes de variación y grados de libertad) para todas las pruebas apropiadas, suponiendo que *E* y *F* no intercatúan, y que son despreciables los tres factores y las interacciones mayores.

15.28 En un experimento que implica sólo 16 intentos, se varían siete factores en dos niveles. Se utiliza un experimento factorial 2^7 con una fracción $\frac{1}{8}$, con los contrastes definitorios *ACD*, *BEF* y *CEG*. Los datos son los siguientes:

Comb. de trat.	Respuesta	Comb. de trat.	Respuesta
(1	31.6	*acg*	31.1
ad	28.7	*cdg*	32.0
abce	33.1	*beg*	32.8
cdef	33.6	*adefg*	35.3
acef	33.7	*efg*	32.4
bcde	34.2	*abdeg*	35.3
abdf	32.5	*bcdfg*	35.6
bf	27.8	*abcfg*	35.1

Realice un análisis de varianza sobre los siete efectos principales, suponiendo que las interacciones son despreciables. Use un nivel de significancia de 0.05.

15.29 Se lleva a cabo un experimento de manera que un ingeniero adquiera conocimiento acerca de la influencia de la temperatura de sellado *A*, temperatura de enfriamiento de una barra *B*, porcentaje de aditivo de polietileno *C*, y presión *D*, sobre la resistencia del sello (en gramos por pulgada) de un lote de envoltura para pan. Se emplea un experimento factorial 2^4 con fracción $\frac{1}{2}$, con el contraste definitorio de *ABCD*. A continuación se presentan los datos. Ejecute un análisis de varianza sobre los efectos principales, y las interacciones de dos factores; suponga que todas las interacciones de tres y más factores son despreciables. Use $\alpha = 0.05$.

A	*B*	*C*	*D*	Respuesta
−1	−1	−1	−1	6.6
1	−1	−1	1	6.9
−1	1	−1	1	7.9
1	1	−1	−1	6.1
−1	−1	1	1	9.2
1	−1	1	−1	6.8
−1	1	1	−1	10.4
1	1	1	1	7.3

15.30 En un experimento realizado en el Departamento de Ingeniería Mecánica y analizado por el Centro de Consultoría en Estadística del Instituto Politécnico y Universidad Estatal de Virginia, un sensor detecta una carga eléctrica cada vez que las aspas de una turbina completan una rotación. Luego, el sensor mide la amplitud de la corriente eléctrica. Seis factores son rpm *A*, temperatura *B*, espacio entre las aspas *C*, espacio entre las aspas y la carcasa *D*, ubicación de la entrada *E*, y ubicación del detector *F*. Se utiliza un experimento factorial 2^6 con fracción de $\frac{1}{4}$, y los contrastes definitorios son *ABCE* y *BCDF*. Los datos son los siguientes:

A	*B*	*C*	*D*	*E*	*F*	Respuesta
−1	−1	−1	−1	−1	−1	3.89
1	−1	−1	−1	1	−1	10.46
−1	1	−1	−1	1	1	25.98
1	1	−1	−1	−1	1	39.88
−1	−1	1	−1	1	1	61.88
1	−1	1	−1	−1	1	3.22
−1	1	1	−1	−1	−1	8.94
1	1	1	−1	1	−1	20.29
−1	−1	−1	1	−1	1	32.07
1	−1	−1	1	1	1	50.76
−1	1	−1	1	1	−1	2.80
1	1	−1	1	−1	−1	8.15
−1	−1	1	1	1	−1	16.80
1	−1	1	1	−1	−1	25.47
−1	1	1	1	−1	1	44.44
1	1	1	1	1	1	2.45

Lleve a cabo un análisis de varianza sobre los efectos principales y las interacciones de dos factores, si se acepta que las interacciones de tres factores o más son despreciables. Use $\alpha = 0.05$.

15.31 En un estudio denominado *Durability of Rubber to Steel Adhesively Bonded Joints*, efectuado por el Departamento de Ciencias del Ambiente y Mecánica, y analizado por el Centro de Consultoría en Estadística del Instituto Politécnico y Universidad Estatal de Virginia, un experimentador mide el número de roturas en un sello adhesivo. Se planteó que en esta rotura era

posible que influyeran la concentración de agua marina A, la temperatura B, el ph C, el voltaje D, y la tensión E. Se emplearon un experimento factorial 2^5 con fracción $\frac{1}{2}$ y el contraste definitorio $ABCDE$. Los datos son los siguientes:

A	B	C	D	E	Respuesta
−1	−1	−1	−1	1	462
1	−1	−1	−1	−1	746
−1	1	−1	−1	−1	714
1	1	−1	−1	1	1070
−1	−1	1	−1	−1	474
1	−1	1	−1	1	832
−1	1	1	−1	1	764
1	1	1	−1	−1	1087
−1	−1	−1	1	−1	522
1	−1	−1	1	1	854
−1	1	−1	1	1	773
1	1	−1	1	−1	1068
−1	−1	1	1	1	572
1	−1	1	1	−1	831
−1	1	1	1	−1	819
1	1	1	1	1	1104

Ejecute un análisis de varianza sobre los efectos principales, y las interacciones de dos factores; suponga que las interacciones de tres o más factores son despreciables. Use $\alpha = 0.05$.

15.32 Considere un diseño $2^{5\text{-}1}$ con los factores A, B, C, D y E. Construya el diseño comenzando con uno 2^4 y use $E = ABCD$ como el generador. Obtenga todos los alias.

15.33 Hay seis factores y sólo pueden usarse 8 puntos de diseño. Construya uno 2^{6-3}, comenzando con 2^3, y utilice $D = AB$, $E = -AC$ y $F = BC$ como generadores.

15.34 Considere el ejercicio 15.33. Construya otro 2^{6-3} que sea diferente del diseño elegido en el ejercicio 15.33.

15.35 Para el ejercicio 15.33, proporcione todos los alias para los seis efectos principales.

15.36 En Myers y Montgomery (2002) se analiza una aplicación en la cual a un ingeniero le interesan los efectos del agrietamiento de una aleación de titanio. Los tres factores son A, temperatura; B, contenido de titanio; y C, cantidad de refinador en grano. Los siguientes datos son una parte del diseño y la respuesta, que es la longitud de las grietas inducida en la muestra de la aleación.

A	B	C	Respuesta
−1	−1	−1	0.5269
1	1	−1	2.3380
1	−1	1	4.0060
−1	1	1	3.3640

a) ¿Cuál es la relación definitoria?

b) Dé los alias de los tres efectos principales, con la suposición de que las interacciones de dos factores pueden ser reales.

c) Si se acepta que las interacciones son despreciables, ¿cuál será el factor principal más importante?

d) Para el factor obtenido en el inciso *c*), ¿en qué nivel sugeriría el lector que el factor estuviera para la producción final, alto o bajo?

e) ¿En qué niveles sugeriría el lector que los demás factores estuvieran para la producción final?

f) ¿Qué riesgos hay en las recomendaciones que el lector hizo en los incisos *e*) y *f*)? Responda con amplitud.

15.10 Fracciones superiores y diseños exploratorios

Algunas situaciones industriales requieren que el analista determine cuáles factores, de entre un número grande de ellos, tienen un efecto sobre alguna respuesta importante. Los factores pueden ser cualitativos o variables de clase, variables de regresión, o una mezcla de ambas. El procedimiento analítico quizás requiera análisis de varianza, regresión, o ambos. Es frecuente que el modelo de regresión que se emplee implique sólo los efectos lineales principales; aunque es posible estimar algunas interacciones. La situación exige la exploración de variables, y los diseños experimentales resultantes se denominan **diseños exploratorios**. Es claro que los candidatos viables son los diseños ortogonales de dos niveles saturados o casi saturados.

Resolución del diseño

Con frecuencia, se clasifican los diseños ortogonales de dos niveles, de acuerdo con su **resolución**, y ésta queda determinada por la siguiente definición.

Definición 15.1: La **resolución** de un diseño ortogonal de dos niveles es la longitud de la interacción más pequeña (menos compleja), de entre el conjunto de contrastes definitorios.

Si el diseño se construye como factorial completo o fraccionario [ya sea 2^k o 2^{k-p} $(p = 1, 2, \ldots, k-1)$], el concepto de resolución del diseño es una ayuda para determinar el efecto de los alias. Por ejemplo, un diseño con resolución II sería de poca utilidad, ya que habría al menos una instancia de alias entre un efecto principal y otro. Un diseño con resolución III tendría todos sus efectos principales (lineales) ortogonales entre sí. No obstante, habrá algunos alias entre los efectos lineales y las interacciones de dos factores. Entonces, queda claro que si el analista está interesado en estudiar los efectos principales (lineales en el caso de la regresión) y no hay interacciones de dos factores, entonces se requiere un diseño cuya resolución sea de al menos III.

15.11 Construcción de diseños con resoluciones III y IV, con 8, 16 y 32 puntos de diseño

Para 2 a 7 variables con 8 puntos de diseño, es posible construir diseños útiles con resoluciones III y IV. Se comienza sencillamente con un factorial 2^3 que haya sido saturado simbólicamente con interacciones.

$$\begin{array}{c} x_1 \quad x_2 \quad x_3 \quad x_1x_2 \quad x_1x_3 \quad x_2x_3 \quad x_1x_2x_3 \\ \left[\begin{array}{rrrrrrr} -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}\right] \end{array}$$

Es claro que un diseño con resolución III se puede construir tan sólo reemplazando las columnas de las interacciones por efectos principales nuevos para las siete variables. Por ejemplo, si se define

$$\begin{aligned} x_4 &= x_1x_2 \qquad \text{(contraste definitorio } ABD\text{)} \\ x_5 &= x_1x_3 \qquad \text{(contraste definitorio } ACE\text{)} \\ x_6 &= x_2x_3 \qquad \text{(contraste definitorio } BCF\text{)} \\ x_7 &= x_1x_2x_3 \qquad \text{(contraste definitorio } ABCG\text{)} \end{aligned}$$

y se obtiene una fracción 2^{-4} de un factorial 2^7. Las expresiones anteriores identifican los contrastes definitorios elegidos. Resultan once contrastes definitorios adicionales y todos contienen al menos tres letras. Así, el diseño es de resolución III. Es claro que si se comienza con un *subconjunto* de columnas aumentadas y se concluye con un diseño que implica menos de siete variables de diseño, el resultado es un diseño de resolución III en menos de siete variables.

Es posible construir un conjunto similar de diseños posibles para 16 puntos de diseño, comenzando con uno 2^4 saturado con interacciones. Las definiciones de las variables que corresponden a estas interacciones producen diseños de resolución III

Tabla 15.19: Algunos diseños 2^{k-p} de resoluciones III, IV y V

Número de factores	Diseño	Número de puntos	Generadores
3	2^{3-1}_{III}	4	$C = \pm\, AB$
4	2^{4-1}_{IV}	8	$D = \pm\, ABC$
5	2^{5-2}_{IV}	8	$D = \pm\, AB;\ E = \pm\, AC$
6	2^{6-1}_{VI}	32	$F = \pm BCD$
	2^{6-2}_{VI}	16	$E = \pm\, ABC;\ F = \pm BCD$
	2^{6-3}_{III}	8	$D = \pm\, AB;\ F = \pm\, BC;\ E = \pm AC$
7	2^{7-1}_{VII}	64	$G = \pm\, ABCDEF$
	2^{7-2}_{IV}	32	$E = \pm\, ABC;\ G = \pm\, ABDE$
	2^{7-3}_{IV}	16	$E = \pm\, ABC;\ F = \pm\, BCD;\ G = \pm\, ACD$
	2^{7-4}_{III}	8	$D = \pm\, AB;\ E = \pm\, AC;\ F = \pm\, BC;\ G = \pm\, ABC$
8	2^{8-2}_{V}	64	$G = \pm\, ABCD;\ H = \pm\, ABEF$
	2^{8-3}_{IV}	32	$F = \pm\, ABC;\ G = \pm\, ABD;\ H = \pm\, BCDE$
	2^{8-4}_{IV}	16	$E = \pm\, BCD;\ F = \pm\, ACD;\ G = \pm\, ABC;\ H = \pm\, ABD$

a través de 15 variables. En forma similar, se pueden construir diseños que contengan 32 corridas, comenzando con uno 2^5.

La tabla 15.19 proporciona al usuario los lineamientos para construir 8, 16, 32 y 64 diseños puntuales, con resoluciones III, IV e incluso V. La tabla da el número de factores, el número de corridas y los generadores que se emplean para producir los diseños 2^{k-p}. El generador dado se emplea para **aumentar el factorial completo** que contiene $k - p$ factores.

La técnica de duplicación

Es posible ampliar los diseños de resolución III ya descritos, para producir un diseño de resolución IV, usando la **técnica de duplicación**. En esta técnica se duplica el tamaño del diseño sumando los *negativos* de la matriz de diseño construida según se describió anteriormente. La tabla 15.20 muestra un diseño de resolución IV con 16 corridas en 7 variables, construido con la técnica de duplicación. Es evidente que podemos construir diseños con resolución IV que incluyan hasta 15 variables, con el empleo de la técnica de duplicación sobre diseños desarrollada con el diseño 2^4 saturado.

Este diseño se construye "duplicando" una fracción $\frac{1}{8}$ de uno 2^6. La última columna se agrega como un séptimo factor. En la práctica, es frecuente que la última columna juegue el papel de variable bloqueadora. No es raro que la técnica de duplicación se emplee en experimentación secuencial, en la cual se analizan los datos del diseño inicial de resolución III. Entonces, con base en el análisis, el experimentador percibe si es necesario un diseño con resolución IV. Como resultado, quizá sea necesaria una variable bloqueadora debido a la separación en el tiempo que ocurre entre las dos partes del experimento. Además de la variable bloqueadora, el diseño final es un experimento 2^6 con fracción $\frac{1}{4}$.

Tabla 15.20: Diseño de resolución IV con dos niveles, con la técnica de duplicación

x_1	x_2	x_3	$x_4 = x_1x_2$	$x_5 = x_1x_3$	$x_6 = x_2x_3$	x_7
−1	−1	−1	1	1	1	−1
1	−1	−1	−1	−1	1	−1
−1	1	−1	−1	1	−1	−1
−1	−1	1	1	−1	−1	−1
1	1	−1	1	−1	−1	−1
1	−1	1	−1	1	−1	−1
−1	1	1	−1	−1	1	−1
1	1	1	1	1	1	−1
Duplicación						
1	1	1	−1	−1	−1	1
−1	1	1	1	1	−1	1
1	−1	1	1	−1	1	1
1	1	−1	−1	1	1	1
−1	−1	1	−1	1	1	1
−1	1	−1	1	−1	1	1
1	−1	−1	1	1	−1	1
−1	−1	−1	−1	−1	−1	1

15.12 Otros diseños de resolución III con dos niveles; los diseños de Plackett-Burman

Una familia de diseños desarrollada por Plackett y Burman (véase la Bibliografía) vino a llenar los huecos que existían con los factoriales fraccionarios. Éstos son útiles con muestras de tamaño 2^r (es decir, incluyen muestras de tamaño 4, 8, 16, 32, 64, ...). Los diseños de Plackett y Burman requieren $2r$ puntos de diseño, por lo que se dispone de diseños de tamaño 12, 20, 24, 28, etcétera. Estos diseños de Plackett-Burman de dos niveles tienen resolución III y son muy fáciles de construir. Para cada tamaño de muestra se dan "renglones básicos". El número de renglones de signos + y − es $n - 1$. Para construir las columnas de la matriz de diseño, se comienza con el renglón básico y se hace una permutación cíclica sobre las columnas, hasta que queden formadas k (el número deseado de variables) columnas. Después, se llena el último renglón con signos negativos. El resultado será un diseño de resolución III con k variables ($k = 1, 2, \ldots, N$). Los renglones básicos son los siguientes:

$N = 12$ + + − + + + − − − + −

$N = 16$ + + + + − + − + + − − + − − −

$N = 20$ + + − − + + + + − + − + − − − − + + −

$N = 24$ + + + + + − + − + + − − + + − − + − + − − − −

Ejemplo 15.11: Construya un diseño exploratorio de dos niveles con 6 variables que contengan 12 puntos de diseño.

Solución: Comience con el renglón básico en la columna inicial. La segunda columna se forma llevando la entrada inferior de la primera columna a la parte superior de la segunda,

y repitiendo la primera. La tercera columna se forma del mismo modo, utilizando las entradas de la segunda columna. Cuando haya un número suficiente de columnas, **sencillamente se llena el último renglón con signos negativos**. El diseño resultante es como sigue:

$$\begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ \left[\begin{array}{cccccc} + & - & + & - & - & - \\ + & + & - & + & - & - \\ - & + & + & - & + & - \\ + & - & + & + & - & + \\ + & + & - & + & + & - \\ + & + & + & - & + & + \\ - & + & + & + & - & + \\ - & - & + & + & + & - \\ - & - & - & + & + & + \\ + & - & - & - & + & + \\ - & + & - & - & - & + \\ - & - & - & - & - & - \end{array}\right] \end{array}$$

Los diseños de Plackett-Burman son populares en la industria para situaciones de exploración. Como diseños de resolución III, todos los efectos lineales son ortogonales. Para cualquier tamaño de muestra, el usuario dispone de un diseño para $k = 2, 3, \ldots, N - 1$ variables.

La estructura de alias para el diseño de Plackett-Burman es muy complicada, por lo que el usuario no puede construir el diseño con un control completo sobre ella, como sí era posible en el caso de diseños 2^k o 2^{k-p}. Sin embargo, en el caso de modelos de regresión, el diseño de Plackett-Burman acepta interacciones (aunque no sean ortogonales) cuando se dispone de suficientes grados de libertad.

15.13 Diseño de parámetros robustos

En este capítulo se destacó el concepto del empleo del diseño de experimentos (DE) para adquirir conocimientos sobre procesos de ingeniería y científicos. En el caso en que un proceso incluya un producto, es posible usar el DE para mejorar el producto o la calidad. Como se dijo en el capítulo 1, se ha dado mucha importancia al uso de métodos estadísticos en la mejoría de productos. Durante las décadas de 1980 y 1990 un aspecto importante de tal mejoramiento fue reforzar la calidad en los procesos y productos, en la etapa de investigación o de diseño del proceso. Es frecuente que se requiera del DE en el desarrollo de procesos con las propiedades siguientes:

1. Insensibles (robustos) a las condiciones ambientales.
2. Insensibles (robustos) a factores que dificultan el control.
3. Proporcionan variación mínima en cuanto al desempeño.

Estos métodos se denominan con frecuencia *diseño de parámetros robustos* (véase Taguchi, Taguchi y Wu, y Kackar, en la Bibliografía). En este contexto, el término *diseño* se refiere al diseño de los procesos o sistema; en tanto que *parámetros* se refiere a los parámetros en el sistema. Éstos son lo que hemos estado llamando *factores* o *variables*.

Queda claro que las metas 1, 2 y 3 mencionadas son muy nobles. Por ejemplo, un ingeniero petrolero puede tener una buena mezcla de gasolina que se desempeñe muy bien en condiciones ideales y estables. Sin embargo, el desempeño quizá se deteriorara debido a cambios en las condiciones ambientales, como tipo de conductor, factores climáticos, tipo de motor, etcétera. Un científico en una compañía de alimentos quizá tenga una mezcla para pasteles que sea muy buena, a menos que el usuario no siga con exactitud las instrucciones acerca de temperatura del horno, tiempo de cocción, y otras parecidas. Un producto o proceso cuyo desempeño sea consistente cuando se expone a esas condiciones ambientales cambiantes se denomina **producto robusto** o **proceso robusto**. [Véase Myers y Montgomery (2002), en la Bibliografía.]

Variables de control y ruido

Taguchi hace énfasis en utilizar dos tipos de variables de diseño en un estudio. Ésos son los factores de control y los factores de ruido.

Definición 15.2: Los **factores de control** son variables que se pueden controlar tanto en el experimento como en el proceso. Los **factores de ruido** son variables que pueden controlarse o no en el experimento, pero no pueden controlarse en el proceso (o al menos no bien).

Un enfoque importante es usar en el mismo experimento variables de control y variables de ruido, como efectos fijos. En este sentido, es frecuente usar diseños o arreglos ortogonales.

Meta del diseño de parámetros robustos: La meta del diseño de parámetros robustos es elegir los niveles de variables de control (es decir, el diseño del proceso) que sean más robustas (insensibles) a los cambios en las variables de ruido.

Debe observarse que los *cambios en las variables de ruido* en realidad implican cambios durante el proceso: cambios en el campo, en el ambiente, en el manejo o uso por parte del consumidor, etcétera.

El arreglo de productos

Un enfoque hacia el diseño de experimentos, que incluye variables tanto de control como de ruido, es el uso de un plan experimental que requiera un diseño ortogonal tanto para las variables de control como de ruido, por separado. Entonces, el experimento completo es tan sólo el producto o cruce de estos dos diseños ortogonales. El siguiente es un ejemplo sencillo de un arreglo de productos con dos variables de control y dos de ruido.

Ejemplo 15.12: En el artículo "The Taguchi Approach to Parameter Design", por D. M. Byrne y S. Taguchi, en *Quality Progress*, de diciembre de 1987, los autores estudian un ejemplo interesante en el que se busca un método para ensamblar un conector electrométrico a un tubo de nailon, que genera el rendimiento necesario para el empuje apropiado para lograr una aplicación en un motor de automóvil. El objetivo es encontrar condiciones controlables que maximicen la fuerza del empuje. Entre las variables controlables están A, el espesor de la pared del conector; y B, la profundidad de inserción. Hay varias variables que durante la operación rutinaria no se pueden

controlar, aunque durante el experimento sí. Entre ellas están C, condiciones del tiempo; y D, condiciones de temperatura. Se toman tres niveles para cada variable de control, y dos para cada variable de ruido. El resultado del arreglo cruzado es el siguiente. El arreglo de control es de 3×3, y el de ruido es un factorial 2^2 que resulta familiar con (1), c, d y cd, que representan las combinaciones de factores. El propósito del factor de ruido es crear la *clase de variabilidad en la respuesta, fuerza de empuje, que podría esperarse con el proceso durante la operación cotidiana*. En la tabla 15.21 se muestra el diseño.

Tabla 15.21: Diseño para el ejemplo 15.12

		***B* (profundidad)**		
		Baja	**Media**	**Alta**
***A* (espesor de pared)**	**Delgado**	(1)	(1)	(1)
		c	c	c
		d	d	d
		cd	cd	cd
	Medio	(1)	(1)	(1)
		c	c	c
		d	d	d
		cd	cd	cd
	Grueso	(1)	(1)	(1)
		c	c	c
		d	d	d
		cd	cd	cd

Análisis

Hay varios procedimientos para analizar el arreglo de productos. El enfoque citado por Taguchi y adoptado por muchas compañías de Estados Unidos relacionadas con procesos de manufactura, implica, inicialmente, la formación de un estadístico resumido con cada combinación en el arreglo de control. Dicho estadístico resumido se denomina *razón señal a ruido*. Suponga que se denota con $y_1, y_2, \ldots, y_n$, un conjunto común de corridas experimentales para el arreglo de ruido en una combinación de arreglo de control fijo. La tabla 15.22 describe algunas de las razones comunes SN.

Tabla 15.22: Razones SN comunes para distintos objetivos

Objetivos	**Razón *SN***
Maximizar la respuesta	$SN_L = -10 \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{y_i^2} \right)$
Lograr el objetivo	$SN_T = 10 \log \left(\frac{\bar{y}^2}{s^2} \right)$
Minimizar la respuesta	$SN_s = -10 \log \left(\frac{1}{n} \sum_{i=1}^{n} y_i^2 \right)$

Para cada uno de los casos anteriores, se desea encontrar la combinación de las variables de control que *maximice SN*.

Ejemplo 15.13: **Estudio de caso** En un experimento que se describe en *Understanding Industrial Designed Experiments*, por Schmidt y Launsby (véase la Bibliografía), en una planta de ensamble de circuitos integrados se lleva a cabo la optimización de un proceso de soldadura. Las partes se insertan a mano o en forma automática en una tarjeta con un circuito impreso en ella. Una vez que se han insertado, la tarjeta se coloca en una máquina soldadora de onda que se emplea para conectar todos los elementos del circuito. Las tarjetas se colocan en una banda y pasan por una serie de etapas. Se sumergen en una mezcla en movimiento para quitar el óxido. Para minimizar el pandeo se calientan antes de aplicar la soldadura. Ésta se realiza conforme las tarjetas se mueven a través de la onda de soldar. El objetivo del experimento es minimizar el número de defectos de soldadura por millón de uniones. El factor y los niveles de control se dan en la tabla 15.23.

Tabla 15.23: Factores de control para el ejemplo 15.13

Factor	$(-\mathbf{1})$	$(+\mathbf{1})$
A, temperatura de la vasija de soldar (°F)	480	510
B, velocidad de la banda (pies/min)	7.2	10
C, densidad de la mezcla	0.9°	1.0°
D, temperatura de precalentamiento	150	200
E, altura de onda (pulg)	0.5	0.6

Es fácil controlar estos factores en el nivel experimental, pero difícil en extremo en el nivel de proceso o de planta.

Factores de ruido: Tolerancias de los factores de control

Es frecuente que en los procesos como éste, uno de los factores naturales de ruido sean las tolerancias de los factores de control. Por ejemplo, en el proceso real *on line*, la temperatura de la vasija de soldar y la velocidad de la banda son difíciles de controlar. Se sabe que el control de la temperatura está dentro de ± 5 °F, y que el control de la velocidad de la banda está dentro de ± 0.2 pies/min. Es muy concebible que la variabilidad de la respuesta del producto (desempeño de la soldadura) se incremente debido a la incapacidad de controlar esos dos factores en sus niveles nominales. El tercer factor de ruido es el tipo de ensamble involucrado. En la práctica, se emplean dos tipos de ensambles. Así, se tienen los factores de ruido que se dan en la tabla 15.24.

Tanto el arreglo de control (arreglo interior) y el de ruido (arreglo exterior) se eligieron para ser factoriales fraccionarios: el primero es $\frac{1}{4}$ de uno 2^5, y el segundo $\frac{1}{2}$ de uno 2^3. El *arreglo cruzado* y los valores de respuesta se presentan en la tabla 15.25. Las tres columnas primeras del arreglo interior representan un 2^3. Las columnas están formadas por $D = -AC$ y $E = -BC$. Así, las interacciones definitorias para el arreglo interior son ACD, BCE y ADE. El arreglo exterior es una fracción de 2^3 fraccionario de resolución III. Observe que cada punto del arreglo interior contiene corridas del arreglo exterior. Así, se observan cuatro valores de respuesta en cada combinación del arreglo de control. La figura 15.16 muestra gráficas que revelan el efecto que tienen la temperatura y la densidad sobre la respuesta media.

Tabla 15.24: Factores de ruido para el ejemplo 15.13

Factor	(−1)	(+1)
A*, tolerancia de la temperatura de la vasija de soldar (°F) (desviación de la nominal)	−5°	−5°
B*, tolerancia de deslizamiento de la banda (pies/min) (desviación del ideal)	−0.2	+0.2
C*, tipo de ensamble	1	2

Tabla 15.25: Arreglos cruzados y valores de la respuesta para el ejemplo 15.13

Arreglo interior					Arreglo exterior				
A	B	C	D	E	(1)	a^*b^*	a^*c^*	b^*c^*	SN_S
1	1	1	−1	−1	194	197	193	275	−46.75
1	1	−1	1	1	136	136	132	136	−42.61
1	−1	1	−1	1	185	261	264	264	−47.81
1	−1	−1	1	−1	47	125	127	42	−39.51
−1	1	1	1	−1	295	216	204	293	−48.15
−1	1	−1	−1	1	234	159	231	157	−45.97
−1	−1	1	1	1	328	326	247	322	−45.76
−1	−1	−1	−1	−1	186	187	105	104	−43.59

La temperatura y la densidad del flujo son los factores más importantes. Parecen influir tanto en $(SN)_S$ como en $\bar{y}$. Por fortuna, la *temperatura elevada* y la *densidad baja del flujo* son preferibles tanto para $(SN)_S$ como para la respuesta media. Así, las condiciones "óptimas" son

$$\text{temperatura de la soldadura} = 510\ °\text{F}, \qquad \text{densidad del flujo} = 0.9°.$$

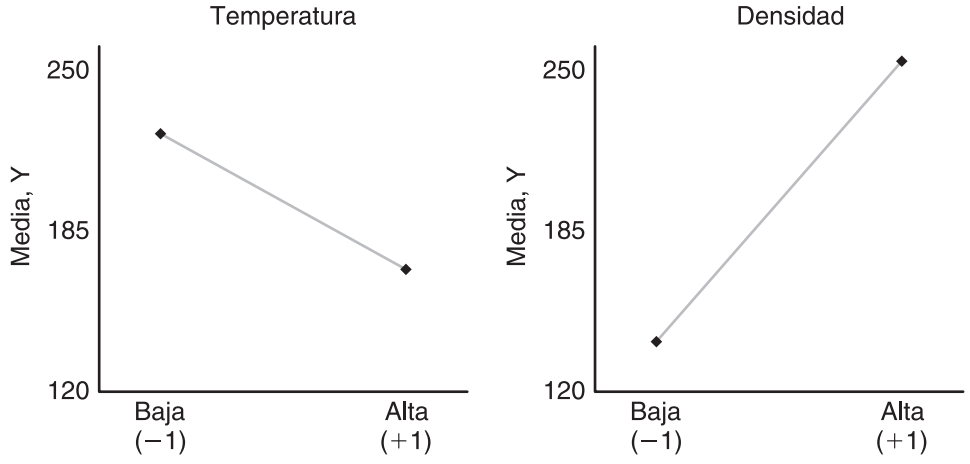


Figura 15.16: Gráfica que muestra la influencia de los factores sobre la respuesta media.

Enfoques alternativos al diseño de parámetros robustos

Un enfoque que sugieren muchos estudiosos consiste en modelar la media y la varianza muestrales por separado, en vez de combinar los dos conceptos separados mediante una razón señal a ruido. Con frecuencia, el modelado separado ayuda al experimentador a obtener una mejor comprensión del proceso que interviene. En el ejemplo siguiente se ilustra este enfoque con el experimento del proceso de soldadura.

Ejemplo 15.14: Considere los datos del ejemplo 15.13. Un análisis alternativo es ajustar modelos separados para la media $\bar{y}$ y la desviación estándar muestral. Suponga que para los factores de control se usa el código habitual $+1$ y -1. Con base en la importancia aparente de la temperatura de la vasija de soldar x_1, y la densidad del flujo x_2, el modelo de la regresión lineal sobre la respuesta (número de errores por millón de uniones) produce el modelo

$$\hat{y} = 197.125 - 27.5x_1 + 57.875x_2.$$

Para encontrar el nivel más robusto de la temperatura y densidad del flujo, es conveniente establecer un compromiso entre la respuesta media y la variabilidad, lo cual requiere modelar esta última. Una herramienta importante para hacerlo es la transformación logarítmica (véase Bartlett y Kendall o Carroll y Ruppert):

$$\ln s^2 = \gamma_0 + \gamma_1(x_1) + \gamma_2(x_2).$$

Este proceso de modelado produce el siguiente resultado:

$$\widehat{\ln s^2} = 6.7692 - 0.8178x_1 + 0.6877x_2.$$

El análisis que es importante para el científico o el ingeniero consiste en utilizar los dos modelos en forma simultánea. Es de mucha utilidad un enfoque gráfico. La figura 15.17 muestra al mismo tiempo gráficas sencillas de la media y desviación estándar. Como era de esperar, la ubicación de la temperatura y la densidad del flujo que minimizan el número medio de errores es la misma que aquella que minimiza la variabilidad, es decir, temperatura alta y densidad del flujo baja. El enfoque gráfico de la respuesta múltiple permite que el usuario perciba intercambios entre la media del proceso y su variabilidad. Para este ejemplo, el ingeniero quizás esté insatisfecho con las condiciones extremas de la temperatura de la soldadora y la densidad de flujo. La figura ofrece una estimación de las condiciones de la media y la variabilidad que indican cuánto se pierde conforme se aleja de las condiciones óptimas a otras intermedias.

Ejercicios

15.37 Use los datos de limpieza del carbón del ejercicio 15.2 de la página 622 para ajustar un modelo del tipo

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

donde los niveles son

x_1: porcentaje de sólidos: 8; 12
x_2: tasa de flujo: 150; 250 gal/min
x_3: pH: 5; 6

Centre y dé escala a las variables de las unidades de diseño. Asimismo, realice una prueba para la falta de ajuste, y haga comentarios acerca de lo adecuado del modelo de regresión lineal.

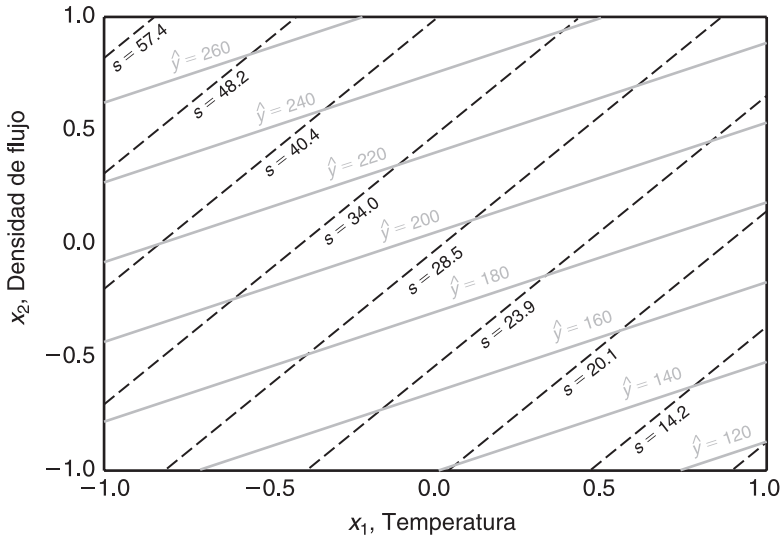


Figura 15.17: Media y desviación estándar para el ejemplo 15.14.

15.38 Se utiliza un plan factorial 2^5 para construir un modelo de regresión que contenga coeficientes de primer orden y términos de modelado para todas las interacciones de dos factores. Para cada factor se realizan corridas duplicadas. Construya la tabla de análisis de varianza que muestre los grados de libertad para la regresión, falta de ajuste y error puro.

15.39 Considere el factorial 2^7 con la fracción $\frac{1}{16}$ que se estudió en la sección 15.11. Liste los 11 contrastes definitorios adicionales.

15.40 Construya un diseño de Plackett-Burman para 10 variables que contengan 24 corridas experimentales.

Ejercicios de repaso

15.41 Se emplea un diseño de Plackett-Burman para estudiar las propiedades reológicas de los copolímeros de peso molecular elevado. En el experimento se fijan dos niveles para cada una de seis variables. La respuesta es la viscosidad del polímero. Los datos fueron analizados en el Centro de Consultoría en Estadística del Instituto Politécnico y Universidad Estatal de Virginia, por personal del Departamento de Ingeniería Química de la universidad. Las variables son las siguientes: química del bloque duro x_1, tasa de flujo de nitrógeno x_2, tiempo de calentamiento x_3, porcentaje de compresión x_4, observaciones alta y baja x_5, deformación porcentual x_6. A continuación se presentan los datos. Construya una ecuación de regresión que relacione la viscosidad con los niveles de las seis variables. Realice pruebas t para todos los efectos principales. Recomiende los factores que deban conservarse para estudios futuros y aquellos que no. Emplee la media cuadrática residual (5 grados de libertad) como medida del error experimental.

Obs.	x_1	x_2	x_3	x_4	x_5	x_6	y
1	1	−1	1	−1	−1	−1	194,700
2	1	1	−1	1	−1	−1	588,400
3	−1	1	1	−1	1	−1	7,533
4	1	−1	1	1	−1	1	514,100
5	1	1	−1	1	1	−1	277,300
6	1	1	1	−1	1	1	493,500
7	−1	1	1	1	−1	1	8,969
8	−1	−1	1	1	1	−1	18,340
9	−1	−1	−1	1	1	1	6,793
10	1	−1	−1	−1	1	1	160,400
11	−1	1	−1	−1	−1	1	7,008
12	−1	−1	−1	−1	−1	−1	3,637

15.42 Una compañía petrolera grande del suroeste lleva a cabo experimentos de manera regular para probar aditivos de los fluidos de perforación. La viscosidad plástica es una medición reológica que refleja el espesor del fluido. A éste se agregan varios polímeros con la finalidad de incrementar su viscosidad. Los que siguen

son datos de dos polímeros que se usaron en dos niveles cada uno, y la viscosidad medida. La concentración de los polímeros se indica como "baja" y "alta". Haga un análisis de un experimento factorial 2^2. Pruebe los efectos e interacción de los dos polímeros.

	Polímero 1			
Polímero 2	Baja		Alta	
Baja	3	3.5	11.3	12.0
Alta	11.7	12.0	21.7	22.4

15.43 Se analiza un experimento factorial 2^2 en el Centro de Consultoría en Estadística del Instituto Politécnico y Universidad Estatal de Virginia. El cliente es un miembro del Department of Housing, Interior Design, and Resource Management. A éste le interesa comparar el arranque en frío contra el precalentamiento de los hornos, en términos de la energía total que se transmite al producto. Además, se comparan las condiciones de la convección en modo regular. Se hicieron cuatro corridas experimentales con cada una de las cuatro combinaciones de los factores. A continuación se presentan los datos del experimento:

	Precalentamiento		**Frío**	
Modo de convección	618 629	619.3 611	575 574	573.7 572
Modo regular	581 581	585.7 595	558 562	562 566

Haga un análisis de varianza para estudiar la interacción y los efectos principales. Saque sus conclusiones.

15.44 Construya un diseño que incluya 12 corridas en las que se varíen 2 factores con 2 niveles cada uno. El lector está restringido a utilizar bloques de tamaño dos, y debe ser capaz de realizar pruebas de significancia sobre ambos efectos principales y el efecto de la interacción.

15.45 En el estudio denominado *The Use of Regression Analysis for Correcting Matrix Effects in the X-Ray Fluorescence Analysis of Pyrotechnic Compositions*, publicado en *Proceedings of the Tenth Conference on the Design of Experiments in Army Research Development and Testing*, ARO-D Report 65-3 (1965), se describe un experimento donde se hicieron variar las concentraciones de 4 componentes de una mezcla impulsora y los pesos de partículas finas y gruesas en el compuesto acuoso. Los factores A, B, C y D, cada uno en dos niveles, representan las concentraciones de los 4 componentes; y los factores E y F, también en dos niveles, representan los pesos de las partículas finas y gruesas que hay en el compuesto. El objetivo del análisis es determinar si las razones de la intensidad de los rayos X, asociadas con el componente 1 del combustible, estaban influidas en forma significativa por la variación de las concentraciones de los distintos componentes y los pesos de las partículas según su tamaño en la mezcla. Se utilizó un experimento factorial 2^6 con fracción $\frac{1}{8}$, con los contrastes definitorios ADE, BCE y ACF. Los datos siguientes representan el total de un par de lecturas de la intensidad:

Lote	**Combinación de tratamientos**	**Total de la razón de intensidad**
1	$abef$	2.2480
2	$cdef$	1.8570
3	(1)	2.2428
4	ace	2.3270
5	bde	1.8830
6	$abcd$	1.8078
7	adf	2.1424
8	bcf	1.9122

El error cuadrático medio agrupado con ocho grados de libertad es 0.02005. Analice los datos con el empleo de un nivel de significancia de 0.05, para determinar si las concentraciones de los componentes y los pesos de las partículas finas y gruesas, presentes en el compuesto, tienen una influencia significativa sobre las razones de intensidad asociadas con el componente 1. Suponga que no existe interacción entre los 6 factores.

15.46 Haga el esquema de bloques para un experimento factorial 2^7 con ocho bloques, cada uno de tamaño 16, usando $ABCD$, $CDEFG$ y BDF, como contrastes definitorios. Indique cuáles interacciones resultan sacrificadas por completo en el experimento.

15.47 Utilice la tabla 15.19 para construir un diseño de 16 corridas con 8 factores, que tenga resolución IV.

15.48 En el diseño del ejercicio de repaso 15.47, compruebe que el diseño en efecto tiene resolución IV.

15.49 Construya un diseño que contenga nueve puntos de diseño, sea ortogonal, contenga un total de 12 corridas, 3 grados de libertad para el error de repetición, y permita una prueba de falta de ajuste para la curvatura cuadrática pura.

15.50 Considere un diseño 2^{3-1}_{III} con 2 corridas centrales. Considere $\bar{y}_f$ como la respuesta promedio en el parámetro de diseño, y $\bar{y}_0$ como la respuesta promedio en el centro del diseño. Suponga que el verdadero modelo de la regresión es

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2.$$

a) Proporcione (y compruebe) $E(\bar{y}_f - \bar{y}_0)$.
b) Explique lo que haya aprendido del resultado del inciso *a*).

15.14 Nociones erróneas y riesgos potenciales; relación con el material de otros capítulos

En el empleo de experimentos factoriales fraccionarios, una de las consideraciones más importantes a la que el analista debe estar atento es la *resolución del diseño*. Un diseño de resolución baja es más pequeño (y, por lo tanto, menos costoso) que otro de resolución mayor. Sin embargo, se paga un precio por el diseño más barato. El diseño de menor resolución tiene alias más pesados que otro de resolución mayor. Por ejemplo, si el investigador tiene la sospecha de que las interacciones de dos factores son importantes, entonces no debería emplearse la resolución III. Un diseño de resolución III es estrictamente un **plan de efectos principales**.

Capítulo 16

Estadística no paramétrica

16.1 Pruebas no paramétricas

La mayoría de los procedimientos de prueba de hipótesis que se presentaron en los capítulos anteriores se basan en la suposición de que las muestras aleatorias se seleccionan de poblaciones normales. Por fortuna, la mayor parte de estas pruebas aún son confiables cuando experimentamos ligeras desviaciones de la normalidad, en particular cuando el tamaño de la muestra es grande. Tradicionalmente, tales procedimientos de prueba se denominan **métodos paramétricos**. En este capítulo consideramos varios procedimientos de prueba alternativos, llamados **no paramétricos** o **métodos de distribución libre**, que a menudo no suponen conocimiento de ninguna clase acerca de las distribuciones de las poblaciones fundamentales, excepto quizá que éstas son continuas.

Los procedimientos no paramétricos o de distribución libre se utilizan con mayor frecuencia por los analistas de datos. Hay muchas aplicaciones en la ciencia y la ingeniería donde los datos se reportan no como valores de un continuo, sino más bien en una **escala ordinal** tal que es bastante natural asignar rangos a los datos. De hecho, en este capítulo el lector notará muy pronto que los métodos de distribución libre que se describen aquí implican un *análisis de rangos*. La mayoría de los analistas encuentran que los cálculos que se intervienen en los métodos no paramétricos son muy atractivos e intuitivos.

Un ejemplo donde se aplica una prueba no paramétrica es el siguiente. Dos jueces deben clasificar cinco marcas de cerveza de alta calidad mediante la asignación de un grado de 1 a la marca que se considera que tiene la mejor calidad global, un grado de 2 a la segunda mejor, etcétera. Entonces, se puede utilizar una prueba no paramétrica para determinar donde existe algún acuerdo entre los dos jueces.

También debemos señalar que hay varias desventajas asociadas con las pruebas no paramétricas. En primer lugar, no utilizan toda la información que proporciona la muestra y, por ello, una prueba no paramétrica será menos eficiente que el procedimiento paramétrico correspondiente, cuando ambos métodos son aplicables. En consecuencia, para lograr la misma potencia, una prueba no paramétrica requerirá un tamaño muestral mayor que el que requeriría la prueba no paramétrica correspondiente.

Como indicamos antes, ligeras divergencias de la normalidad tienen como resultado desviaciones menores del ideal para las pruebas paramétricas estándar. Esto es

particularmente cierto para la prueba t y la prueba F. En el caso de la prueba t y la prueba F, el valor P citado tal vez sea ligeramente erróneo si existe una trasgresión moderada de la suposición de normalidad.

En resumen, si se puede aplicar tanto una prueba paramétrica como una no paramétrica al mismo conjunto de datos, deberíamos aplicar la técnica paramétrica más eficiente. Sin embargo, debemos reconocer que las suposiciones de normalidad a menudo no se pueden justificar, y que no siempre tenemos mediciones cuantitativas. Es una fortuna que los estadísticos nos brinden diversos procedimientos no paramétricos útiles. Armado con las técnicas no paramétricas, el analista de datos tiene más municiones para adaptar una variedad más amplia de situaciones experimentales. Se debe señalar que incluso bajo las suposiciones de la teoría normal estándar, las eficiencias de las técnicas no paramétricas son notablemente cercanas a las del procedimiento paramétrico correspondiente. Por otro lado, las divergencias serias de la normalidad hacen que el método no paramétrico se vuelva más eficiente que el procedimiento paramétrico.

Prueba de signo

El lector debería recordar que los procedimientos que se estudian en la sección 10.7, para probar la hipótesis nula de que $\mu = \mu_0$, son válidos sólo si la población es aproximadamente normal o si la muestra es grande. Sin embargo, si $n < 30$ y la población es decididamente no normal, debemos recurrir a una prueba no paramétrica.

La prueba de signo se utiliza para probar hipótesis sobre una *mediana* poblacional. En el caso de muchos de los procedimientos no paramétricos, la media se reemplaza por la mediana como el **parámetro de ubicación** pertinente bajo prueba. Recuerde que la mediana muestral se definió en la sección 1.4. La contraparte poblacional, que se denota con $\tilde{\mu}$, tiene una definición análoga. Dada una variable aleatoria X, $\tilde{\mu}$ se define de modo que $P(X > \tilde{\mu}) \leq 0.5$ y $P(X < \tilde{\mu}) \leq 0.5$. En el caso continuo,

$$P(X > \tilde{\mu}) = P(X < \tilde{\mu}) = 0.5.$$

Por supuesto, si la distribución es simétrica, la media y la mediana poblacionales son iguales. Al probar la hipótesis nula H_0 de que $\tilde{\mu} = \tilde{\mu}_0$ contra una alternativa adecuada, sobre la base de una muestra aleatoria de tamaño n, reemplazamos cada valor de la muestra que exceda a $\tilde{\mu}_0$ con un signo *más*, y cada valor de la muestra menor que $\tilde{\mu}_0$ con un signo *menos*. Si la hipótesis nula es verdadera y la población es simétrica, la suma de los signos más debería ser aproximadamente igual a la suma de los signos menos. Cuando un signo aparece con más frecuencia de lo que debería, con base sólo en el azar, rechazamos la hipótesis de que la mediana poblacional $\tilde{\mu}$ es igual a $\tilde{\mu}_0$.

En teoría la prueba de signo se aplica tan sólo en situaciones donde $\tilde{\mu}_0$ no puede ser igual al valor de cualquiera de las observaciones. Aunque hay una probabilidad cero de obtener una observación muestral exactamente igual a $\tilde{\mu}_0$ cuando la población es continua, no obstante, en la práctica ocurrirá con frecuencia un valor muestral igual a $\tilde{\mu}_0$ debido a una falta de precisión al registrar los datos. Cuando se observan valores muestrales iguales a $\tilde{\mu}_0$ se excluyen del análisis y, en consecuencia, se reduce el tamaño de la muestra.

El estadístico de prueba adecuado para la prueba de signo es la variable aleatoria binomial X, que representa el número de signos más en nuestra muestra aleatoria. Si la hipótesis nula de que $\tilde{\mu} = \tilde{\mu}_0$ es verdadera, la probabilidad de que un valor mues-

tral tenga como resultado un signo más o uno menos es igual a 1/2. Por lo tanto, para probar la hipótesis nula de que $\tilde{\mu} = \tilde{\mu}_0$, en realidad probamos la hipótesis nula de que el número de signos más es un valor de una variable aleatoria que tiene la distribución binomial con el parámetro $p = 1/2$. Los valores P para las alternativas unilateral y bilateral se pueden calcular entonces con el uso de esta distribución binomial. Por ejemplo, al probar

$$\begin{aligned} H_0&: \ \tilde{\mu} = \tilde{\mu}_0, \\ H_1&: \ \tilde{\mu} < \tilde{\mu}_0, \end{aligned}$$

rechazaremos H_0 a favor de H_1 sólo si la proporción de signos más es bastante menor que 1/2; es decir, cuando el valor x de nuestra variable aleatoria es pequeño. Por lo tanto, si el valor P que se calcula

$$P = P(X \leq x \text{ cuando } p = 1/2)$$

es menor que o igual a algún nivel de significancia α preestablecido, rechazamos H_0 a favor de H_1. Por ejemplo, cuando $n = 15$ y $x = 3$, encontramos de la tabla A.1 que

$$P = P(X \leq 3 \text{ cuando } p = 1/2) = \sum_{x=0}^{3} b\left(x; 15, \frac{1}{2}\right) = 0.0176,$$

por lo que la hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$ se puede rechazar en realidad en el nivel de significancia de 0.05 pero no en el nivel 0.01.

Para probar la hipótesis

$$\begin{aligned} H_0&: \ \tilde{\mu} = \tilde{\mu}_0, \\ H_1&: \ \tilde{\mu} > \tilde{\mu}_0, \end{aligned}$$

rechazamos H_0 a favor de H_1 sólo si la proporción de signos más es bastante mayor que 1/2; es decir, cuando x es grande. De aquí, si el valor P calculado

$$P = P(X \geq x \text{ cuando } p = 1/2)$$

es menor que α, rechazamos H_0 a favor de H_1. Finalmente, para probar la hipótesis

$$\begin{aligned} H_0&: \ \tilde{\mu} = \tilde{\mu}_0, \\ H_1&: \ \tilde{\mu} \neq \tilde{\mu}_0, \end{aligned}$$

rechazamos H_0 a favor de H_1 cuando la proporción de signos más es significativamente menor o mayor que 1/2. Esto, por supuesto, es equivalente a que x sea bastante pequeña o bastante grande. Por lo tanto, si $x < n/2$ y el valor P calculado

$$P = 2P(X \leq x \text{ cuando } p = 1/2)$$

es menor o igual que α, o si $x > n/2$ y el valor P calculado

$$P = 2P(X \geq x \text{ cuando } p = 1/2)$$

es menor o igual que α, rechazamos H_0 a favor de H_1.

Siempre que $n > 10$, las probabilidades binomiales con $p = 1/2$ se pueden aproximar a partir de la curva normal, ya que $np = nq > 5$. Suponga, por ejemplo, que deseamos probar la hipótesis

$$H_0\text{: } \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\text{: } \tilde{\mu} < \tilde{\mu}_0,$$

en el nivel de significancia $\alpha = 0.05$ para una muestra aleatoria de tamaño $n = 20$ que produce $x = 6$ signos más. Utilizando la aproximación de la curva normal con

$$\tilde{\mu} = np = (20)(0.5) = 10$$

y

$$\sigma = \sqrt{npq} = \sqrt{(20)(0.5)(0.5)} = 2.236,$$

encontramos que

$$z = \frac{6.5 - 10}{2.236} = -1.57.$$

Por lo tanto,

$$P = P(X \leq 6) \approx P(Z < -1.57) = 0.0582,$$

que conduce a no rechazar la hipótesis nula.

Ejemplo 16.1: Los siguientes datos representan el número de horas que un compensador opera antes de que requiera una recarga:

$$1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, 1.7.$$

Utilice la prueba de signo para probar la hipótesis en el nivel de significancia de 0.05 de que este compensador específico opera con una mediana de 1.8 horas antes de requerir una recarga.

Solución:

1. H_0: $\tilde{\mu} = 1.8$.

2. H_1: $\tilde{\mu} \neq 1.8$.

3. $\alpha = 0.05$.

4. Estadística de prueba: variable binomial X con $p = \frac{1}{2}$.

5. Cálculos: Al reemplazar cada valor con el símbolo "+" si excede 1.8, con el símbolo "−" si es menor que 1.8 y descartar las mediciones que sean iguales a 1.8, obtenemos la siguiente secuencia

$$- \quad + \quad - \quad - \quad + \quad - \quad - \quad + \quad - \quad -$$

para la que $n = 10$, $x = 3$ y $n/2 = 5$. Por lo tanto, de la tabla A.1 el valor P que se calcula es

$$P = 2P(X \leq 3 \text{ cuando } p = \frac{1}{2}) = 2\sum_{x=0}^{3} b(x; 10, \frac{1}{2}) = 0.3438 > 0.05.$$

6. Decisión: no rechace la hipótesis nula y concluya que la mediana del tiempo de operación no es significativamente diferente de 1.8 horas.

También se puede utilizar la prueba de signo para probar la hipótesis nula $\tilde{\mu}_1 - \tilde{\mu}_2 = d_0$ para observaciones pareadas. Aquí reemplazamos cada diferencia, d_i, con un signo más o un signo menos, dependiendo si la diferencia ajustada, $d_i - d_0$, es positiva o negativa. A lo largo de esta sección suponemos que las poblaciones son simétricas. No obstante, aun si las poblaciones fueran asimétricas podríamos llevar a cabo el mismo procedimiento de prueba, pero las hipótesis se refieren a las medianas poblacionales en vez de a las medias.

Ejemplo 16.2: Una compañía de taxis intenta decidir si el uso de llantas radiales en vez de llantas regulares con cinturón mejora la economía del combustible. Se equipan 16 automóviles con llantas radiales y se manejan por un recorrido de prueba establecido. Sin cambiar de conductores, se equipan los mismos autos con las llantas regulares con cinturón y se manejan una vez más por el recorrido de prueba. El consumo de gasolina, en kilómetros por litro, se presenta en la tabla 16.1.

¿Con el nivel de significancia de 0.05 podemos concluir que los automóviles equipados con llantas radiales obtienen mejores economías de combustible, que los equipados con llantas regulares con cinturón?

Tabla 16.1: Datos para el ejemplo 16.2

Automóvil	1	2	3	4	5	6	7	8
Llantas radiales	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0
Llantas con cinturón	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8
Automóvil	9	10	11	12	13	14	15	16
Llantas radiales	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
Llantas con cinturón	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8

Solución: Sean $\tilde{\mu}_1$ y $\tilde{\mu}_2$ los kilómetros por litro promedio para los automóviles equipados con llantas radiales y con cinturón, respectivamente.

1. H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 = 0$.

2. H_1: $\tilde{\mu}_1 - \tilde{\mu}_2 > 0$.

3. $\alpha = 0.05$.

4. Estadístico de prueba: variable binomial X con $p = 1/2$.

5. Cálculos: Después de reemplazar cada diferencia positiva con un símbolo "+" y cada diferencia negativa con un símbolo "−", y después descartar las dos diferencias cero, obtenemos la secuencia

$$+ \; - \; + \; + \; - \; + \; + \; + \; + \; + \; + \; + \; - \; +$$

para la que $n = 14$ y $x = 11$. Con la aproximación de la curva normal, encontramos que

$$z = \frac{10.5 - 7}{\sqrt{(14)(0.5)(0.5)}} = 1.87,$$

y entonces

$$P = P(X \geq 11) \approx P(Z > 1.87) = 0.0307.$$

6. Decisión: Rechace H_0 y concluya que, en promedio, las llantas radiales mejoran la economía de combustible.

La prueba de signo no sólo es uno de nuestros procedimientos no paramétricos más fáciles de aplicar, pues tiene la ventaja adicional de ser aplicable a datos dicotómicos que no se pueden registrar en una escala numérica, pero que se pueden representar mediante respuestas positivas y negativas. Por ejemplo, la prueba de signo se aplica en experimentos donde se registra una respuesta cualitativa como "éxito" o "fracaso"; y en experimentos de tipo sensorial donde se registra un signo más o un signo menos, dependiendo de si el catador del sabor identifica de manera correcta o incorrecta el ingrediente que se desea.

Intentaremos realizar comparaciones entre varios de los procedimientos no paramétricos y las pruebas paramétricas correspondientes. En el caso de la prueba de signo la competencia es, desde luego, la prueba t. Si se muestrea de una distribución normal, el uso de la prueba t tendrá como resultado la potencia más grande de la prueba. Si la distribución sólo es simétrica, aunque no sea normal, se prefiere la prueba t en términos de potencia, a menos que la distribución tenga "colas muy pesadas" en comparación con la distribución normal.

16.2 Prueba de rango con signo

El lector debería notar que la prueba de signo tan sólo utiliza los signos más y menos de las diferencias entre las observaciones y $\tilde{\mu}_0$ en el caso de una muestra, o los signos más y menos de las diferencias entre los pares de observaciones en el caso de la muestra pareada; aunque no toma en consideración la magnitud de tales diferencias. Una prueba que utiliza dirección y magnitud, propuesta en 1945 por Frank Wilcoxon, se llama ahora comúnmente **prueba de rango con signo de Wilcoxon**.

El analista podría extraer más información de los datos en una forma no paramétrica, si resulta razonable aplicar una restricción adicional a la distribución de la que se toman los datos. La prueba de rango con signo de Wilcoxon se aplica en el caso de una **distribución continua simétrica**. Bajo esta condición se prueba la hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$. Primero restamos $\tilde{\mu}_0$ de cada valor muestral y descartamos todas las diferencias iguales a cero. Se clasifican entonces las diferencias restantes sin importar el signo. Se asigna un rango de 1 a la menor diferencia absoluta (es decir, sin signo), un rango de 2 a la siguiente más pequeña, y así sucesivamente. Cuando el valor absoluto de dos o más diferencias es el mismo, se asigna a cada uno el promedio de los rangos que se asignarían si las diferencias fueran distinguibles. Por ejemplo, si las diferencias quinta y sexta más pequeñas son iguales en valor absoluto, a cada una se le asignaría un rango de 5.5. Si la hipótesis $\tilde{\mu} = \tilde{\mu}_0$ es verdadera, el total de los rangos que corresponden a las diferencias positivas debería ser casi igual al total de los rangos que corresponden a las diferencias negativas. Representemos estos totales con w_+ y w_-, respectivamente. Designamos el menor de w_+ y w_- con w.

Al seleccionar muestras repetidas esperaríamos que variaran w_+ y w_- y, por lo tanto, w. De esta manera consideramos w_+, w_- y w como valores de las correspondientes variables aleatorias W_+, W_- y W. La hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$ se puede rechazar a favor de la alternativa $\tilde{\mu} < \tilde{\mu}_0$ sólo si w_+ es pequeña y w_- es grande. Asimismo, la alternativa $\tilde{\mu} > \tilde{\mu}_0$ se puede aceptar sólo si w_+ es grande y w_- es pequeña. Para una alternativa bilateral podemos rechazar H_0 a favor de H_1 si w_+ o w_- y, por lo tanto, w son suficientemente pequeñas. De esta manera, no importa cuál hipótesis

alternativa sea, rechazamos la hipótesis nula cuando el valor del estadístico adecuado W_+, W_- o W es suficientemente pequeño.

Dos muestras con observaciones pareadas

Para probar la hipótesis nula de que se muestrean dos poblaciones simétricas continuas con $\tilde{\mu}_1 = \tilde{\mu}_2$ para el caso de una muestra pareada, clasificamos las diferencias de las observaciones pareadas sin importar el signo y procedemos como en el caso de una sola muestra. Los diversos procedimientos de prueba para los casos de una sola muestra y de una muestra pareada, se resumen en la tabla 16.2.

Tabla 16.2: Prueba de rango con signo

H_0	H_1	Calcule
$\tilde{\mu} = \tilde{\mu}_0$	$\tilde{\mu} < \tilde{\mu}_0$	w_+
	$\tilde{\mu} > \tilde{\mu}_0$	w_-
	$\tilde{\mu} \neq \tilde{\mu}_0$	w
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	w_+
	$\tilde{\mu}_1 > \tilde{\mu}_2$	w_-
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	w

No es difícil mostrar que siempre que $n < 5$ y el nivel de significancia no exceda 0.05 para una prueba de una cola, o 0.10 para una prueba de dos colas, todos los valores posibles de w_+, w_- o w conducirán a la aceptación de la hipótesis nula. Sin embargo, cuando $5 \leq n \leq 30$, la tabla A.17 muestra valores críticos aproximados de W_+ y W_- para niveles de significancia iguales a 0.01, 0.025 y 0.05 para una prueba de una cola, y valores críticos de W para niveles de significancia iguales a 0.02, 0.05 y 0.10 para una prueba de dos colas. Se rechaza la hipótesis nula si el valor calculado w_+, w_- o w es **menor o igual que** el valor tabulado apropiado. Por ejemplo, cuando $n = 12$, la tabla A.17 muestra que se requiere un valor de $w_+ \leq 17$ para que la alternativa unilateral $\tilde{\mu} < \tilde{\mu}_0$ sea significativa en el nivel 0.05.

Ejemplo 16.3: Repita el ejemplo 16.1 usando la prueba de rango con signo.

Solución:

1. H_0: $\tilde{\mu} = 1.8$.

2. H_1: $\tilde{\mu} \neq 1.8$.

3. $\alpha = 0.05$.

4. Región crítica: Como $n = 10$, después de descartar la medición que es igual a 1.8, la tabla A.17 muestra que la región crítica es $w \leq 8$.

5. Cálculos: Al restar 1.8 de cada medición y después clasificar las diferencias sin hacer caso del signo, tenemos

d_i	−0.3	0.4	−0.9	−0.5	0.2	−0.2	−0.3	0.2	−0.6	−0.1
Rangos	5.5	7	10	8	3	3	5.5	3	9	1

Ahora bien, $w_+ = 13$ y $w_- = 42$, de manera que $w = 13$, el menor de w_+ y w_-.

6. Decisión: Como antes, no rechace H_0 y concluya que el tiempo promedio de operación no es significativamente diferente de 1.8 horas.

La prueba de rango con signo también se puede utilizar para probar la hipótesis nula de que $\tilde{\mu}_1 - \tilde{\mu}_2 = d_0$. En este caso las poblaciones no necesitan ser simétricas. Como con la prueba de signo, restamos d_0 de cada diferencia, clasificamos las diferencias ajustadas sin importar el signo y aplicamos el mismo procedimiento anterior.

Ejemplo 16.4: Se afirma que un estudiante universitario de último año puede aumentar su calificación en el área del campo de especialidad del examen de registro de graduados en al menos 50 puntos, si de antemano se le ofrecen problemas de muestra. Para probar esta afirmación, se dividen 20 estudiantes del último año en 10 pares, de manera que cada par tenga casi el mismo promedio de puntos de calidad general en sus 3 primeros años en la universidad. Los problemas y respuestas de muestra se proporcionan al azar a un miembro de cada par 1 semana antes del examen. Las calificaciones del examen se presentan en la tabla 16.3:

Tabla 16.3: Datos para el ejemplo 16.4

	Par									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Con problemas de muestra	531	621	663	579	451	660	591	719	543	575
Sin problemas de muestra	509	540	688	502	424	683	568	748	530	524

Pruebe la hipótesis nula en el nivel de significancia de 0.05 de que los problemas de muestra aumentan las calificaciones en 50 puntos, contra la hipótesis alternativa de que el aumento es menor a 50 puntos.

Solución: Representemos con $\tilde{\mu}_1$ y $\tilde{\mu}_2$ la calificación media de todos los estudiantes que resuelven el examen en cuestión con y sin problemas de muestra, respectivamente.

1. H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 = 50$.

2. H_1: $\tilde{\mu}_1 - \tilde{\mu}_2 < 50$.

3. $\alpha = 0.05$.

4. Región crítica: Como $n = 10$, la tabla A.17 muestra que la región crítica es $w_+ \leq 11$.

5. Cálculos:

	Par									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
d_i	22	81	−25	77	27	−23	23	−29	13	51
$d_i - d_0$	−28	31	−75	27	−23	−73	−27	−79	−37	1
Rangos	5	6	9	3.5	2	8	3.5	10	7	1

Encontramos ahora que $w_+ = 6 + 3.5 + 1 = 10.5$.

6. Decisión: Rechace H_0 y concluya que los problemas de muestra, "en promedio", no aumentan las calificaciones de registro de graduados en 50 puntos.

Aproximación normal para muestras grandes

Cuando $n \geq 15$, la distribución muestral de W_+ (o W_-) se aproxima a la distribución normal con media

$$\mu_{W_+} = \frac{n(n+1)}{4} \qquad \text{y varianza} \qquad \sigma^2_{W_+} = \frac{n(n+1)(2n+1)}{24}.$$

Por lo tanto, cuando n excede el valor más grande en la tabla A.17, se utiliza el estadístico

$$Z = \frac{W_+ - \mu_{W_+}}{\sigma_{W_+}}$$

para determinar la región crítica para nuestra prueba.

Ejercicios

16.1 Los siguientes datos representan el tiempo, en minutos, que un paciente tiene que esperar durante 12 visitas al consultorio de una doctora antes de ser atendido por ésta:

17	15	20	20	32	28
12	26	25	25	35	24

Utilice la prueba de signo al nivel de significancia de 0.05 para probar la afirmación de la doctora, de que la media del tiempo de espera para sus pacientes no es mayor que 20 minutos antes de entrar al consultorio.

16.2 Los siguientes datos representan el número de horas de vuelo de entrenamiento que reciben 18 estudiantes para piloto, de cierto instructor, antes de su primer vuelo solos:

9	12	18	14	12	14	12	10	16
11	9	11	13	11	13	15	13	14

Con las probabilidades binomiales de la tabla A.1, realice una prueba de signo al nivel de significancia de 0.02 para probar la afirmación del instructor, de que la mediana del tiempo que se requiere antes de que sus estudiantes vuelen solos es 12 horas de vuelo de entrenamiento.

16.3 Un inspector de alimentos examina 16 latas de cierta marca de jamón para determinar el porcentaje de impurezas externas. Se registraron los siguientes datos:

2.4	2.3	3.1	2.2	2.3	1.2	1.0	2.4
1.7	1.1	4.2	1.9	1.7	3.6	1.6	2.3

Con la aproximación normal a la distribución binomial, realice una prueba de signo al nivel de significancia de 0.05, para probar la hipótesis nula de que la mediana del porcentaje de impurezas en esta marca de jamón es 2.5%, contra la alternativa de que la mediana del porcentaje de impurezas no es 2.5%.

16.4 Un proveedor de pintura afirma que un nuevo aditivo reducirá el tiempo de secado de su pintura acrílica. Para probar esta afirmación, se pintaron 12 paneles de madera: una mitad de cada panel con pintura que contiene el aditivo regular, y la otra con pintura que contiene el nuevo aditivo. Los tiempos de secado, en horas, se registran a continuación:

	Tiempo de secado (horas)	
Panel	**Nuevo aditivo**	**Aditivo regular**
1	6.4	6.6
2	5.8	5.8
3	7.4	7.8
4	5.5	5.7
5	6.3	6.0
6	7.8	8.4
7	8.6	8.8
8	8.2	8.4
9	7.0	7.3
10	4.9	5.8
11	5.9	5.8
12	6.5	6.5

Utilice la prueba de signo en el nivel 0.05 para probar la hipótesis nula de que el nuevo aditivo no es mejor que el aditivo regular para reducir el tiempo de secado de este tipo de pintura.

16.5 Se afirma que una nueva dieta reducirá 4.5 kilogramos el peso de una persona, en promedio, en un periodo de 2 semanas. Se registran los pesos de 10 mujeres que siguen esta dieta, antes y después de un periodo de 2 semanas, y se obtienen los siguientes datos:

Mujer	**Peso antes**	**Peso después**
1	58.5	60.0
2	60.3	54.9
3	61.7	58.1
4	69.0	62.1
5	64.0	58.5
6	62.6	59.9
7	56.7	54.4
8	63.6	60.2
9	68.2	62.3
10	59.4	58.7

Utilice la prueba de signo al nivel de significancia de 0.05 para probar la hipótesis de que la dieta reduce la mediana del peso en 4.5 kilogramos, contra la hipótesis alternativa de que la mediana de la diferencia en pesos es menor que 4.5 kilogramos.

16.6 Se comparan dos tipos de instrumentos para medir la cantidad de monóxido de azufre en la atmósfera en un experimento de contaminación atmosférica. Se registraron las siguientes lecturas diarias en un periodo de 2 semanas:

	Monóxido de azufre	
Día	**Instrumento *A***	**Instrumento *B***
1	0.96	0.87
2	0.82	0.74
3	0.75	0.63
4	0.61	0.55
5	0.89	0.76
6	0.64	0.70
7	0.81	0.69
8	0.68	0.57
9	0.65	0.53
10	0.84	0.88
11	0.59	0.51
12	0.94	0.79
13	0.91	0.84
14	0.77	0.63

Usando la aproximación normal a la distribución binomial, realice una prueba de signo para determinar si los diferentes instrumentos conducen a diferentes resultados. Utilice un nivel de significancia de 0.05.

16.7 Las siguientes cifras indican la presión sanguínea sistólica de 16 corredores antes y después de una carrera de 8 kilómetros:

Corredor	**Antes**	**Después**
1	158	164
2	149	158
3	160	163
4	155	160
5	164	172
6	138	147
7	163	167
8	159	169
9	165	173
10	145	147
11	150	156
12	161	164
13	132	133
14	155	161
15	146	154
16	159	170

Utilice una prueba de signo al nivel de significancia de 0.05 para probar la hipótesis nula de que correr 8 kilómetros aumenta la mediana de la presión sanguínea sistólica en 8 puntos contra la alternativa de que el aumento en la mediana es menor que 8 puntos.

16.8 Analice los datos del ejercicio 16.1 usando la prueba de rango con signo.

16.9 Analice los datos del ejercicio 16.2 usando la prueba de rango con signo.

16.10 Los pesos de 5 personas antes de que dejen de fumar y cinco semanas después de dejar de fumar, en kilogramos, son los siguientes:

	Individual				
	1	**2**	**3**	**4**	**5**
Antes	66	80	69	52	75
Después	71	82	68	56	73

Utilice la prueba de rango con signo para observaciones pareadas para probar la hipótesis, en el nivel de significancia de 0.05, de que dejar de fumar no tiene efecto en el peso de una persona, contra la alternativa de que el peso aumenta si se deja de fumar.

16.11 Repita el ejercicio 16.5 usando la prueba de rango con signo.

16.12 Los siguientes son los números de recetas surtidas por dos farmacias en un periodo de 20 días:

Día	**Farmacia *A***	**Farmacia *B***
1	19	17
2	21	15
3	15	12
4	17	12
5	24	16
6	12	15
7	19	11
8	14	13
9	20	14
10	18	21
11	23	19
12	21	15
13	17	11
14	12	10
15	16	20
16	15	12
17	20	13
18	18	17
19	14	16
20	22	18

Utilice la prueba de rango con signo al nivel de significancia de 0.01 para determinar si las dos farmacias, "en promedio", surten el mismo número de recetas, contra la alternativa de que la farmacia A surte más recetas que la farmacia B.

16.13 Repita el ejercicio 16.7 con la prueba de rango con signo.

16.14 Repita el ejercicio 16.6 con la prueba de rango con signo.

16.3 Prueba de la suma de rangos de Wilcoxon

Como indicamos antes, el procedimiento no paramétrico, por lo general, es una alternativa adecuada para la prueba de la teoría normal cuando no es válida la suposición de normalidad. Cuando interesa probar la igualdad de las medias de dos distribuciones continuas que evidentemente no son normales, y las muestras son independientes (es decir, no hay pareamiento de observaciones), la **prueba de la suma de rangos de Wilcoxon** o **prueba de dos muestras de Wilcoxon** es una alternativa apropiada a la prueba t de dos muestras que se describe en el capítulo 10.

Probaremos la hipótesis nula H_0 de que $\tilde{\mu}_1 = \tilde{\mu}_2$ contra alguna alternativa adecuada. Primero seleccionamos una muestra aleatoria de cada una de las poblaciones. Sea n_1 el número de observaciones en la muestra más pequeña y n_2 el número de observaciones en la muestra más grande. Cuando las muestras son de igual tamaño, n_1 y n_2 se pueden asignar de manera aleatoria. Hay que ordenar las $n_1 + n_2$ observaciones de las muestras combinadas en orden ascendente y sustituir un rango de $1, 2, \ldots, n_1 + n_2$ para cada observación. En el caso de empates (observaciones idénticas), reemplazamos las observaciones por la media de los rangos que tendrían las observaciones si fueran distinguibles. Por ejemplo, si la séptima y octava observaciones son idénticas, asignaríamos un rango de 7.5 a cada una de las dos observaciones.

La suma de los rangos que corresponden a las n_i observaciones en la muestra más pequeña se denota con w_1. De manera similar, el valor w_2 representa la suma de los n_2 rangos que corresponden a la muestra más grande. El total $w_1 + w_2$ depende sólo del número de observaciones en las dos muestras y de ninguna manera resulta afectado por los resultados del experimento. De aquí, si $n_1 = 3$ y $n_2 = 4$, entonces $w_1 + w_2 = 1 + 2 + \cdots + 7 = 28$, sin importar los valores numéricos de las observaciones. En general,

$$w_1 + w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2},$$

la suma aritmética de los enteros $1, 2, \ldots, n_1 + n_2$. Una vez que se determina w_1, puede ser más fácil encontrar w_2 mediante la fórmula

$$w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - w_1.$$

Al elegir muestras repetidas de tamaños n_1 y n_2, esperaríamos que variaran w_1 y, por lo tanto, w_2. Consideramos entonces w_1 y w_2 como valores de las variables aleatorias W_1 y W_2, respectivamente. La hipótesis nula $\tilde{\mu}_1 = \tilde{\mu}_2$ se rechazará a favor de la alternativa $\tilde{\mu}_1 < \tilde{\mu}_2$ sólo si w_1 es pequeña y w_2 es grande. Asimimo, la alternativa $\tilde{\mu}_1 > \tilde{\mu}_2$ se puede aceptar sólo si w_1 es grande y w_2 es pequeña. Para una prueba de dos colas, podemos rechazar H_0 a favor de H_1 si w_1 es pequeña y w_2 es grande, o si w_1 es grande y w_2 es pequeña. En otras palabras, se acepta la alternativa $\tilde{\mu}_1 < \tilde{\mu}_2$ si w_1 es suficientemente pequeña; la alternativa $\tilde{\mu}_1 > \tilde{\mu}_2$ se acepta si w_2 es suficientemente pequeña; y la alternativa $\tilde{\mu}_1 \neq \tilde{\mu}_2$ se acepta si el mínimo de w_1 y w_2 es suficientemente pequeño. En la práctica real, por lo general, basamos nuestra decisión en el valor

$$u_1 = w_1 - \frac{n_1(n_1 + 1)}{2} \qquad \text{o} \qquad u_2 = w_2 - \frac{n_2(n_2 + 1)}{2}$$

del estadístico relacionado U_1 o U_2, o en el valor u del estadístico U, el mínimo de U_1 y U_2. Dichos estadísticos simplifican la construcción de tablas de valores críticos,

pues U_1 y U_2 tienen distribuciones muestrales simétricas y toman valores en el intervalo de 0 a n_1n_2 tales que $u_1 + u_2 = n_1n_2$.

De las fórmulas para u_1 y u_2 vemos que u_1 será pequeña cuando w_1 es pequeña, y u_2 será pequeña cuando w_2 sea pequeña. En consecuencia, la hipótesis nula se rechazará siempre que los estadísticos apropiados U_1, U_2 o U tomen un valor menor o igual que el valor crítico deseado dado en la tabla A.18. Los diversos procedimientos de prueba se resumen en la tabla 16.4.

Tabla 16.4: Prueba de la suma de rangos

H_0	H_1	Calcule
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	u_1
	$\tilde{\mu}_1 > \tilde{\mu}_2$	u_2
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	u

La tabla A.18 da valores críticos de U_1 y U_2 para niveles de significancia iguales a 0.001, 0.002, 0.01, 0.02, 0.025 y 0.05 para una prueba de una sola cola, y valores críticos de U para niveles de significancia iguales a 0.002, 0.02, 0.05 y 0.10 para una prueba de dos colas. Si el valor observado de u_1, u_2 o u es **menor o igual que** el valor crítico tabulado, se rechaza la hipótesis nula en el nivel de significancia que se indica en la tabla. Suponga, por ejemplo, que deseamos probar la hipótesis nula de que $\tilde{\mu}_1 = \tilde{\mu}_2$ contra la alternativa unilateral de que $\tilde{\mu}_1 < \tilde{\mu}_2$ en el nivel de significancia 0.05 para muestras aleatorias de tamaño $n_1 = 3$ y $n_2 = 5$, que dan el valor $w_1 = 8$. Se sigue que

$$u_1 = 8 - \frac{(3)(4)}{2} = 2.$$

Nuestra prueba de una sola cola se basa en el estadístico U_1. Con la tabla A.18, rechazamos la hipótesis nula de medias iguales cuando $u_1 \leq 1$. Como $u_1 = 2$ no cae en la región de aceptación, no se puede rechazar la hipótesis nula.

Ejemplo 16.5: Se encuentra que el contenido de nicotina de dos marcas de cigarrillos, medido en miligramos, es el siguiente:

Marca A	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
Marca B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

Pruebe la hipótesis, en el nivel de significancia de 0.05, de que el contenido promedio de nicotina de las dos marcas es igual, contra la alternativa de que son diferentes.

Solución:

1. H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$.

2. H_1: $\tilde{\mu}_1 \neq \tilde{\mu}_2$.

3. $\alpha = 0.05$.

4. Región crítica: $u \leq 17$ (de la tabla A.18).

5. Cálculos: Las observaciones se acomodan en orden ascendente y se les asignan rangos del 1 al 18.

Datos originales	Rangos	Datos originales	Rangos
0.6	1	4.0	10.5*
1.6	2	4.0	10.5
1.9	3	4.1	12
2.1	4*	4.8	13*
2.2	5	5.4	14.5*
2.5	6	5.4	14.5
3.1	7	6.1	16*
3.3	8*	6.2	17
3.7	9*	6.3	18*

*Los rangos con asterisco pertenecen a la muestra A.

Ahora

$$w_1 = 4 + 8 + 9 + 10.5 + 13 + 14.5 + 16 + 18 = 93,$$

y

$$w_2 = \frac{(18)(19)}{2} - 93 = 78.$$

Por lo tanto,

$$u_1 = 93 - \frac{(8)(9)}{2} = 57, \qquad u_2 = 78 - \frac{(10)(11)}{2} = 23.$$

6. Decisión: No rechace la hipótesis nula H_0 y concluya que no hay diferencia significativa en el contenido promedio de nicotina en las dos marcas de cigarrillos.

Teoría normal de aproximación para dos muestras

Cuando n_1 y n_2 exceden 8, la distribución muestral de U_1 (o U_2) se aproxima a la distribución normal con media

$$\mu_{U_1} = \frac{n_1 n_2}{2} \qquad \text{y varianza} \qquad \sigma^2_{U_1} = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

En consecuencia, cuando n_2 es mayor que 20, el valor máximo en la tabla A.18, y n_1 es al menos 9, se puede utilizar el estadístico

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

para nuestra prueba, con la región crítica que cae ya sea en alguna o en ambas colas de la distribución normal estándar, dependiendo de la forma de H_1.

El uso de la prueba de suma de rangos de Wilcoxon no se restringe a poblaciones no normales. Se puede utilizar en vez de la prueba t de dos muestras cuando las poblaciones son normales, aunque la potencia será menor. La prueba de suma de rangos de Wilcoxon siempre es superior a la prueba t para poblaciones decididamente no normales.

16.4 Prueba de Kruskal-Wallis

En los capítulos 13, 14 y 15, la técnica de análisis de varianza resalta como técnica analítica para probar la igualdad de $k \geq 2$ medias poblacionales. De nuevo, sin embargo, el lector debería recordar que se debe suponer la normalidad, con la finalidad de que la prueba F sea teóricamente correcta. En esta sección investigamos una alternativa no paramétrica al análisis de varianza.

La **prueba de Kruskal-Wallis**, también llamada **prueba H de Kruskal-Wallis**, es una generalización de la prueba de la suma de rangos para el caso de $k > 2$ muestras. Se utiliza para probar la hipótesis nula H_0 de que k muestras independientes provienen de poblaciones idénticas. Presentada en 1952 por W. H. Kruskal y W. A. Wallis, la prueba constituye un procedimiento no paramétrico para probar la igualdad de las medias, en el análisis de varianza de un factor, cuando el experimentador desea evitar la suposición de que las muestras se seleccionaron de poblaciones normales.

Sea n_i $(i = 1, 2, \ldots, k)$ el número de observaciones en la i-ésima muestra. Primero, combinamos todas las k muestras y acomodamos las $n = n_1 + n_2 + \cdots + n_k$ observaciones en orden ascendente, y sustituimos el rango apropiado de $1, 2, \ldots, n$ para cada observación. En el caso de empates (observaciones idénticas), seguimos el procedimiento acostumbrado de reemplazar las observaciones por las medias de los rangos que tendrían las observaciones si fueran distinguibles. La suma de los rangos que corresponde a las n_i observaciones en la i-ésima muestra se denota mediante la variable aleatoria R_i. Consideremos ahora el estadístico

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1),$$

que se aproxima muy bien mediante una distribución chi cuadrada con $k - 1$ grados de libertad, cuando H_0 es verdadera y si cada muestra consiste en al menos 5 observaciones. El hecho de que h, el supuesto valor de H, sea grande cuando las muestras independientes provienen de poblaciones que no son idénticas nos permite establecer el siguiente criterio de decisión para probar H_0:

Prueba de Kruskal-Wallis

Para probar la hipótesis nula H_0 de que k muestras independientes provienen de poblaciones idénticas, calcule

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1),$$

donde r_i es el valor supuesto de R_i, para $i = 1, 2, \ldots, k$. Si h cae en la región crítica $H > \chi_\alpha^2$ con $v = k - 1$ grados de libertad, rechace H_0 con el nivel de significancia α; de otra manera, no rechace H_0.

Ejemplo 16.6: En un experimento para determinar cuál de tres diferentes sistemas de misiles es preferible, se mide la tasa de utilización del propulsor. Los datos, después de codificarlos, se presentan en la tabla 16.5. Utilice la prueba de Kruskal-Wallis y un nivel de significancia $\alpha = 0.05$, para probar la hipótesis de que las tasas de utilización del propulsor son las mismas para los tres sistemas de misiles.

Tabla 16.5: Tasas de utilización del propulsor

Sistema de misiles								
1			**2**			**3**		
24.0	16.7	22.8	23.2	19.8	18.1	18.4	19.1	17.3
19.8	18.9		17.6	20.2	17.8	17.3	19.7	18.9
						18.8	19.3	

Solución: **1.** H_0: $\mu_1 = \mu_2 = \mu_3$.

2. H_1: las tres medias son diferentes.

3. $\alpha = 0.05$.

4. Región crítica: $h > \chi^2_{0.05} = 5.991$, para $v = 2$ grados de libertad.

5. Cálculos: En la tabla 16.6 convertimos las 19 observaciones a rangos y sumamos los rangos para cada sistema de misiles.

Tabla 16.6: Rangos para las tasas de utilización del propulsor

Sistema de misiles		
1	**2**	**3**
19	18	7
1	14.5	11
17	6	2.5
14.5	4	2.5
9.5	16	13
$r_1 = 61.0$	5	9.5
	$r_2 = 63.5$	8
		12
		$r_3 = 65.5$

Ahora, al sustituir $n_1 = 5$, $n_2 = 6$, $n_3 = 8$ y $r_1 = 61.0$, $r_2 = 63.5$, $r_3 = 65.5$, nuestro estadístico de prueba H toma el valor

$$h = \frac{12}{(19)(20)}\left(\frac{61.0^2}{5} + \frac{63.5^2}{6} + \frac{65.5^2}{8}\right) - (3)(20) = 1.66.$$

6. Decisión: Como $h = 1.66$ no cae en la región crítica $h > 5.991$, tenemos insuficiente evidencia para rechazar la hipótesis de que las tasas de utilización del propulsor son las mismas para los tres sistemas de misiles.

Ejercicios

16.15 Un fabricante de cigarrillos afirma que el contenido de alquitrán de la marca de cigarrillos B es menor que la de la marca A. Para probar esta afirmación, se registran las siguientes determinaciones de contenido de alquitrán, en miligramos:

Marca A	1	12	9	13	11	14
Marca B	8	10	7			

Utilice la prueba de suma de rangos con $\alpha = 0.05$ para probar si tal afirmación es válida.

16.16 Para averiguar si un nuevo suero detendrá la leucemia, se seleccionan 9 pacientes, quienes ya alcanzaron una etapa avanzada de la enfermedad. Cinco pacientes reciben el tratamiento y cuatro no. Los tiempos de supervivencia, en años, a partir del momento en que comienza el experimento son

Con tratamiento	2.1	5.3	1.4	4.6	0.9
Sin tratamiento	1.9	0.5	2.8	3.1	

Utilice la prueba de suma de rangos, en el nivel de significancia de 0.05, para determinar si el suero es eficaz.

16.17 Los siguientes datos representan el número de horas que operan dos diferentes tipos de calculadoras científicas de bolsillo, antes de que necesiten recargarse.

Calculadora A	5.5	5.6	6.3	4.6	5.3	5.0	6.2	5.8	5.1
Calculadora B	3.8	4.8	4.3	4.2	4.0	4.9	4.5	5.2	4.5

Utilice la prueba de la suma de rangos con $\alpha = 0.01$ para determinar si la calculadora A opera más tiempo que la calculadora B con una carga completa de la batería.

16.18 Se fabrica un hilo para pesca usando dos procesos. Para determinar si hay una diferencia en la resistencia media a la rotura de los hilos, se seleccionan 10 piezas de cada proceso y después se prueba dicha resistencia. Los resultados son los siguientes:

Proceso 1	10.4	9.8	11.5	10.0	9.9
	9.6	10.9	11.8	9.3	10.7
Proceso 2	8.7	11.2	9.8	10.1	10.8
	9.5	11.0	9.8	10.5	9.9

Utilice la prueba de suma de rangos con $\alpha = 0.1$ para determinar si hay una diferencia entre las resistencias medias a la rotura de los hilos fabricados por los dos procesos.

16.19 De una clase de matemáticas de 12 estudiantes con capacidades iguales, quienes utilizan material programado, se seleccionan 5 al azar y el profesor les da instrucción adicional. Los resultados del examen final son los siguientes:

	Calificación						
Con instrucción adicional	87	69	78	91	80		
Sin instrucción adicional	75	88	64	82	93	79	67

Utilice la prueba de la suma de rangos con $\alpha = 0.05$ para determinar si la instrucción adicional afecta la calificación promedio.

16.20 Los siguientes datos representan los pesos, en kilogramos, del equipaje personal que llevan, en diferentes vuelos, un miembro de un equipo de béisbol y un jugador de un equipo de baloncesto.

Peso del equipaje (kilogramos)

Jugador de béisbol			**Jugador de baloncesto**	
16.3	20.0	18.6	15.4	16.3
18.1	15.0	15.4	17.7	18.1
15.9	18.6	15.6	18.6	16.8
14.1	14.5	18.3	12.7	14.1
17.7	19.1	17.4	15.0	13.6
16.3	13.6	14.8	15.9	16.3
13.2	17.2	16.5		

Utilice la prueba de la suma de rangos con $\alpha = 0.05$, para probar la hipótesis nula de que los dos atletas llevan la misma cantidad de equipaje en promedio, contra la hipótesis alternativa de que los pesos promedio del equipaje para los dos atletas son diferentes.

16.21 Los siguientes datos representan los tiempos de operación, en horas, para tres tipos de calculadoras científicas de bolsillo, antes de que requieran recarga:

Calculadora

A			***B***			***C***		
4.9	6.1	4.3	5.5	5.4	6.2	6.4	6.8	5.6
4.6	5.2		5.8	5.5	5.2	6.5	6.3	6.6
				4.8				

Utilice la prueba de Kruskal-Wallis, en el nivel de significancia 0.01, para probar la hipótesis de que los tiempos de operación para las tres calculadoras son iguales.

16.22 En el ejercicio 13.8 de la página 523 utilice la prueba de Kruskal-Wallis en el nivel de significancia de 0.05, para determinar si los solventes químicos orgánicos difieren de manera significativa en la tasa de absorción.

16.5 Pruebas de corridas

Al aplicar los diversos conceptos estadísticos que se presentan a lo largo de este libro, siempre se supone que nuestros datos muestrales se reúnen mediante algún procedimiento aleatorio. Las **pruebas de corridas**, que se basan en el orden en el que se obtienen las observaciones muestrales, es una técnica útil para probar la hipótesis nula H_0 de que las observaciones en realidad se extraen al azar.

Para ilustrar las pruebas de corridas, supongamos que se encuesta a 12 personas para saber si utilizan cierto producto. Se cuestionaría seriamente la supuesta aleatoriedad de la muestra si las 12 personas fueran del mismo sexo. Designaremos un hombre y una mujer con los símbolos M y F, respectivamente, y registraremos los resultados de acuerdo con su sexo en el orden en que suceden. Una secuencia común para el experimento sería

$$\underbrace{M\ M}\ \underbrace{F\ F\ F}\ \underbrace{M}\ \underbrace{F\ F}\ \underbrace{M\ M\ M\ M},$$

donde agrupamos las subsecuencias de símbolos similares. Tales agrupamientos se llaman **corridas**.

Definición 16.1: Una **corrida** es una subsecuencia de uno o más símbolos idénticos que representan una propiedad común de los datos.

Sin importar si las mediciones de nuestra muestra representan datos cualitativos o cuantitativos, la prueba de corridas divide los datos en dos categorías mutuamente excluyentes: masculino o femenino; defectuoso o no defectuoso; caras o cruces; arriba o abajo de la mediana; etcétera. En consecuencia, una secuencia siempre estará limitada a dos símbolos distintos. Sea n_1 el número de símbolos asociado con la categoría que ocurre menos, y n_2 el número de símbolos que pertenecen a la otra categoría. Entonces, el tamaño de la muestra $n = n_1 + n_2$.

Para los $n = 12$ símbolos en nuestra encuesta tenemos cinco corridas, con la primera que contiene dos M, la segunda tres F, etcétera. Si el número de corridas es mayor o menor que el que esperaríamos al azar, se debería rechazar la hipótesis de que la muestra se extrajo al azar. Ciertamente, una muestra que tiene como resultado sólo dos corridas,

$$M\ M\ M\ M\ M\ M\ M\ F\ F\ F\ F\ F,$$

o la inversa, es más improbable que ocurra a partir de un proceso de selección aleatoria. Tal resultado indica que las primeras 7 personas entrevistadas fueron todas hombres, seguidas de cinco mujeres. Asimismo, si la muestra tiene como resultado el número máximo de 12 corridas, como en la secuencia alternada

$$M\ F\ M\ F\ M\ F\ M\ F\ M\ F\ M\ F,$$

de nuevo sospecharíamos del orden en que se seleccionaron los individuos para la encuesta.

La prueba de corridas para la aleatoriedad se basa en la variable aleatoria V, el número total de corridas que suceden en la secuencia completa de nuestro experimento. En la tabla A.19, se dan valores de $P(V \leq v *$ cuando H_0 es verdadera) para

$v* = 2, 3, \ldots,$ 20 corridas, y valores de n_1 y n_2 menores o iguales que 10. Los valores P tanto para pruebas de una cola como de dos colas se pueden obtener usando estos valores tabulados.

En la encuesta anterior presentamos un total de 5 F y 7 M. De aquí, con $n_1 = 5$, $n_2 = 7$, y $v = 5$, de la tabla A.19 notamos para una prueba de dos colas que el valor P es

$$P = 2P(V \leq 5 \text{ cuando } H_0 \text{ es real}) = 0.394 > 0.05.$$

Es decir, el valor $v = 5$ es razonable en el nivel de significancia de 0.05 cuando H_0 es verdadera y, por lo tanto, no tenemos suficiente evidencia para rechazar la hipótesis de aleatoriedad en nuestra muestra.

Cuando el número de corridas es grande, por ejemplo, si $v = 11$ y $n_1 = 5$ y $n_2 = 7$, entonces el valor P en una prueba de dos colas es

$$P = 2P(V \geq 11 \text{ cuando } H_0 \text{ es real}) = 2[1 - P(V \leq 10 \text{ cuando } H_0 \text{ es real})]$$
$$2(1 - 0.992) = 0.016 < 0.05,$$

que nos lleva a rechazar la hipótesis de que los valores de la muestra ocurren al azar.

La prueba de corridas también sirve para detectar desviaciones en la aleatoriedad de una secuencia de mediciones cuantitativas en el tiempo, ocasionadas por tendencias o periodicidades. Al reemplazar cada medición en el orden en que se obtienen por un símbolo *más* si caen por arriba de la mediana, por un símbolo *menos* si caen por debajo de la mediana, y al omitir todas las mediciones que son exactamente iguales a la mediana, generamos una secuencia de símbolos más y menos que se prueban por su aleatoriedad como se ilustra en el siguiente ejemplo.

Ejemplo 16.7: Se ajusta una máquina para servir de adelgazador de pintura acrílica en un contenedor. ¿Diría que la cantidad de adelgazador de pintura que despacha la máquina varía de forma aleatoria, si se mide el contenido de los siguientes 15 contenedores y se encuentra que es 3.6, 3.9, 4.1, 3.6, 3.8, 3.7, 3.4, 4.0, 3.8, 4.1, 3.9, 4.0, 3.8, 4.2 y 4.1 litros? Utilice un nivel de significancia de 0.1.

Solución:

1. H_0: La secuencia es aleatoria.

2. H_1: La secuencia no es aleatoria.

3. $\alpha = 0.1$.

4. Estadístico de prueba: V, número total de corridas.

5. Cálculos: Para la muestra dada encontramos $\bar{x} = 3.9$. Al reemplazar cada medición por el símbolo "+", si cae por arriba de 3.9, por el símbolo "−" si cae por debajo de 3.9, y omitimos las dos mediciones que son iguales a 3.9, obtenemos la secuencia

$$- \; + \; - \; - \; - \; - \; + \; - \; + \; + \; - \; + \; +$$

para la que $n_1 = 6$, $n_2 = 7$ y $v = 6$. Por lo tanto, de la tabla A.19, el valor P calculado es

$$P = 2P(V \geq 8 \text{ cuando } H_0 \text{ es real}) = 2(0.5) = 1.$$

6. Decisión: No rechace la hipótesis de que la secuencia de mediciones varía de forma aleatoria.

La prueba de corridas, aunque menos poderosa, también se utiliza como una alternativa para la prueba de dos muestras de Wilcoxon, para probar la afirmación de que dos muestras aleatorias provienen de poblaciones que tienen la misma distribución y, por lo tanto, medias iguales. Si las poblaciones son simétricas, el rechazo de la afirmación de distribuciones iguales es equivalente a aceptar la hipótesis alternativa de que las medias no son iguales. Al llevar a cabo la prueba, primero combinamos las observaciones de ambas muestras y las acomodamos en orden ascendente. Asignamos ahora la letra A a cada observación tomada de una de las poblaciones; y la letra B, a cada observación de la segunda población. Así se genera una secuencia que consiste en los símbolos A y B. Si las observaciones de una población empatan con las observaciones de la otra población, la secuencia de símbolos A y B que se genera no será única y, en consecuencia, es poco probable que el número de corridas sea único. Los procedimientos para quitar los empates, por lo general, tienen como resultado tediosos cálculos adicionales, y por tal razón se preferiría la aplicación de la prueba de la suma de rangos de Wilcoxon siempre que ocurran dichas situaciones.

Para ilustrar el uso de las corridas en la prueba de medias iguales, considere los tiempos de sobrevivencia de los pacientes de leucemia del ejercicio 16.16 de la página 686 para los que tenemos

0.5	0.9	1.4	1.9	2.1	2.8	3.1	4.6	5.3
B	A	A	B	A	B	B	A	A

de donde resultan $v = 6$ corridas. Si las dos poblaciones simétricas tienen medias iguales, las observaciones de las dos muestras estarán entremezcladas, lo cual dará como resultado muchas corridas. Sin embargo, si las medias poblacionales son significativamente diferentes, esperaríamos que la mayoría de las observaciones para una de las dos muestras fueran más pequeñas que las de la otra muestra. En el caso extremo donde las poblaciones no se traslapan, obtendríamos una secuencia de la forma

$$A\ A\ A\ A\ A\ B\ B\ B\ B \text{ o } B\ B\ B\ B\ A\ A\ A\ A\ A$$

y en cualquier caso sólo hay dos corridas. En consecuencia, la hipótesis de medias poblacionales iguales se rechazará en el nivel de significancia α sólo cuando v es suficientemente pequeña, de modo que

$$P = P(V \leq v \text{ cuando } H_0 \text{ es real}) \leq \alpha,$$

lo que implica una prueba de una cola.

De regreso a los datos del ejercicio 16.16 de la página 686 para los que $n_1 = 4$, $n_2 = 5$ y $v = 6$, de la tabla A.19 encontramos que

$$P = P(V \leq 6 \text{ cuando } H_0 \text{ es real}) = 0.786 > 0.05$$

y, por lo tanto, no se rechaza la hipótesis nula de medias iguales. De aquí concluimos que el nuevo suero no prolonga la vida al no detener la leucemia.

Cuando n_1 y n_2 aumentan en tamaño, la distribución de muestreo de V se aproxima a la distribución normal con media

$$\mu_V = \frac{2n_1n_2}{n_1 + n_2} + 1 \qquad \text{y varianza} \qquad \sigma_V^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}.$$

En consecuencia, cuando n_1 y n_2 son ambos mayores que 10, se puede utilizar el estadístico

$$Z = \frac{V - \mu_V}{\sigma_V}$$

para establecer la región crítica para la prueba de corridas.

16.6 Límites de tolerancia

Los límites de tolerancia para una distribución normal de mediciones se presenta en el capítulo 9. En esta sección consideramos un método para construir intervalos de tolerancia que sean independientes de la forma de la distribución subyacente. Como se podría sospechar, para un grado de confianza razonable serán considerablemente más grandes que los que se construyen cuando se supone la normalidad, y el tamaño de la muestra que se requiere es, por lo general, muy grande. Los límites de tolerancia no paramétricos se establecen en función de las observaciones más grande y más pequeña en nuestra muestra.

Límites de tolerancia bilaterales

Para cualquier distribución de mediciones, los límites de tolerancia bilaterales se indican mediante las observaciones más grandes en una muestra de tamaño n, donde n se determina de manera que se afirme con una confianza de $(1-\gamma)100\%$ que, **al menos**, la proporción $1-\alpha$ de la distribución está incluida entre los extremos de la muestra.

La tabla A.20 da los tamaños muestrales que se requieren para valores seleccionados de γ y $1-\alpha$. Por ejemplo, cuando $\gamma = 0.01$ y $1-\alpha = 0.95$, debemos elegir una muestra aleatoria de tamaño $n = 130$, con la finalidad de tener 99% de confianza de que al menos 95% de la distribución de mediciones está incluida entre los extremos de la muestra.

En vez de determinar el tamaño muestral n de modo que una proporción específica de mediciones esté contenida entre los extremos de la muestra, en muchos procesos industriales es deseable determinar el tamaño de la muestra, de forma que una proporción fija de la población caiga por debajo de la observación más grande (o por arriba de la más pequeña) de la muestra. Tales límites se llaman límites de tolerancia unilaterales.

Límites de tolerancia unilaterales

Para cualquier distribución de mediciones, un límite de tolerancia unilateral se determina mediante la observación más pequeña (o más grande) en una muestra de tamaño n, donde n se determina de manera que se pueda asegurar con $(1-\gamma)100\%$ que, **al menos**, la proporción $1-\alpha$ de la distribución excederá la más pequeña (será menor que la mayor) observación de la muestra.

La tabla A.21 muestra los tamaños muestrales requeridos correspondientes a valores seleccionados de γ y $1-\alpha$. De aquí, cuando $\gamma = 0.05$ y $1-\alpha = 0.70$, debemos elegir una muestra de tamaño $n = 9$, para tener una confianza de 95% de que 70% de nuestra distribución de mediciones excederá la observación más pequeña de la muestra.

16.7 Coeficiente de correlación de rango

En el capítulo 11 utilizamos el coeficiente de correlación muestral r para medir la relación lineal entre dos variables continuas X y Y. Si los rangos 1, 2,..., n se asignan a las observaciones x en orden de magnitud y de manera similar a las observaciones y, y si estos rangos se sustituyen después con los valores numéricos reales en la fórmula para el coeficiente de correlación del capítulo 11, obtenemos la contraparte no paramétrica del coeficiente de correlación convencional. Un coeficiente de correlación calculado de esta forma se conoce como **coeficiente de correlación de rangos de Spearman**, y se denota con r_s. Cuando no hay empates entre ambos conjuntos de mediciones, la fórmula para r_s se reduce a una expresión mucho más simple que incluye las diferencias d_i entre los rangos asignados a los n pares de x y y, que establecemos ahora.

Coeficiente de correlación de rango

Una medición no paramétrica de la asociación entre dos variables X y Y está dada por el **coeficiente de correlación de rango**

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2,$$

donde d_i es la diferencia entre los rangos asignados x_i y y_i, y n es el número de pares de datos.

En la práctica la fórmula anterior también se usa cuando hay empates entre las observaciones x o y. Los rangos para observaciones empatadas se asignan como en la prueba de rango con signo al promediar los rangos que se habrían asignado si las observaciones fueran distinguibles.

El valor de r_s, por lo general, estará cercano al valor que se obtiene al encontrar r con base en mediciones numéricas y se interpreta casi en la misma forma. Como antes, el valor de r_s irá de -1 a $+1$. Un valor de $+1$ o -1 indica una asociación perfecta entre X y Y, el signo más ocurre para rangos idénticos y el signo menos para rangos inversos. Cuando r_s es cercano a cero, concluiríamos que las variables no están correlacionadas.

Ejemplo 16.8: Las cifras que se listan en la tabla 16.7, publicadas por la Comisión Federal de Comercio, muestran los miligramos de alquitrán y nicotina que se encuentra en 10 marcas de cigarrillos. Calcule el coeficiente de correlación de rangos para medir el grado de relación entre el contenido de alquitrán y nicotina en cigarrillos.

Tabla 16.7: Contenidos de alquitrán y nicotina

Marca de cigarrillo	Contenido de alquitrán	Contenido de nicotina
Viceroy	14	0.9
Marlboro	17	1.1
Chesterfield	28	1.6
Kool	17	1.3
Kent	16	1.0
Raleigh	13	0.8
Old Gold	24	1.5
Philip Morris	25	1.4
Oasis	18	1.2
Players	31	2.0

Solución: Representemos con X y Y los contenidos de alquitrán y nicotina, respectivamente. Primero asignamos rangos a cada conjunto de medidas, con el rango de 1 asignado al número más bajo en cada conjunto, el rango de 2 al segundo número más bajo en cada conjunto, y así sucesivamente, hasta que se asigna el rango 10 al número más grande. La tabla 16.8 muestra los rangos individuales de las mediciones y las diferencias en rangos para los 10 pares de observaciones.

Tabla 16.8: Rangos para los contenidos de alquitrán y nicotina

Marca de cigarrillo	x_i	y_i	d_i
Viceroy	2	2	0
Marlboro	4.5	4	0.5
Chesterfield	9	9	0
Kool	4.5	6	−1.5
Kent	3	3	0
Raleigh	1	1	0
Old Gold	7	8	−1
Philip Morris	8	7	1
Oasis	6	5	1
Players	10	10	0

Al sustituir en la fórmula para r_s, encontramos que

$$r_s = 1 - \frac{(6)(5.50)}{(10)(100-1)} = 0.967,$$

lo que indica una correlación positiva alta entre la cantidad de alquitrán y de nicotina que se encuentra en los cigarrillos.

Hay algunas ventajas al usar r_s en vez de r. Por ejemplo, ya no suponemos que la relación fundamental entre X y Y es lineal y, por lo tanto, cuando los datos poseen una relación curvilínea distinta, el coeficiente de correlación de rangos probablemente será más confiable que la medición convencional. Una segunda ventaja del uso del coeficiente de correlación de rangos es el hecho de que no se hacen suposiciones de normalidad con respecto a las distribuciones de X y Y. Quizá la mayor ventaja ocurre cuando se es incapaz de hacer mediciones numéricas significativas y, sin embargo, se pueden establecer rangos. Tal es el caso, por ejemplo, cuando diferentes jueces clasifican a un grupo de individuos de acuerdo con algún atributo. El coeficiente de correlación de rangos se puede utilizar en esta situación como una medida de la consistencia de los dos jueces.

Para probar la hipótesis de que $\rho = 0$ con el uso de un coeficiente de correlación de rangos, se necesita considerar la distribución muestral de los valores r_s, bajo la suposición de no correlación. En la tabla A.22 aparecen valores críticos calculados para $\alpha = 0.05$, 0.025, 0.01 y 0.005. La elaboración de esta tabla es similar a la tabla de valores críticos para la distribución t, excepto para la columna izquierda, que ahora da el número de pares de observaciones en vez de los grados de libertad. Como la distribución de los valores r_s es simétrica alrededor de cero cuando $\rho = 0$, el valor r_s que deja un área de α a la izquierda es igual al negativo del valor r_s que deja un área α a la derecha. Para una hipótesis alternativa bilateral, la región crítica de tamaño α cae igualmente en las dos colas de la distribución. Para una prueba en la que la hipótesis alternativa es negativa, la región crítica está completamente en la cola izquierda de la distribución y, cuando la alternativa es positiva, la región crítica se coloca por completo en la cola derecha.

Ejemplo 16.9: Refiérase al ejemplo 16.8 y pruebe la hipótesis de que la correlación entre la cantidad de alquitrán y nicotina en los cigarrillos es cero contra la alternativa de que es mayor que cero. Utilice un nivel de significancia de 0.01.

Solución: **1.** H_0: $\rho = 0$.

2. H_1: $\rho > 0$.

3. $\alpha = 0.01$.

4. Región crítica: $r_s > 0.745$, de la tabla A.22.

5. Cálculos: Del ejemplo 16.8, $r_s = 0.967$.

6. Decisión: Rechace H_0 y concluya que hay una correlación significativa entre la cantidad de alquitrán y nicotina que se encuentra en los cigarrillos.

Con la suposición de no correlación, se puede mostrar que la distribución de los valores r_s se aproxima a una distribución normal con una media 0 y desviación estándar de $1/\sqrt{n-1}$ conforme n aumenta. En consecuencia, cuando n excede los valores dados en la tabla A.22, se podría probar la correlación de significancia mediante el cálculo de

$$z = \frac{r_s - 0}{1/\sqrt{n-1}} = r_s\sqrt{n-1}$$

y la comparación con los valores críticos de la distribución normal estándar que se muestran en la tabla A.3.

Ejercicios

16.23 Se selecciona una muestra aleatoria de 15 adultos que viven en una pequeña ciudad, con la finalidad de estimar la proporción de votantes que favorecen a cierto candidato para alcalde. También se le preguntó a cada individuo si era graduado universitario. Se obtiene la siguiente secuencia, al hacer que Y y N designen las respuestas de "sí" y "no" a la pregunta sobre instrucción:

$$N\ N\ N\ N\ Y\ Y\ N\ Y\ Y\ N\ Y\ N\ N\ N\ N$$

Utilice la prueba de corridas en el nivel de significancia de 0.1, para determinar si la secuencia apoya la afirmación de que la muestra se seleccionó al azar.

16.24 Se utiliza un proceso de plateado para cubrir cierto tipo de charola de servicio. Cuando el proceso está bajo control, el espesor de la plata sobre la charola variará de forma aleatoria siguiendo una distribución normal con una media de 0.02 milímetros y una desviación estándar de 0.005 milímetros. Suponga que las siguientes 12 charolas examinadas muestran los siguientes espesores de plata: 0.019, 0.021, 0.020, 0.019, 0.020, 0.018, 0.023, 0.021, 0.024, 0.022, 0.023, 0.022. Utilice la prueba de corridas para determinar si las fluctuaciones en el espesor de una charola a otra son aleatorias. Sea $\alpha = 0.05$.

16.25 Use la prueba de corridas para probar si hay una diferencia en el tiempo promedio de operación para las dos calculadoras del ejercicio 16.17 en la página 686.

16.26 En una línea de producción industrial, los artículos se inspeccionan de forma periódica en busca de defectuosos. Lo siguiente es una secuencia de artículos defectuosos, D, y no defectuosos, N, producidos por esta línea:

$$D\ D\ N\ N\ N\ D\ N\ N\ D\ D\ N\ N\ N\ N$$
$$N\ D\ D\ D\ N\ N\ D\ N\ N\ N\ N\ D\ N\ D$$

Utilice la teoría de muestras grandes para la prueba de corridas, con un nivel de significancia de 0.05, para determinar si los defectuosos ocurren o no al azar.

16.27 Suponga que las mediciones del ejercicio 1.14 de la página 28 se registran en renglones sucesivos de izquierda a derecha conforme se reúnen, utilice la prueba de corridas, con $\alpha = 0.05$, para probar la hipótesis de que los datos representan una secuencia aleatoria.

16.28 ¿Qué tan grande se requiere que sea una muestra para tener 95% de confianza de que al menos 85% de la distribución de medidas se incluye entre los extremos de la muestra?

16.29 ¿Cuál es la probabilidad de que el rango de una muestra aleatoria de tamaño 24 incluya al menos 90% de la población?

16.30 ¿Qué tan grande se requiere que sea una muestra para tener 99% de confianza de que al menos 80% de la población sea menor que la observación más grande de la muestra?

16.31 ¿Cuál es la probabilidad de que al menos 95% de una población exceda el valor más pequeño en una muestra aleatoria de tamaño $n = 135$?

16.32 La siguiente tabla da las calificaciones registradas de 10 estudiantes en un examen de mitad del semestre y la del examen final en un curso de cálculo:

Estudiante	Examen de mitad del semestre	Examen final
L.S.A.	84	73
W.P.B.	98	63
R.W.K.	91	87
J.R.L.	72	66
J.K.L.	86	78
D.L.P.	93	78
B.L.P.	80	91
D.W.M.	0	0
M.N.M.	92	88
R.H.S.	87	77

a) Calcule el coeficiente de correlación de rangos.

b) Pruebe la hipótesis nula de que $\rho = 0$ contra la alternativa de que $\rho > 0$. Utilice $\alpha = 0.025$.

16.33 Con referencia a los datos del ejercicio 11.1 de la página 397,

a) calcule el coeficiente de correlación de rangos;

b) en el nivel de significancia 0.05 pruebe la hipótesis nula de que $\rho = 0$ contra la alternativa de que $\rho \neq 0$. Compare sus resultados con los que se obtienen en el ejercicio 11.53 de la página 438.

16.34 Calcule el coeficiente de correlación de rangos para la precipitación pluvial diaria y la cantidad de partículas eliminadas en el ejercicio 11.9 de la página 399.

16.35 Con referencia a los pesos y tamaños del tórax de los infantes del ejercicio 11.52 en la página 438,

a) calcule el coeficiente de correlación de rangos;

b) en el nivel de significancia 0.025 pruebe la hipótesis de que $\rho = 0$ contra la alternativa de que $\rho > 0$.

16.36 Un grupo de consumidores prueba la calidad general de nueve marcas de hornos de microondas. Los rangos asignados por el grupo y los precios de venta sugeridos son los siguientes:

Fabricante	Clasificación del grupo	Precio sugerido
A	6	$480
B	9	395
C	2	575
D	8	550
E	5	510
F	1	545
G	7	400
H	4	465
I	3	420

¿Existe una relación significativa entre la calidad y el precio de un horno de microondas? Utilice un nivel de significancia de 0.05.

16.37 En un desfile de regreso a clases dos jueces califican ocho carros alegóricos en el siguiente orden:

	Carro alegórico							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Juez A	5	8	4	3	6	2	7	1
Juez B	7	5	4	2	8	1	6	3

a) Calcule la correlación de rangos.

b) Pruebe la hipótesis nula de que $\rho = 0$ contra la alternativa de que $\rho > 0$. Use $\alpha = 0.05$.

16.38 En el artículo titulado "Risky Assumptions" de Paul Slovic, Baruch Fischoff y Sarah Lichtenstein, publicado en *Psychology Today* (junio de 1980), miembros de la Liga de Mujeres Votantes y expertos profesionalmente implicados en la evaluación de riesgos clasificaron el riesgo de muerte, en Estados Unidos, para 30 actividades y tecnologías. Las puntuaciones se presentan en la tabla 16.9.

a) Calcule el coeficiente de correlación de rangos.

b) Pruebe la hipótesis nula de cero correlación entre las clasificaciones de la Liga de Mujeres Votantes y de los expertos contra la alternativa de que la correlación no es cero. Utilice un nivel de significancia de 0.05.

Tabla 16.9: Datos de puntuación para el ejercicio 16.38

Riesgo de la actividad o tecnología	Votantes	Expertos	Riesgo de la actividad o tecnología	Votantes	Expertos
Energía nuclear	1	20	**Vehículos de motor**	2	1
Armas de fuego	3	4	**Tabaquismo**	4	2
Motocicletas	5	6	**Bebidas alcohólicas**	6	3
Aviación privada	7	12	**Trabajo policiaco**	8	17
Pesticidas	9	8	**Cirugía**	10	5
Bombero	11	18	**Construcción**	12	13
Cacería	13	23	**Latas de aerosol**	14	26
Montañismo	15	29	**Bicicletas**	16	15
Aviación comercial	17	16	**Energía eléctrica**	18	9
Natación	19	10	**Anticonceptivos**	20	11
Esquí	21	30	**Rayos X**	22	7
Fútbol americano	23	27	**Ferrocarriles**	24	19
Conservadores de alimentos	25	14	**Colorantes de alimentos**	26	21
Podadoras	27	28	**Antibióticos**	28	24
Electrodomésticos	29	22	**Vacunas**	30	25

Ejercicios de repaso

16.39 Un estudio de una compañía química compara las propiedades de desecación de dos diferentes polímeros. Se utilizaron 10 lodos diferentes y se permitió que ambos polímeros secaran cada lodo. El secado libre se midió en ml/min.

a) Utilice la prueba de signos en el nivel 0.05 para probar la hipótesis nula de que el polímero *A* tiene la misma mediana de secado que el polímero *B*.

b) Utilice la prueba de rangos con signo para probar la hipótesis del inciso *a*).

Tipo de lodo	Polímero *A*	Polímero *B*
1	12.7	12.0
2	14.6	15.0
3	18.6	19.2
4	17.5	17.3
5	11.8	12.2
6	16.9	16.6
7	19.9	20.1
8	17.6	17.6
9	15.6	16.0
10	16.0	16.1

16.40 En el ejercicio de repaso 13.58 de la página 568, use la prueba de Kruskal-Wallis, en el nivel de significancia de 0.05, para determinar si los análisis químicos realizados por los cuatro laboratorios dan, en promedio, los mismos resultados.

16.41 Use los datos del ejercicio 13.12 de la página 533 para ver si la cantidad mediana de pérdida de nitrógeno en la transpiración es diferente para los tres niveles de proteína dietética.

Capítulo 17

Control estadístico de la calidad

17.1 Introducción

La noción del uso de las técnicas de muestreo y de análisis estadístico en un escenario de producción tiene sus comienzos en la década de 1920. El objetivo de este concepto altamente exitoso es la reducción sistemática de la variabilidad y el aislamiento asociado de las fuentes de dificultades *durante la producción*. En 1924 Walter A. Shewhart de la empresa Bell Telephone Laboratories desarrolló el concepto de una gráfica de control. Sin embargo, no fue sino hasta la Segunda Guerra Mundial que se generalizó el uso de gráficas de control. Esto se debió a la importancia de mantener la calidad en los procesos de producción durante ese periodo. En las décadas de 1950 y 1960 el desarrollo del control de calidad y el área general de seguridad de la calidad crecieron de manera rápida, en particular con el surgimiento del programa espacial en Estados Unidos. En Japón hubo un amplio y exitoso uso del control de calidad gracias a los esfuerzos de W. Edwards Deming, quien trabajó como consultor en Japón después de la Segunda Guerra Mundial. El control de calidad ha sido, y es, un elemento importante en el desarrollo de la industria y de la economía nipones.

El control de calidad recibe una creciente atención como herramienta de administración en la que importantes características de un producto se observan, evalúan y comparan con algún tipo de estándar. Los diversos procedimientos en el control de calidad implican un uso considerable de los procedimientos de muestreo y principios estadísticos, que ya estudiamos en capítulos anteriores. Los usuarios principales del control de calidad son, por supuesto, las corporaciones industriales. Resulta claro que un programa eficaz de control de calidad aumenta tanto la calidad del artículo que se produce como las utilidades. Esto es particularmente cierto en la actualidad, pues los productos se fabrican en volúmenes altos. Antes del movimiento hacia los métodos de control, la calidad a menudo sufría a causa de la falta de eficiencia que, por supuesto, incrementa los costos.

La gráfica de control

El propósito de una gráfica de control es determinar si el desempeño de un proceso se mantiene en un nivel aceptable de calidad. Se espera, desde luego, que cualquier proceso experimente una variabilidad natural, es decir, variabilidad debida esencial-

mente a fuentes de variación poco importantes e incontrolables. Por otro lado, un proceso puede experimentar tipos más serios de variabilidad en mediciones de desempeño claves. Estas fuentes de variabilidad pueden surgir de uno de varios tipos de "causas asignables" no aleatorias, como errores del operador o indicadores mal ajustados en una máquina. Un proceso que opera en dicho estado se denomina **fuera de control**. Se dice que un proceso que experimenta sólo variaciones aleatorias está en **control estadístico**. Desde luego, un proceso de producción exitoso puede operar en un estado de control durante un periodo largo. Se supone que durante este periodo el proceso elabora un producto aceptable. Sin embargo, quizás haya un "corrimiento" gradual o repentino que requiera detección.

Una gráfica de control tiene la finalidad de ser un dispositivo para detectar el estado no aleatorio o fuera de control de un proceso. Por lo general, la gráfica de control toma la forma que se indica en la figura 17.1. Es importante que el corrimiento se detecte de forma rápida, de manera que se pueda corregir el problema. Evidentemente, si la detección es lenta se producen muchos artículos defectuosos o fuera de las especificaciones, lo cual da como resultado un desperdicio significativo y un incremento en los costos.

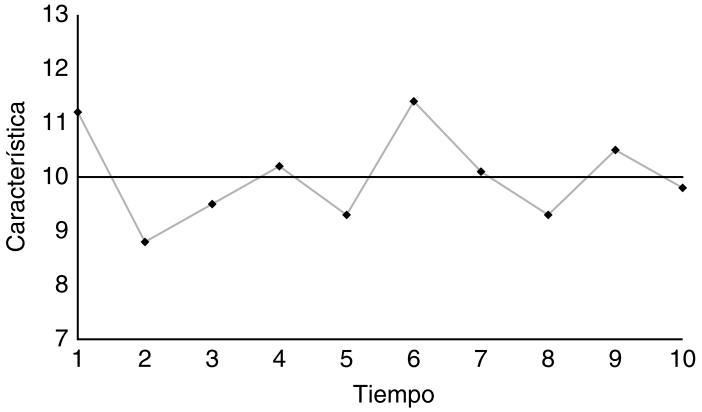


Figura 17.1: Gráfica de control típica.

Se deben considerar algunos tipos de características de la calidad y las unidades del proceso se deben muestrear conforme pasa el tiempo. Por ejemplo, la característica puede ser la circunferencia de un cojinete de motor. La línea central representa el valor promedio de la característica cuando el proceso está controlado. Los puntos que se indican en la figura representarían los resultados de, digamos, promedios muestrales de tal característica, con las muestras tomadas respecto al tiempo. Los límites de control superior e inferior se eligen de modo que se esperaría que todos los puntos muestrales queden cubiertos por estos límites, si el proceso está controlado. Como resultado, la forma general de los puntos graficados respecto al tiempo determina si se concluye que el proceso está dentro de control. La evidencia para estar "dentro de control" se obtiene de un patrón aleatorio de puntos, con todos los valores graficados dentro de los límites de control. Cuando un punto cae fuera de los límites de control, esto se toma como evidencia de un proceso que está fuera de control, y se sugiere una búsqueda de la causa. Además, un patrón no aleatorio de puntos se puede considerar sospechoso y en realidad una indicación de que se necesita una investigación de la acción correctiva adecuada.

17.2 Naturaleza de los límites de control

Las ideas fundamentales en las que se basan las gráficas de control son similares en estructura a la prueba de hipótesis. Los límites de control se establecen para controlar la probabilidad de cometer el error de concluir que el proceso está fuera de control, cuando de hecho no lo está. Esto corresponde a la probabilidad de cometer un error tipo I, si probáramos la hipótesis nula de que el proceso está bajo control. Por otro lado, debemos estar atentos al error del segundo tipo; a saber, no encontrar el proceso fuera de control cuando de hecho sí lo está (error tipo II). De esta manera, la elección de los límites de control es similar a la elección de una región crítica.

Como en el caso de la prueba de hipótesis, es importante el tamaño de la muestra en cada punto. La consideración del tamaño de la muestra depende, en gran medida, de la sensibilidad o potencia de detección del estado fuera de control. En esta aplicación, la noción de *potencia* es muy similar a la situación de la prueba de hipótesis. Es claro que cuanto más grande sea la muestra en cada periodo, más rápida será la detección de un proceso fuera de control. En un sentido, los límites de control en realidad definen lo que el usuario considera como estar *bajo control*. En otras palabras, evidentemente la anchura dada por los límites de control debe depender en cierto sentido de la variabilidad del proceso. Como resultado, el cálculo de los límites de control dependerá por completo de manera natural de los datos que se tomen de los resultados del proceso. De esta forma, cualquier control de calidad debe tener su comienzo en el cálculo a partir de una muestra preliminar o conjunto de muestras, que establecerán tanto la línea central como los límites de control de calidad.

17.3 Propósitos de la gráfica de control

Un propósito evidente de la gráfica de control es la mera vigilancia del proceso, es decir, determinar si se necesitan realizar cambios. Además, la constante obtención sistemática de datos a menudo permite a la administración evaluar la capacidad del proceso. Claramente, si una sola característica de desempeño es importante, el muestreo y la estimación continuos de la media y la desviación estándar de la característica de desempeño ofrece la actualización de lo que el proceso puede hacer en términos de desempeño medio y variación aleatoria. Esto es valioso aun si el proceso permanece bajo control durante periodos largos. La estructura sistemática y formal de la gráfica a menudo puede prevenir una reacción desmesurada ante cambios que representen sólo fluctuaciones aleatorias. En efecto, en muchas situaciones, los cambios realizados por una reacción desmesurada llega a crear problemas serios, difíciles de resolver.

Las características de calidad de las gráficas de control caen, por lo general, en *dos* categorías: **variables** y **atributos**. Como resultado, los tipos de gráficas de control con frecuencia tienen las mismas clasificaciones. En el caso de las gráficas de tipo variables, la característica, por lo general, es una medición sobre un continuo, como diámetro, peso, etcétera. Para la gráfica de atributos, la característica refleja si el producto individual *concuerda* (es o no defectuoso). Las aplicaciones de estas dos situaciones distintas son evidentes.

En el caso de la gráfica de variables, se debe ejercer control sobre la tendencia central y la variabilidad. Un analista de control de calidad se debe preocupar de si existe *en promedio* un corrimiento de los valores de la característica de desempeño. Además, siempre habrá interés acerca de si algún cambio en las condiciones del proceso tiene como resultado una disminución en la precisión (es decir, un aumento

en la variabilidad). Para tratar con estos dos conceptos son esenciales gráficas de control separadas. La tendencia central está controlada por la *gráfica* $\bar{X}$, donde las medias de muestras relativamente pequeñas se grafican en la gráfica de control. La variabilidad alrededor de la media se controla mediante el *rango* en la muestra, o la *desviacion estándar* muestral. En el caso de muestreo de atributos, la *proporción de defectuosos* de una muestra a menudo es la cantidad que se grafica. En la siguiente sección analizamos el desarrollo de gráficas de control para la característica de desempeño tipo variable.

17.4 Gráficas de control para variables

Un ejemplo es una forma relativamente fácil para entender los rudimentos de la gráfica $\bar{X}$ para variables. Suponga que las gráficas de control de calidad se deben utilizar en un proceso de fabricación de cierta parte de un motor. Suponga que la media del proceso es $\mu = 50$ mm y que la desviación estándar es $\sigma = 0.01$ mm. Suponga que se muestrean grupos de 5 cada hora, y que los valores de la *media muestral* $\bar{X}$ se registran y grafican como en la figura 17.2. Los límites para las gráficas $\bar{X}$ se basan en la desviación estándar de la variable aleatoria $\bar{X}$. Sabemos del material del capítulo 8 que para el promedio de observaciones independientes en una muestra de tamaño n,

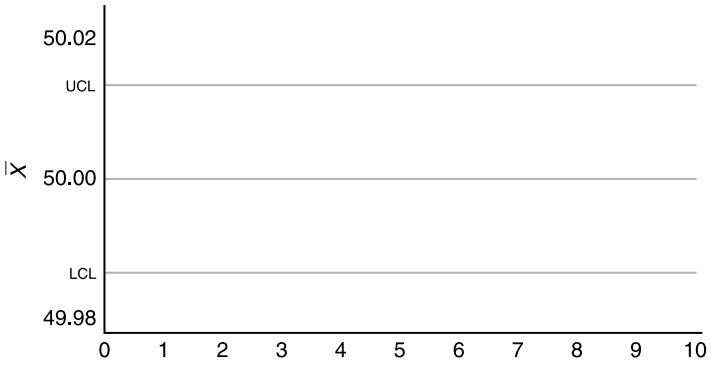


Figura 17.2: Los límites de control 3σ para el ejemplo de la parte del motor.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

donde σ es la desviación estándar de una observación individual. Los límites de control se determinan de manera que tengan como resultado una pequeña probabilidad de que un valor dado de $\bar{X}$ esté fuera de los límites dado que, en realidad, el proceso está bajo control (es decir, $\mu = 50$). Si recurrimos al teorema del límite central, tenemos que bajo la condición de que el proceso está bajo control,

$$\bar{X} \sim N\left(50, \frac{0.01}{\sqrt{5}}\right).$$

Como resultado, $(1 - \alpha)100\%$ de los valores $\bar{X}$ cae dentro de los límites cuando el proceso está bajo control si utilizamos los límites

$$\text{LCL} = \mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 50 - z_{\alpha/2}(0.0045), \qquad \text{UCL} = \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 50 + z_{\alpha/2}(0.0045).$$

Aquí, LCL y UCL representan límite de control inferior y límite de control superior, respectivamente. Con frecuencia las gráficas $\bar{X}$ se basan en límites que se denominan como límites "tres-sigma", con referencia, por supuesto, a $z_\alpha/2 = 3$ y a límites que se convierten en

$$\mu \pm 3\frac{\sigma}{\sqrt{n}}.$$

En nuestro ejemplo los límites superior e inferior son

$$\text{LCL} = 50 - 3(0.0045) = 49.9865, \qquad \text{UCL} = 50 + 3(0.0045) = 50.0135.$$

De esta forma, si vemos la estructura de los límites 3σ desde el punto de vista de la prueba de hipótesis, para un punto muestral dado, la probabilidad de que el valor $\bar{X}$ caiga fuera de los límites de control es 0.0026, dado que el proceso está bajo control. Ésta es la probabilidad de que el analista determine *de manera errónea* que el proceso está fuera de control (véase la tabla A.3).

El ejemplo anterior no sólo ilustra la gráfica $\bar{X}$ para variables, sino también debería proporcionar al lector una idea de la naturaleza de las gráficas de control en general. La línea central, por lo general, refleja el valor ideal de un parámetro importante. Los límites de control se establecen a partir del conocimiento de las propiedades de muestreo del estadístico, que estima el parámetro en cuestión. Muy a menudo éstos implican un múltiplo de la desviación estándar del estadístico. Se ha vuelto una práctica general utilizar límites 3σ. En el caso de la gráfica $\bar{X}$ que se proporciona aquí, el teorema del límite central brinda al usuario una buena aproximación de la probabilidad de determinar en falso que el proceso está fuera de control. En general, sin embargo, quizás el usuario no confíe en la normalidad del estadístico de la línea central. Como resultado, tal vez no se conozca la probabilidad exacta de un "error tipo I". A pesar de esto, es casi estándar el uso de límites $k\sigma$. Mientras el uso de los límites 3σ es amplio, a veces el usuario puede desear desviarse de esta aproximación. Un múltiplo menor de σ quizá sea apropiado cuando es importante detectar de forma rápida una situación fuera de control. Debido a consideraciones económicas, puede resultar costoso permitir que un proceso continúe funcionando fuera de control, incluso por periodos cortos; mientras que el costo de la búsqueda y corrección de las causas imputables puede ser relativamente pequeño. Es claro, en este caso, que son adecuados los límites de control que son más estrictos que los límites 3σ.

Subgrupos racionales

Los valores muestrales a ser usados en un esfuerzo de control de calidad se dividen en subgrupos con una *muestra* que representa un subgrupo. Como indicamos antes, el orden en el tiempo de producción es en realidad una base natural para la selección de los subgrupos. Se puede ver el esfuerzo de control de calidad de manera muy simple como **1.** muestreo, **2.** detección de un estado fuera de control y **3.** búsqueda de las causas imputables que puedan ocurrir con el tiempo. La selección de la base para estos grupos muestrales parece ser bastante directa. La elección de estos subgrupos de información muestral podría tener un efecto importante en el éxito del programa de control de calidad. Estos subgrupos con frecuencia se denominan **subgrupos racionales**. Generalmente, si el analista se interesa en la detección de un *corrimiento*

de la ubicación, se considera que los subgrupos se deben elegir de manera que la variabilidad dentro del subgrupo sea pequeña, y que las causas asignables, si se presentaran, puedan tener la mayor posibilidad de detección. Así, deseamos elegir los subgrupos de forma que se maximice la variabilidad entre subgrupos. La elección de unidades en un subgrupo que se producen en tiempos muy cercanos, por ejemplo, es una aproximación razonable. Por otro lado, las gráficas de control a menudo se utilizan para controlar la variabilidad, en cuyo caso el estadístico de desempeño es la *variabilidad dentro de la muestra*. Por ello, es más importante elegir los subgrupos racionales para maximizar la variabilidad dentro de la muestra. En este caso, las observaciones en los subgrupos se deberían comportar más como una muestra aleatoria y esta variabilidad dentro de las muestras necesita ser una descripción de la variabilidad del proceso.

Es importante notar que las gráficas de control sobre la variabilidad se deben establecer antes del desarrollo de las gráficas sobre el centro de ubicación (digamos, gráficas $\bar{X}$). Cualquier gráfica de control sobre el centro de ubicación en realidad dependerá de la variabilidad. Por ejemplo, vimos un ejemplo de la gráfica de tendencia central y ésta depende de σ. En las secciones que siguen, se presentará una estimación de σ a partir de los datos.

Gráficas $\bar{X}$ con parámetros estimados

Ilustramos con anterioridad las nociones de la gráfica $\bar{X}$ que usa el teorema del límite central, y empleamos valores *conocidos* de la media y desviación estándar del proceso. Como se indicó, se utilizan los límites de control

$$\text{LCL} = \mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad \text{UCL} = \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

y un valor $\bar{X}$ que cae fuera de estos límites se considera evidencia de que la media μ cambió y por ello el proceso quizás esté fuera de control.

En muchas situaciones prácticas no es razonable suponer que conocemos μ y σ. Como resultado, se deben proporcionar estimaciones de los datos que se obtienen cuando el proceso está bajo control. Típicamente, las estimaciones se determinan durante un periodo en el que se reúne *información de origen o de inicio*. Se elige una base para subgrupos racionales y los datos se reúnen con muestras de tamaño n en cada subgrupo. Los tamaños muestrales, por lo general, son pequeños, digamos, 4, 5 o 6, y se toman k muestras, con k al menos igual a 20. Durante este periodo en el que se supone que el proceso está bajo control, el usuario establece estimaciones de μ y σ, sobre las que se basa la gráfica de control. La información importante reunida durante este periodo incluye las medias muestrales en el subgrupo, la media general y el rango de la muestra en cada subgrupo. En los siguientes párrafos señalaremos cómo se utiliza esta información para producir la gráfica de control.

Una parte de la información muestral de estas k muestras toma la forma $\bar{X}_1$, $\bar{X}_2, \ldots, \bar{X}_k$, donde la variable aleatoria $\bar{X}_i$ es el promedio de los valores en la i-ésima muestra. Evidentemente, el promedio global es la variable aleatoria

$$\bar{\bar{X}} = \frac{1}{k}\sum_{i=1}^{k}\bar{X}_i.$$

Éste es el estimador adecuado de la media del proceso y, como resultado, es la línea central en la gráfica de control $\bar{X}$. En aplicaciones de control de calidad a menudo es conveniente estimar σ a partir de la información relacionada con los *rangos* en

las muestras, en vez de las desviaciones estándar de las muestras. Para la i-ésima muestra definamos

$$R_i = X_{\text{máx},i} - X_{\text{mín},i}$$

como el rango para los datos en la i-ésima muestra. Aquí $X_{\text{máx},i}$ y $X_{\text{mín},i}$ son, respectivamente, la observación más grande y la más pequeña en la muestra. La estimación apropiada de σ es una función del rango promedio

$$\bar{R} = \frac{1}{k}\sum_{i=1}^{k} R_i.$$

Una estimación de σ, digamos $\hat{\sigma}$, se obtiene mediante

$$\hat{\sigma} = \frac{\bar{R}}{d_2},$$

donde d_2 es una constante que depende del tamaño de la muestra. Los valores de d_2 se muestran en la tabla A.23.

El uso del rango para producir una estimación de σ tiene sus raíces en aplicaciones del tipo de control de calidad, en particular debido a que el rango era muy fácil de calcular (en la época anterior al periodo en que el tiempo de cálculo no se considera una dificultad). La suposición de normalidad de las observaciones individuales está implícita en la gráfica $\bar{X}$. Por supuesto, la existencia del teorema del límite central es ciertamente útil a este respecto. Bajo la suposición de normalidad, usamos una variable aleatoria que se denomina rango relativo, dada por

$$W = \frac{R}{\sigma}.$$

Resulta que los momentos de W son funciones simples del tamaño muestral n (véase la referencia a Montgomery, 2000, en la bibliografía). El valor esperado de W a menudo se denomina d_2. Así, al tomar el valor esperado de W anterior,

$$\frac{E(R)}{\sigma} = d_2.$$

Como resultado, se comprende con facilidad el fundamento para la estimación de $\hat{\sigma} = \bar{R}/d_2$. Se sabe bien que el método del rango produce un estimador eficiente de σ en muestras relativamente pequeñas. Esto hace que el estimador sea en particular atractivo en aplicaciones de control de calidad, ya que los tamaños muestrales en los subgrupos, por lo general, son pequeños. El uso del método del rango para la estimación de σ tiene como resultado gráficas de control con los siguientes parámetros:

$$\text{UCL} = \bar{\bar{X}} + \frac{3\bar{R}}{d_2\sqrt{n}}, \qquad \text{línea central} = \bar{\bar{X}}, \qquad \text{LCL} = \bar{\bar{X}} - \frac{3\bar{R}}{d_2\sqrt{n}}.$$

Al definir la cantidad

$$A_2 = \frac{3}{d_2\sqrt{n}},$$

tenemos que

$$\text{UCL} = \bar{\bar{X}} + A_2\bar{R}, \qquad \text{LCL} = \bar{\bar{X}} - A_2\bar{R}.$$

Para simplificar la estructura, el usuario de las gráficas $\bar{X}$ a menudo encuentra valores tabulados de A_2. En la tabla A.23 se dan tabulaciones de los valores de A_2 para varios tamaños muestrales.

Gráficas *R* para control de variación

Hasta aquí todos los ejemplos y detalles trataron sobre el intento del analista de control de calidad de detectar condiciones fuera de control producidas por un *corrimiento en la media*. Los límites de control se basan en la distribución de la variable aleatoria $\bar{X}$ y dependen de la suposición de normalidad de las observaciones individuales. Es importante para el control que se aplique a la variabilidad, así como al centro de ubicación. De hecho, muchos expertos consideran que el control de variabilidad de la característica de desempeño es más importante y que se debe establecer antes de considerar el centro de ubicación. La variabilidad del proceso se puede controlar usando *gráficas del rango muestral*. Una gráfica de los rangos muestrales respecto al tiempo se denomina **gráfica *R***. Se puede utilizar la misma estructura general como en el caso de la gráfica $\bar{X}$, con $\bar{R}$ como *línea central* y los límites de control dependen de una estimación de la desviación estándar de la variable aleatoria R. Así, como en el caso de la gráfica $\bar{X}$, establecen límites 3σ donde "3σ" implica $3\sigma_R$. La cantidad σ_R se debe estimar a partir de los datos, justo como se estima $\sigma_{\bar{X}}$.

La estimación de σ_R, la desviación estándar, también se basa en la distribución del rango relativo

$$W = \frac{R}{\sigma}.$$

La desviación estándar de W es una función conocida del tamaño muestral y, por lo general, se denota como d_3. Entonces,

$$\sigma_R = \sigma d_3.$$

Podemos reemplazar ahora σ por $\hat{\sigma} = \bar{R}/d_2$, y de esta forma el estimador de σ_R es

$$\hat{\sigma}_R = \frac{\bar{R}d_3}{d_2}.$$

Así las cantidades que definen la gráfica R son

$$\text{UCL} = \bar{R}D_4, \qquad \text{línea central} = \bar{R}, \qquad \text{LCL} = \bar{R}D_3,$$

donde las constantes D_4 y D_3 (dependiendo sólo en n) son

$$D_4 = 1 + 3\frac{d_3}{d_2}, \qquad D_3 = 1 - 3\frac{d_3}{d_2}.$$

Las constantes D_4 y D_3 se tabulan en la tabla A.23.

Gráficas $\bar{X}$ y R para variables

Se controla un proceso de fabricación de partes componentes para misiles, con la resistencia a la tensión, en libras por pulgada cuadrada, como característica de desempeño. Se toman muestras de tamaño 5 cada hora y se reportan 25 muestras. Los datos se muestran en la tabla 17.1.

Tabla 17.1: Información muestral sobre datos de resistencia a la tensión

Número de muestra	Observaciones					$\bar{X}_i$	R_i
1	1515	1518	1512	1498	1511	1510.8	20
2	1504	1511	1507	1499	1502	1504.6	12
3	1517	1513	1504	1521	1520	1515.0	17
4	1497	1503	1510	1508	1502	1504.0	13
5	1507	1502	1497	1509	1512	1505.4	15
6	1519	1522	1523	1517	1511	1518.4	12
7	1498	1497	1507	1511	1508	1504.2	14
8	1511	1518	1507	1503	1509	1509.6	15
9	1506	1503	1498	1508	1506	1504.2	10
10	1503	1506	1511	1501	1500	1504.2	11
11	1499	1503	1507	1503	1501	1502.6	8
12	1507	1503	1502	1500	1501	1502.6	7
13	1500	1506	1501	1498	1507	1502.4	9
14	1501	1509	1503	1508	1503	1504.8	8
15	1507	1508	1502	1509	1501	1505.4	8
16	1511	1509	1503	1510	1507	1508.0	8
17	1508	1511	1513	1509	1506	1509.4	7
18	1508	1509	1512	1515	1519	1512.6	11
19	1520	1517	1519	1522	1516	1518.8	6
20	1506	1511	1517	1516	1508	1511.6	11
21	1500	1498	1503	1504	1508	1502.6	10
22	1511	1514	1509	1508	1506	1509.6	8
23	1505	1508	1500	1509	1503	1505.0	9
24	1501	1498	1505	1502	1505	1502.2	7
25	1509	1511	1507	1500	1499	1505.2	12

Como indicamos antes, es importante en un principio establecer condiciones de la variabilidad "bajo control". La línea central calculada de la gráfica R es

$$\bar{R} = \frac{1}{25}\sum_{i=1}^{25} R_i = 10.72.$$

De la tabla A.23 encontramos que para $n = 5$, $D_3 = 0$ y $D_4 = 2.115$. Como resultado, los límites de control para la gráfica R son

$$\text{LCL} = \bar{R}D_3 = (10.72)(0) = 0,$$
$$\text{UCL} = \bar{R}D_4 = (10.72)(2.114) = 22.6621.$$

En la figura 17.3 se muestra la gráfica R. Ninguno de los rangos graficados cae fuera de los límites de control. Como resultado, no hay indicación de una situación fuera de control.

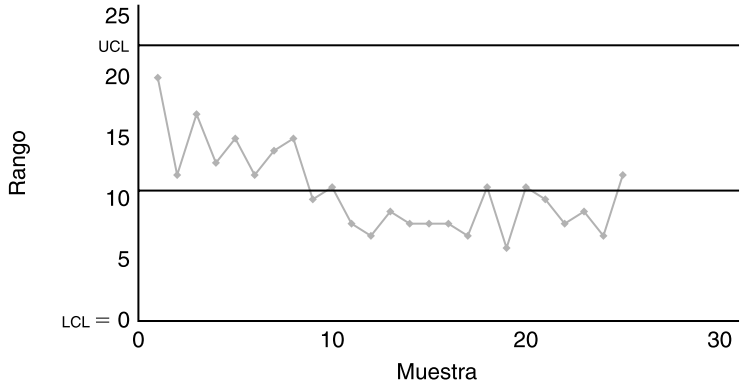


Figura 17.3: Gráfica R para el ejemplo de resistencia a la tensión.

Ahora se puede construir la gráfica $\bar{X}$ para las lecturas de la resistencia a la tensión. La línea central es

$$\bar{\bar{X}} = \frac{1}{25}\sum_{i=1}^{25}\bar{X}_i = 1507.328.$$

Para muestras de tamaño 5, encontramos de la tabla A.23 que $A_2 = 0.577$. Por ello, los límites de control son

$$\text{UCL} = \bar{\bar{X}} + A_2\bar{R} = 1507.328 + (0.577)(10.72) = 1513.5134,$$
$$\text{LCL} = \bar{\bar{X}} - A_2\bar{R} = 1507.328 - (0.577)(10.72) = 1501.1426.$$

En la figura 17.4 se muestra la gráfica $\bar{X}$. Como el lector puede observar, tres valores caen fuera de los límites de control. Como resultado, los límites de control para $\bar{X}$ no se deberían usarse para la línea de control de calidad.

Comentarios adicionales acerca de las gráficas de control para variables

Quizá un proceso parezca estar bajo control y, de hecho, permanecer bajo control durante un periodo largo. ¿Esto necesariamente significa que el proceso opera exitosamente? Un proceso que opera *bajo control* es simplemente uno donde son estables la media y la variabilidad del proceso. Aparentemente, no ocurren cambios serios. "Bajo control" implica que el proceso permanece consistente con variabilidad natural. Las gráficas de control de calidad pueden verse como un método en el que la variabilidad natural inherente rige la amplitud de los límites de control. No hay implicación, sin embargo, de hasta qué punto un proceso bajo control satisface las *especificaciones* predeterminadas que el proceso requiere. Las especificaciones son límites que establece el consumidor. Si la variabilidad natural actual del proceso es

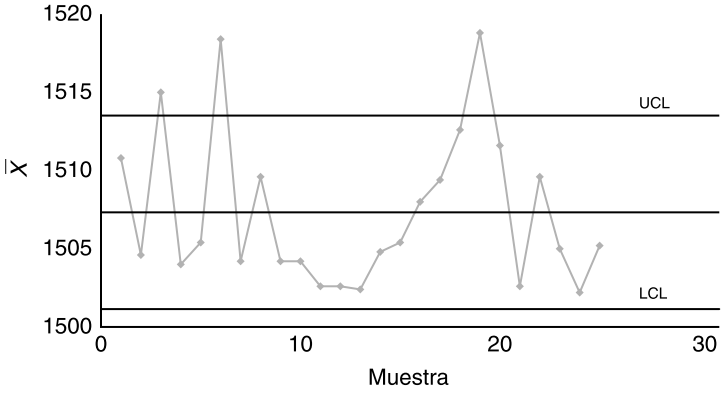


Figura 17.4: Gráfica $\bar{X}$ para el ejemplo de resistencia a la tensión.

mayor que la dictada por la especificación, el proceso no producirá artículos que cumplan las especificaciones con una frecuencia alta, aunque el proceso sea estable y esté bajo control.

Aludimos a la suposición de normalidad sobre las observaciones individuales en una gráfica de control de variables. Para la gráfica $\bar{X}$, si las observaciones individuales son normales, el estadístico $\bar{X}$ es normal. Como resultado, el analista de control de calidad en este caso tiene control sobre la probabilidad de un error tipo I. Si las X individuales no son normales, $\bar{X}$ es aproximadamente normal y por ello existe un control aproximado sobre la probabilidad de un error tipo I, para el caso en el que se conoce σ. Sin embargo, el uso del método del rango para estimar la desviación estándar también depende de la suposición de normalidad. Estudios con respecto a la robustez de la gráfica $\bar{X}$ para desviaciones de la normalidad indican que para las muestras de tamaño $k \geq 4$, la gráfica $\bar{X}$ tiene como resultado un riesgo α cercano al anunciado (véase el trabajo de Montgomery, 2000, y Schilling y Nelson, 1976, en la bibliografía). Indicamos antes que la aproximación $\pm k\sigma_R$ a la gráfica R es una cuestión de conveniencia y tradición. Aun si la distribución de observaciones individuales es normal, la distribución de R no es normal. De hecho, la distribución de R no es ni siquiera simétrica. Los límites de control simétricos $\pm k\sigma_R$ sólo dan una aproximación al riesgo α y, en algunos casos, la aproximación no es particularmente buena.

Elección del tamaño de la muestra (función característica de operación) en el caso de la gráfica $\bar{X}$

Los científicos e ingenieros que tratan con el control de calidad a menudo se refieren a los factores que afectan el *diseño de la gráfica de control*. Los componentes que determinan el diseño de la gráfica incluyen el tamaño de la muestra que se toma en cada subgrupo, la amplitud de los límites de control y la frecuencia del muestreo. Todos estos factores dependen en gran medida de consideraciones económicas y prácticas. La frecuencia de muestreo evidentemente depende del costo del muestreo y del costo que ocurre si el proceso continúa fuera de control durante un periodo largo. Estos mismos factores afectan la amplitud de la región "bajo control". El costo asociado con

la investigación y la búsqueda de causas imputables tiene un impacto sobre la amplitud de la región y sobre la frecuencia de muestreo. Se dedica una cantidad considerable de atención al diseño óptimo de gráficas de control y no se darán aquí detalles más extensos. Se remite al lector al trabajo de Montgomery (2000), que se cita en la bibliografía para un excelente recuento histórico de gran parte de esta investigación.

La elección del tamaño muestral y la frecuencia de muestreo requiere equilibrar los recursos disponibles para estos dos esfuerzos. En muchos casos, el analista quizá necesite hacer cambios en la estrategia hasta que se logre el equilibrio adecuado. El analista siempre debe estar consciente de que si el costo de producción de artículos no adecuados es grande, una alta frecuencia de muestreo con tamaño muestral relativamente pequeño sería una estrategia adecuada.

Se deben tomar en cuenta muchos factores en la elección de un tamaño muestral. En la ilustración y el análisis enfatizamos el uso de $n = 4$, 5 o 6. Estos valores se consideran relativamente pequeños para problemas generales en inferencia estadística, pero quizá tamaños muestrales apropiados para control de calidad. Una justificación, por supuesto, es que el control de calidad es un proceso continuo y los resultados producidos por una muestra o un conjunto de unidades estarán seguidos por resultados de muchas más. Así, el tamaño muestral "efectivo" de todo el esfuerzo de control de calidad es muchas veces mayor que el que se utiliza en un subgrupo. Por lo general, se considera más efectivo *muestrear con frecuencia* con un tamaño muestral pequeño.

El analista puede utilizar la noción de *poder* de una prueba para obtener alguna idea de la efectividad del tamaño muestral que se elige. Esto es importante en particular debido a que los tamaños muestrales pequeños, por lo general, se utilizan en cada subgrupo. Véase los capítulos 10 y 13 para un análisis de la potencia de pruebas formales sobre medias y del análisis de varianza. Aunque las pruebas formales de hipótesis en realidad no se realizan en el control de calidad, se puede tratar la información muestral como si la estrategia en cada subgrupo fuera a probar una hipótesis, ya sea sobre la media poblacional μ o sobre la desviación estándar σ. Es de interés la *probabilidad de detección* de una condición fuera de control para una muestra dada y, quizá más importante, el número esperado de corridas que se requieren para la detección. La probabilidad de detección de una condición fuera de control específica corresponde al poder de una prueba. No es nuestra intención mostrar el desarrollo de la potencia para todos los tipos de gráficas de control que aquí se presentan, sino más bien mostrar el desarrollo de la gráfica $\bar{X}$ y presentar los resultados de potencia para la gráfica R.

Considere la gráfica $\bar{X}$ para σ conocida. Suponga que el estado bajo control tiene $\mu = \mu_0$. Un estudio del papel del tamaño muestral del subgrupo es equivalente a investigar el riesgo β, es decir, la probabilidad de que un valor $\bar{X}$ permanezca dentro de los límites de control dado que, en realidad, ocurre un corrimiento en la media. Suponga que la forma que toma el corrimiento es

$$\mu = \mu_0 + r\sigma.$$

De nuevo, al utilizar la normalidad de $\bar{X}$, tenemos

$$\beta = P\{\text{LCL} \leq \bar{X} \leq \text{UCL} \mid \mu = \mu_0 + r\sigma\}.$$

Para el caso de límites $k\sigma$,

$$\text{LCL} = \mu_0 - \frac{k\sigma}{\sqrt{n}}, \qquad \text{y} \qquad \text{UCL} = \mu_0 + \frac{k\sigma}{\sqrt{n}}.$$

Como resultado, si denotamos con Z la variable aleatoria normal estándar

$$\begin{aligned}\beta &= P\left\{Z < \left[\frac{\mu_0 + k\sigma/\sqrt{n} - \mu}{\sigma/\sqrt{n}}\right]\right\} - P\left\{Z < \left[\frac{\mu_0 - k\sigma/\sqrt{n} - \mu}{\sigma/\sqrt{n}}\right]\right\} \\ &= P\left\{Z < \left[\frac{\mu_0 + k\sigma/\sqrt{n} - (\mu + r\sigma)}{\sigma/\sqrt{n}}\right]\right\} - P\left\{Z < \left[\frac{\mu_0 - k\sigma/\sqrt{n} - (\mu + r\sigma)}{\sigma/\sqrt{n}}\right]\right\} \\ &= P(Z < k - r\sqrt{n}) - P(Z < -k - r\sqrt{n}).\end{aligned}$$

Observe el papel de n, r y k en la expresión para el riesgo β. La probabilidad de no detectar un corrimiento específico con claridad aumenta con un aumento de k, como se esperaba. β disminuye con un aumento de r, la magnitud del corrimiento y disminuye con un aumento en el tamaño muestral n.

Se debería enfatizar que la expresión anterior tiene como resultado el riesgo β (probabilidad de un error tipo II) para el caso de una *sola muestra*. Por ejemplo, suponga que en el caso de una muestra de tamaño 4, ocurre un corrimiento de σ en la media. La probabilidad de detectar el corrimiento (poder) *en la primera muestra a continuación del corrimiento* es (suponga límites 3σ):

$$1 - \beta = 1 - [P(Z < 1) - P(Z < -5)] = 0.1587.$$

Por otro lado, la probabilidad de detectar un corrimiento de 2σ es

$$1 - \beta = 1 - [P(Z < -1) - P(Z < -7)] = 0.8413.$$

Los resultados anteriores ilustran una probabilidad bastante modesta de detectar un corrimiento de magnitud σ y una probabilidad bastante alta de detectar un corrimiento de magnitud 2σ. La presentación completa de cómo se comportan los límites de control, digamos, 3σ para la gráfica $\bar{X}$ que aquí se describe se muestra en la figura 17.5. En vez de graficar el poder, se da una gráfica de β contra r, donde el corrimiento en la media tiene magnitud $r\sigma$. Por supuesto, los tamaños muestrales de $n = 4$, 5, 6 tienen como resultado una probabilidad pequeña de detectar un corrimiento de 1.0σ o incluso 1.5σ sobre la primera muestra después del corrimiento.

Pero si el muestreo se realiza con frecuencia, la probabilidad quizá no sea tan importante como el número promedio o esperado de corridas que se requiere antes de la detección del corrimiento. Una detección rápida es importante y en realidad es posible, aunque la probabilidad de detección sobre la primera muestra no sea alta. Resulta que las gráficas $\bar{X}$ con estas pequeñas muestras tendrá como resultado una detección relativamente rápida. Si β es la probabilidad de no detectar un corrimiento sobre la primera muestra que sigue al corrimiento, entonces la probabilidad de detectar el corrimiento sobre la muestra s-ésima después de éste es (si suponemos muestras independientes):

$$P_s = (1 - \beta)\beta^{s-1}.$$

El lector debe reconocer ésta como una aplicación de la distribución geométrica. El valor promedio o esperado del número de muestras que se requieren para la detección es

$$\sum_{s=1}^{\infty} s\beta^{s-1}(1 - \beta) = \frac{1}{1 - \beta}.$$

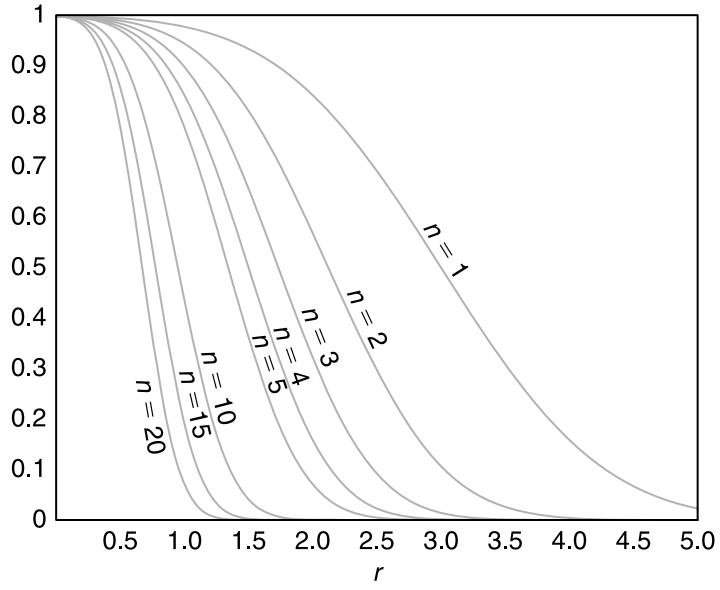


Figura 17.5: Curvas características de operación para la gráfica $\bar{X}$ con límites 3σ. Aquí, β es el error de probabilidad tipo II sobre la primera muestra, después de un corrimiento en la media de $r\sigma$.

Así el número esperado de muestras que se requieren para detectar el corrimiento en la media es el *recíproco del poder* (es decir, la probabilidad de detección de la primera muestra después del corrimiento).

Ejemplo 17.1: Para el analista de control de calidad en cierto esfuerzo de control de calidad es importante detectar con rapidez corrimientos en la media de $\pm\ \sigma$ mientras se utiliza una gráfica de control 3σ con un tamaño muestral $n = 4$. El número esperado de muestras que se requieren después del corrimiento para la detección del estado fuera de control puede ser una ayuda en la evaluación del procedimiento de control de calidad.

De la figura 17.5, para $n = 4$ y $r = 1$, se puede ver que $\beta \approx 0.84$. Si denotamos con s el número de muestras que se requieren para detectar el corrimiento, la media de s es

$$E(s) = \frac{1}{1-\beta} = \frac{1}{0.16} = 6.25.$$

De esta manera, en promedio, se requieren seis subgrupos antes de la detección de un corrimiento de $\pm\ \sigma$.

Elección del tamaño muestral para la gráfica *R*

La curva co de la gráfica R se muestra en la figura 17.6. Como la gráfica R se utiliza para control de la desviación estándar del proceso, el riesgo β se grafica como función de la desviación estándar bajo control, σ_0, y la desviación estándar después

de que el proceso queda fuera de control. La última desviación estándar se denotará con σ_1. Sea

$$\lambda = \frac{\sigma_1}{\sigma_0}.$$

Se grafica β contra λ para varios tamaños muestrales.

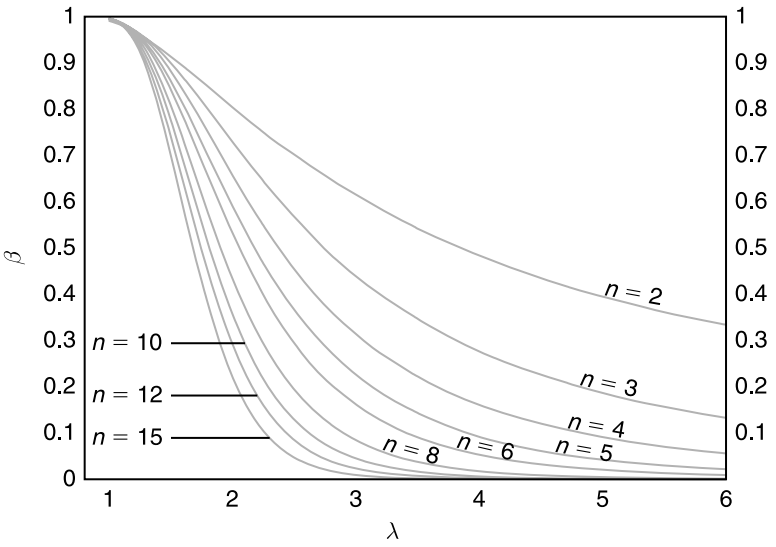


Figura 17.6: Curva característica operativa para las gráficas R con límites de 3σ.

Gráficas $\bar{X}$ y S para variables

Es natural para el estudiante de estadística anticipar el uso de la varianza muestral en la gráfica $\bar{X}$ y en una gráfica para control de la variabilidad. El rango es eficiente como estimador para σ, pero esta eficiencia disminuye conforme el tamaño de la muestra se hace más grande. Para n tan grande como 10, se debe utilizar la estadística familiar

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

en la gráfica de control, tanto para la media como para la variabilidad. El lector debe recordar del capítulo 9 que S^2 es un estimador insesgado para σ^2 pero que S no es insesgado para σ. Es una costumbre corregir S para sesgos en aplicaciones de la gráfica de control. Sabemos, en general, que

$$E(S) \neq \sigma.$$

En el caso en que las X_i sean independientes, y estén distribuidas de forma normal con media μ y varianza σ^2,

$$E(S) = c_4\sigma, \qquad \text{donde} \qquad c_4 = \left(\frac{2}{n-1}\right)^{1/2}\frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}$$

y $\Gamma(\cdot)$ se refiere a la función gamma (véase el capítulo 6). Por ejemplo, para $n = 50$, $c_4 = 3/8\sqrt{2\pi}$. Además, la varianza del estimador S es

$$Var(S) = \sigma^2(1 - c_4^2).$$

Establecimos las propiedades de S que nos permitirán escribir límites de control para $\bar{X}$ y S. Para construir una estructura adecuada, comenzamos con la suposición de que se conoce σ. Después presentamos la estimación de σ a partir de un conjunto de muestras preliminar.

Si se grafica el estadístico S, los parámetros evidentes de la gráfica de control son

$$\text{UCL} = c_4\sigma + 3\sigma\sqrt{1 - c_4^2}, \quad \text{línea central} = c_4\sigma, \quad \text{LCL} = c_4\sigma - 3\sigma\sqrt{1 - c_4^2}.$$

Como de costumbre, los límites de control se definen de manera más sucinta a través del uso de constantes tabuladas. Sean

$$B_5 = c_4 - 3\sqrt{1 - c_4^2}, \qquad B_6 = c_4 + 3\sqrt{1 - c_4^2},$$

entonces, tenemos

$$\text{UCL} = B_6\sigma, \qquad \text{línea central} = c_4\sigma, \qquad \text{LCL} = B_5\sigma.$$

En la tabla A.23 se tabulan los valores de B_5 y B_6 para varios tamaños muestrales.

Ahora, por supuesto, los límites de control anteriores sirven como base para el desarrollo de los parámetros de control de calidad, para la situación que con más frecuencia se observa en la práctica, a saber, en la que se desconoce σ. Debemos suponer una vez más que se toma un conjunto de *muestras base* o muestras preliminares para producir una estimación de σ durante lo que se supone como periodo "bajo control". Las desviaciones estándar muestrales S_1, $S_2, \ldots,$ S_m se obtienen a partir de muestras que son cada una de tamaño n. A menudo se utiliza un estimador insesgado del tipo

$$\frac{\bar{S}}{c_4} = \left(\frac{1}{m}\sum_{i=1}^{m} S_i\right)/c_4$$

$$\bar{S},$$

para σ. Aquí, por supuesto, $\bar{S}$, el valor promedio de la desviación estándar muestral en la muestra preliminar, es la línea central lógica en la gráfica de control para el control de la variabilidad. Los límites de control superior e inferior son estimadores insesgados de los límites de control adecuados para el caso donde se conoce σ. Como

$$E\left(\frac{\bar{S}}{c_4}\right) = \sigma,$$

el estadístico S es una línea central apropiada (como estimador insesgado de $c_4\sigma$) y las cantidades

$$\bar{S} - 3\frac{\bar{S}}{c_4}\sqrt{1 - c_4^2} \qquad \text{y} \qquad \bar{S} + 3\frac{\bar{S}}{c_4}\sqrt{1 - c_4^2}$$

son los límites de control 3σ inferior y superior apropiados, respectivamente. Como resultado, la línea central y los límites para la gráfica S para control de variabilidad son

$$\text{LCL} = B_3\bar{S}, \qquad \text{línea central} = \bar{S}, \qquad \text{UCL} = B_4\bar{S},$$

donde

$$B_3 = 1 - \frac{3}{c_4}\sqrt{1 - c_4^2}, \qquad B_4 = 1 + \frac{3}{c_4}\sqrt{1 - c_4^2}.$$

Las constantes B_3 y B_4 aparecen en la tabla A.23.

Ahora podemos escribir los parámetros de la gráfica $\bar{X}$ correspondiente que implican el uso de la desviación estándar muestral. Supongamos que S y $\bar{X}$ están disponibles de la muestra preliminar base. La línea central continúa siendo $\bar{X}$ y los límites 3σ son simplemente de la forma $\overline{\overline{X}} \pm 3\hat{\sigma}/\sqrt{n}$, donde $\hat{\sigma}$ es un estimador insesgado. Simplemente proporcionamos $\bar{S}/c_4$ como un estimador de σ, y de esta forma tenemos

$$\text{LCL} = \bar{\bar{X}} - A_3\bar{S}, \qquad \text{línea central} = \bar{\bar{X}}, \qquad \text{UCL} = \bar{\bar{X}} + A_3\bar{S},$$

donde

$$A_3 = \frac{3}{c_4\sqrt{n}}.$$

En la tabla A.23 aparece la constante A_3 para varios tamaños muestrales.

Ejemplo 17.2: Se producen contenedores mediante un proceso donde el volumen de los contenedores se sujeta a un control de calidad. Se utilizan 25 muestras de tamaño 5, cada una para establecer los parámetros de control de calidad. En la tabla 17.2 se documenta la información de estas muestras.

De la tabla A.23, $B_3 = 0$, $B_4 = 2.089$, $A_3 = 1.427$. Como resultado, los límites de control para $\bar{X}$ están dados por

$$\bar{\bar{X}} + A_3\bar{S} = 62.3771, \qquad \bar{\bar{X}} - A_3\bar{S} = 62.2740,$$

y los límites de control para la gráfica S son

$$\text{LCL} = B_3\bar{S} = 0, \qquad \text{UCL} = B_4\bar{S} = 0.0754.$$

Las figuras 17.7 y 17.8 muestran las gráficas de control $\bar{X}$ y S, respectivamente, para este ejemplo. En las gráficas se representa la información para las 25 muestras en el conjunto de datos preliminar. Parece que el control se establece después de las primeras muestras.

17.5 Gráficas de control para atributos

Como indicamos antes en este capítulo, muchas aplicaciones industriales de control de calidad requieren que la característica de calidad indique sólo la afirmación de que el artículo "se adapta". En otras palabras, no hay la medición continua que es crucial para el desempeño del artículo. Una ilustración evidente de este tipo de muestreo, que se llama **muestreo por atributos**, es el desempeño de una bombilla de luz: que funciona o no de manera satisfactoria. El artículo es **defectuoso o no defectuoso**. Las piezas metálicas fabricadas pueden tener deformaciones. Los contenedores de una línea de producción pueden tener fugas. En ambos casos, un artículo defectuoso impide su uso por parte del consumidor. La gráfica de control estándar para esta situación es la gráfica p, o gráfica *para la fracción de defectuosos*. Como se podría esperar, la distribución de probabilidad que interviene es la distribución binomial. Se remite al lector al capítulo 5 para información básica de la distribución binomial.

Tabla 17.2: Volumen de muestras de contenedores para 25 muestras en una muestra preliminar (en centímetros cúbicos)

Muestra	**Observaciones**					$\bar{X}_i$	S_i
1	62.255	62.301	62.289	62.289	62.311	62.269	0.0495
2	62.187	62.225	62.337	62.297	62.307	62.271	0.0622
3	62.421	62.377	62.257	62.295	62.222	62.314	0.0829
4	62.301	62.315	62.293	62.317	62.409	62.327	0.0469
5	62.400	62.375	62.295	62.272	62.372	62.343	0.0558
6	62.372	62.275	62.315	62.372	62.302	62.327	0.0434
7	62.297	62.303	62.337	62.392	62.344	62.335	0.0381
8	62.325	62.362	62.351	62.371	62.397	62.361	0.0264
9	62.327	62.297	62.318	62.342	62.318	62.320	0.0163
10	62.297	62.325	62.303	62.307	62.333	62.313	0.0153
11	62.315	62.366	62.308	62.318	62.319	62.325	0.0232
12	62.297	62.322	62.344	62.342	62.313	62.324	0.0198
13	62.375	62.287	62.362	62.319	62.382	62.345	0.0406
14	62.317	62.321	62.297	62.372	62.319	62.325	0.0279
15	62.299	62.307	62.383	62.341	62.394	62.345	0.0431
16	62.308	62.319	62.344	62.319	62.378	62.334	0.0281
17	62.319	62.357	62.277	62.315	62.295	62.313	0.0300
18	62.333	62.362	62.292	62.327	62.314	62.326	0.0257
19	62.313	62.387	62.315	62.318	62.341	62.335	0.0313
20	62.375	62.321	62.354	62.342	62.375	62.353	0.0230
21	62.399	62.308	62.292	62.372	62.299	62.334	0.0483
22	62.309	62.403	62.318	62.295	62.317	62.328	0.0427
23	62.293	62.293	62.342	62.315	62.349	62.318	0.0264
24	62.388	62.308	62.315	62.392	62.303	62.341	0.0448
25	62.328	62.318	62.317	62.295	62.319	62.314	0.0111
						$\bar{\bar{X}} = 62.3256$	
						$\bar{S} = 0.0361$	

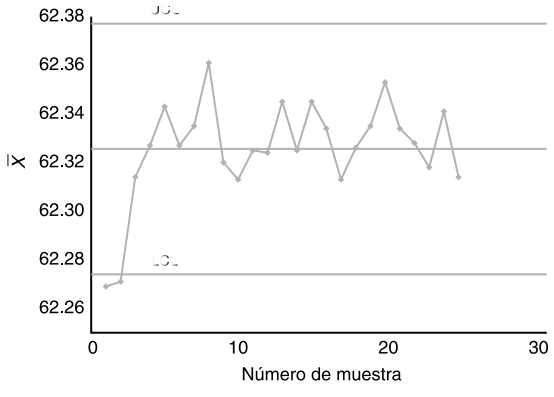


Figura 17.7: Gráfica $\bar{X}$ con límites de control establecidos por los datos del ejemplo 17.2.

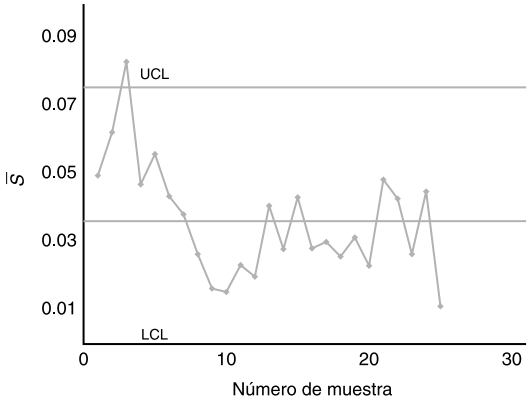


Figura 17.8: Gráfica S con límites de control establecidos por los datos del ejemplo 17.2.

Gráfica p para la fracción de defectuosos

Cualquier artículo fabricado puede tener varias características que son importantes y lo debe examinar un inspector. Sin embargo, todo el desarrollo se enfoca aquí a una sola característica. Suponga que para todos los artículos la probabilidad de un defectuoso es p, y que todos los artículos se producen de forma independiente. Entonces, en una muestra aleatoria de n artículos producidos, con X como el número de artículos defectuosos, tenemos

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

Como se podría sospechar, la media y varianza de la variable aleatoria binomial jugarán un papel importante en el desarrollo de la gráfica de control. El lector debería recordar que

$$E(X) = np \qquad \text{y} \qquad Var(X) = np(1-p).$$

Un estimador insesgado de p es la **fracción de defectuosos** o la **proporción de defectuosos,** $\hat{p}$, donde

$$\hat{p} = \frac{\text{número de defectuosos en la muestra de tamaño } n}{n}.$$

Como en el caso de las gráficas de control de variables, las propiedades de distribución de p son importantes en el desarrollo de la gráfica de control. Sabemos que

$$E(\hat{p}) = p, \qquad Var(\hat{p}) = \frac{p(1-p)}{n}.$$

Aquí aplicamos los mismos principios 3σ que utilizamos para las gráficas de variables. Supongamos inicialmente que se conoce p. La estructura, entonces, de las gráficas de control implica el uso de límites 3σ con

$$\hat{\sigma} = \sqrt{\frac{p(1-p)}{n}}.$$

De esta manera, los límites son

$$\text{LCL} = p - 3\sqrt{\frac{p(1-p)}{n}}, \qquad \text{UCL} = p + 3\sqrt{\frac{p(1-p)}{n}},$$

con el proceso considerado bajo control cuando los valores $\hat{p}$ de la muestra yacen dentro de los límites de control.

En general, por supuesto, no se conoce el valor de p y se debe estimar a partir de un conjunto base de muestras de forma muy similar al caso de μ y σ en las gráficas de variables. Suponga que hay m muestras preliminares de tamaño n. Para una muestra dada, cada una de las n observaciones se reporta como "defectuosa" o "no defectuosa". El estimador insesgado evidente para p que se utiliza en la gráfica de control es

$$\bar{p} = \frac{1}{m} \sum_{i=1}^{m} \hat{p}_i,$$

donde $\bar{p}$ es la proporción de defectuosos en la i-ésima muestra. Como resultado, los límites de control son

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \qquad \text{línea central} = \bar{p}, \qquad \text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

Ejemplo 17.3: Considere los datos que se muestran en la tabla 17.3 sobre el número de componentes electrónicos defectuosos en muestras de tamaño 50. Se tomaron 20 muestras con la finalidad de establecer valores preliminares de la gráfica de control. Las gráficas de control determinadas por este periodo preliminar tendrán una línea central $\bar{p} = 0.088$ y límites de control

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{50}} = -0.0322, \qquad \text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{50}} = 0.2082.$$

Tabla 17.3: Datos para el ejemplo 17.3 para establecer límites de control para gráficas p, muestras de tamaño 50

Muestra	**Número de componentes defectuosos**	**Fracción de defectuosos $\hat{p}_i$**
1	8	0.16
2	6	0.12
3	5	0.10
4	7	0.14
5	2	0.04
6	5	0.10
7	3	0.06
8	8	0.16
9	4	0.08
10	4	0.08
11	3	0.06
12	1	0.02
13	5	0.10
14	4	0.08
15	4	0.08
16	2	0.04
17	3	0.06
18	5	0.10
19	6	0.12
20	3	0.06
		$\bar{p} = 0.088$

Evidentemente, con un valor calculado negativo, el LCL se ajusta a cero. A partir de los valores de los límites de control se hace evidente que el proceso está bajo control durante este periodo preliminar.

Selección del tamaño muestral para la gráfica *p*

La elección del tamaño muestral para la gráfica p para atributos incluye los mismos tipos generales de consideraciones que los de la gráfica para variables. Se requiere un tamaño muestral que sea suficientemente grande para tener una probabilidad alta de detección de una condición fuera de control cuando, de hecho, ocurre un cambio específico en p. No hay un *mejor método* para la elección del tamaño de la muestra. Sin embargo, una aproximación razonable, sugerida por Duncan (véase la bibliografía), consiste en elegir n de modo que haya una probabilidad de 0.5 de detectar un corrimiento de un monto particular en p. La solución que resulta para n es bastante simple. Suponga que se aplica la aproximación normal a la distribución binomial. Deseamos, bajo la condición de que p tiene un corrimiento a, digamos, $p_1 > p_0$, que

$$P(\hat{p} \geq \text{UCL}) = P\left[Z \geq \frac{\text{UCL} - p_1}{\sqrt{p_1(1-p_1)/n}}\right] = 0.5.$$

Como $P(Z > 0) = 0.5$, hacemos

$$\frac{\text{UCL} - p_1}{\sqrt{p_1(1-p_1)/n}} = 0.$$

Al sustituir,

$$p + 3\sqrt{\frac{p(1-p)}{n}} = \text{UCL},$$

tenemos

$$(p - p_1) + 3\sqrt{\frac{p(1-p)}{n}} = 0.$$

Ahora podemos despejar n, el tamaño de cada muestra:

$$n = \frac{9}{\Delta^2}p(1-p),$$

donde, por supuesto, Δ es el "corrimiento" en el valor de p, y p es la probabilidad de un defectuoso sobre la que se basan los límites de control. Sin embargo, si las gráficas de control se basan en límites $k\sigma$ entonces

$$n = \frac{k^2}{\Delta^2}p(1-p).$$

Ejemplo 17.4: Suponga que se diseña una gráfica de control de calidad de atributos con un valor de $p = 0.01$ para la probabilidad bajo control de un defectuoso. ¿Cuál es el tamaño de la muestra por subgrupo que produce una probabilidad de 0.5 de que se detecte un proceso que se corre a $p = p_1 = 0.05$? La gráfica p resultante incluirá límites 3σ.

Solución: Aquí tenemos $\Delta = 0.04$. El tamaño adecuado de la muestra es

$$n = \frac{9}{(0.04)^2}(0.01)(0.99) = 55.68 \approx 56.$$

Gráficas de control para defectuosos (uso del modelo de Poisson)

En el desarrollo anterior supusimos que el artículo bajo consideración es uno que es defectuoso (es decir, no funcional) o no defectuoso. En el último caso es funcional y, por ello, aceptable para el consumidor. En muchas situaciones este enfoque de "defectuoso o no" es demasiado simplista. Las unidades pueden contener defectos o no cumplir con la norma; pero aún así funcionar bastante bien para el consumidor. En realidad, en este caso, sería importante ejercer control sobre el *número de defectos* o *número de diferencias*. Este tipo de esfuerzo de control de calidad encuentra aplicación cuando las unidades son no simplistas o quizá grandes. Por ejemplo, el número de defectos puede ser bastante útil como el objeto de control cuando el artículo o unidad es, digamos, una computadora personal. Otro ejemplo es una unidad definida por 50 pies de tubería fabricada, donde el número de soldaduras defectuosas es el objeto del control de calidad, el número de defectos de 50 pies de tejido para alfombras fabricado o el número de "burbujas" en una hoja grande de vidrio fabricado.

Es claro a partir de lo que describimos aquí, que la distribución binomial no es apropiada. El número total de diferencias en una unidad o el número promedio por unidad se puede usar como la medida para la gráfica de control. Bastante a menudo se supone que el número de diferencias en una muestra de artículos sigue la distribución de Poisson. Este tipo de gráfica con frecuencia se llama **gráfica *C***.

Suponga que el número de defectos X en una unidad de producto sigue la distribución de Poisson con parámetro λ. (Aquí $t = 1$ para el modelo de Poisson.) Recuerde que para la distribución de Poisson,

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots.$$

Aquí, la variable aleatoria X es el número de diferencias. En el capítulo 5 vimos que la media y la varianza de la variable aleatoria de Poisson son ambas λ. De esta forma, si la gráfica de control de calidad se estructurara de acuerdo con los límites 3σ acostumbrados, tendríamos, para λ conocida,

$$\text{UCL} = \lambda + 3\sqrt{\lambda}, \qquad \text{línea central} = \lambda, \qquad \text{LCL} = \lambda - 3\sqrt{\lambda}.$$

Como de costumbre, λ a menudo debe provenir de un estimador de los datos. Una estimación insesgada de λ es el número *promedio* de diferencias por muestra. Denote esta estimación con $\hat{\lambda}$. Así la gráfica de control tiene los límites

$$\text{UCL} = \hat{\lambda} + 3\sqrt{\hat{\lambda}}, \qquad \text{línea central} = \hat{\lambda}, \qquad \text{LCL} = \hat{\lambda} - 3\sqrt{\hat{\lambda}}.$$

Ejemplo 17.5: La tabla 17.4 representa el número de defectos en 20 muestras sucesivas de rollos de hoja metálica, cada una de 100 pies de longitud. Se debe desarrollar una gráfica de control a partir de estos datos preliminares con la finalidad de controlar el número de defectos en tales muestras. La estimación del parámetro de Poisson λ está dada por $\hat{\lambda} = 5.95$. Como resultado, los límites de control sugeridos por estos datos preliminares son:

$$\text{UCL} = \hat{\lambda} + 3\sqrt{\hat{\lambda}} = 13.2678 \qquad \text{y} \qquad \text{LCL} = \hat{\lambda} - 3\sqrt{\hat{\lambda}} = -1.3678,$$

con LCL igualada a cero.

Tabla 17.4: Datos para el ejemplo 17.5; el control incluye el número de defectos en un rollo de hoja metálica

Número de muestra	Número de defectos	Número de muestra	Número de defectos
1	8	11	3
2	7	12	7
3	5	13	5
4	4	14	9
5	4	15	7
6	7	16	7
7	6	17	8
8	4	18	6
9	5	19	7
10	6	20	4
			Promedio 5.95

Tabla 17.5: Datos adicionales del proceso de producción del ejemplo 17.5

Número de muestra	Número de defectos	Número de muestra	Número de defectos
1	3	11	7
2	5	12	5
3	8	13	9
4	5	14	4
5	8	15	6
6	4	16	5
7	3	17	3
8	6	18	2
9	5	19	1
10	2	20	6

La figura 17.9 muestra una gráfica de los datos preliminares con los límites de control.

La tabla 17.5 muestra datos adicionales tomados del proceso de producción. Para cada muestra, se investiga la unidad en la que se basa la gráfica; a saber, 100 pies del metal. Se revela la información de 20 muestras. La figura 17.10 muestra una gráfica de los datos adicionales de producción. Es claro que el proceso está bajo control, al menos a lo largo del periodo en el que se toman los datos.

En el ejemplo 17.5, dejamos muy claro que la unidad de muestreo o de inspección es, a saber, 100 pies de metal. En muchos casos donde el artículo es específico (por ejemplo, una computadora personal o un tipo específico de dispositivo electrónico), la unidad de inspección puede ser un *conjunto de artículos*. Por ejemplo, el analista puede decidir utilizar 10 computadoras en cada subgrupo y de esta forma observar un conteo del número total de defectos que se encuentran. De esta forma la muestra preliminar para la construcción de la gráfica de control incluiría el uso de varias muestras, cada una de 10 computadoras. La elección del tamaño muestral puede depender de muchos factores. A menudo, se puede querer un tamaño muestral que asegure un LCL positivo.

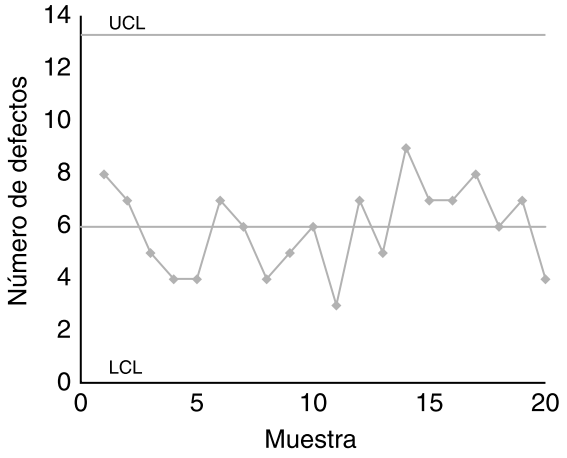


Figura 17.9: Datos preliminares representados en la gráfica de control para el ejemplo 17.5.

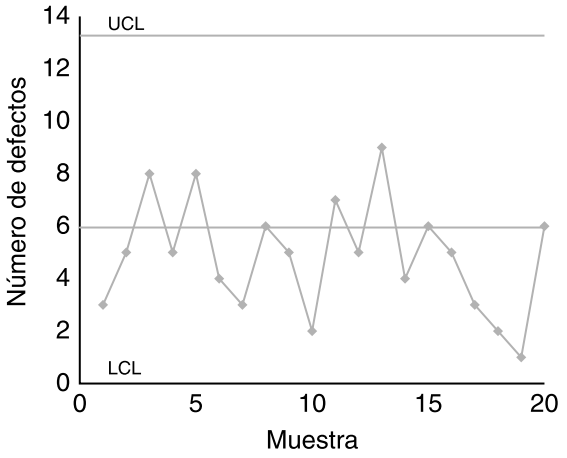


Figura 17.10: Datos adicionales de producción para el ejemplo 17.5.

El analista puede desear utilizar el número promedio de defectos por unidad de muestreo como la medida básica en la gráfica de control. Por ejemplo, para el caso de la computadora personal, sea la variable aleatoria el número total de defectos

$$U = \frac{\text{número total de defectos}}{n}$$

que se mide para cada muestra de, digamos, $n = 10$. Se puede utilizar el método de las funciones generadoras de momento para demostrar que U es una variable aleatoria de Poisson (véase el ejercicio de repaso 17.1), si suponemos que el número de defectos por unidad de muestreo es de Poisson con parámetro λ. De esta manera, la gráfica de control para esta situación se caracteriza por lo siguiente:

$$\text{UCL} = \bar{U} + 3\sqrt{\frac{\bar{U}}{n}}, \qquad \text{línea central} = \bar{U}, \qquad \text{LCL} = \bar{U} - 3\sqrt{\frac{\bar{U}}{n}}.$$

Aquí, por supuesto, $\bar{U}$ es el promedio de los valores U en el conjunto de datos preliminares o base. El término $\bar{U}/n$ se deriva del resultado que

$$E(U) = \lambda, \qquad Var(U) = \frac{\lambda}{n},$$

y por ello $\bar{U}$ es un estimador insesgado de $E(U) = \lambda$ y $\bar{U}/n$ es un estimador insesgado de $Var(U) = \lambda/n$. Este tipo de gráfica de control a menudo se denomina **gráfica *U***.

En todo el desarrollo de esta sección basamos nuestra producción de las gráficas de control en el modelo de probabilidad de Poisson. Este modelo se utiliza en combinación con el concepto 3σ. Como explicamos antes en este capítulo, la noción de límites 3σ tiene sus raíces en la aproximación normal, aunque muchos usuarios consideran que el concepto funciona bien como herramienta pragmática, incluso si la normalidad no es siquiera aproximadamente correcta. La dificultad, por supuesto, es que en ausencia de normalidad, no se puede controlar la probabilidad de una especificación incorrecta de un estado fuera de control. En el caso del modelo de Poisson, cuando λ es pequeña la distribución es bastante asimétrica, condición que llega a producir resultados indeseables si se conserva la aproximación 3σ.

17.6 Gráficas de control de cusum

La desventaja con las gráficas de control tipo Shewhart, que se desarrollan e ilustran en las secciones anteriores, radica en su incapacidad para detectar pequeños cambios en la media. Un mecanismo de control de calidad que recibe una considerable atención en la literatura estadística y amplio uso en la industria es la **gráfica de suma acumulada (cusum)**. El método para la gráfica de suma acumulada es sencillo y su atractivo es que es intuitivo. Debe ser evidente para el lector por qué responde mejor a pequeños cambios en la media. Considere una gráfica de control para la media con un nivel de referencia establecido en el valor W. Considere las observaciones particulares $X_1, X_2, \ldots, X_r$. Las primeras r sumas acumuladas son

$$\begin{aligned} S_1 &= X_1 - W \\ S_2 &= S_1 + (X_2 - W) \\ S_3 &= S_2 + (X_3 - W) \\ &\vdots \\ S_r &= S_{r-1} + (X_r - W). \end{aligned}$$

Es claro que la suma acumulada es simplemente la acumulación de las diferencias del nivel de referencia. Es decir,

$$S_k = \sum_{i=1}^{k} (X_i - W), \qquad k = 1, 2, \ldots$$

La gráfica de suma acumulada es, entonces, una gráfica de S_k contra el tiempo.

Suponga que consideramos que el nivel de referencia w es un valor aceptable de la media μ. Claramente, si no hay corrimiento en μ, la gráfica de suma acumulada debería ser aproximadamente horizontal, con algunas fluctuaciones menores balanceadas alrededor de cero. Ahora, si sólo hay un cambio moderado en la media, debe resultar un cambio relativamente grande en la *pendiente* de la gráfica de suma acumulada, pues cada nueva observación tiene una oportunidad de contribuir con un corrimiento y la medida que se grafica se acumula a estos corrimientos. Por supuesto, la señal de que la media está recorrida yace en la naturaleza de la pendiente de la gráfica de suma acumulada. El propósito de la gráfica es detectar cambios que se alejan del nivel de referencia. Una pendiente diferente de cero (en cualquier dirección) representa un cambio a partir del nivel de referencia. Una pendiente positiva indica un aumento en la media por arriba del nivel de referencia; en tanto que una pendiente negativa señala una disminución.

Las gráficas de suma acumulada a menudo se diseñan con un nivel de *calidad aceptable* definido (AQL) y un nivel de *calidad rechazable* (RQL) prestablecido por el usuario. Ambos representan valores de la media. Éstos se pueden ver como si jugaran papeles similares a los de las medias nula y alternativa en la prueba de hipótesis. Considere una situación donde el analista desea detectar un aumento en el valor de la media del proceso. Usaremos la notación μ_0 para AQL y μ_1 para RQL y sea $\mu_1 > \mu_0$. El nivel de referencia se fija ahora en

$$W = \frac{\mu_0 + \mu_1}{2}.$$

Los valores de S_r ($r = 1, 2, \ldots$) tendrán una pendiente negativa si la media del proceso está en μ_0 y una pendiente positiva si la media del proceso está en μ_1.

Regla de decisión para las gráficas cusum

Como indicamos antes, la pendiente de la gráfica de suma acumulada proporciona la señal de acción para el analista de control de calidad. La regla de decisión requiere la acción si, en el r-ésimo periodo de muestreo,

$$d_r > h,$$

donde h es un valor prestablecido que se llama **longitud del intervalo de decisión** y

$$d_r = S_r - \min_{1 \le i \le r-1} S_i.$$

En otras palabras, se toma la acción si los datos revelan que el valor de la suma acumulada real excede en una cantidad específica al valor previo de la suma acumulada más pequeña.

Una modificación en la mecánica que se describió antes permite la facilidad en el empleo del método. Describimos un procedimiento que grafica las sumas acumuladas y calcula las diferencias. Una modificación simple implica graficar las diferencias de manera directa y permitir la verificación contra el intervalo de decisión. La expresión general para d_r es bastante simple. Para el procedimiento de la suma acumulada, donde se detectan aumentos en la media,

$$d_r = \text{máx}[0, d_{r-1} + (X_r - W)].$$

La elección del valor de h es, por supuesto, muy importante. En este libro elegimos no proporcionar los detalles que aparecen en la literatura que trata de esta elección. Se remite al lector a Ewan y Kemp, 1960 (véase la bibliografía), para una exposición más completa. Una consideración importante es la **longitud esperada de la corrida**. Idealmente, la longitud esperada de la corrida es bastante grande bajo $\mu = \mu_0$ y bastante pequeña cuando $\mu = \mu_1$.

Ejercicios de repaso

17.1 Considere $X_1, X_2, \ldots, X_n$, variables aleatorias de Poisson independientes con parámetros $\mu_1, \mu_2, \ldots, \mu_n$. Utilice las propiedades de las funciones generadoras de momento para mostrar que la variable aleatoria $\sum_{i=1}^{n} X_i$ es una variable aleatoria de Poisson con media $\sum_{i=1}^{n} \mu_i$ y varianza $\sum_{i=1}^{n} \mu_i$.

17.2 Considere los siguientes datos tomados en subgrupos de tamaño 5. Los datos contienen 20 promedios, y rangos del diámetro (en milímetros) de una parte componente importante de un motor. Elabore gráficas $\bar{X}$ y R. ¿El proceso parece estar bajo control?

Muestra	$\bar{X}$	R
1	2.3972	0.0052
2	2.4191	0.0117
3	2.4215	0.0062
4	2.3917	0.0089
5	2.4151	0.0095
6	2.4027	0.0101
7	2.3921	0.0091
8	2.4171	0.0059
9	2.3951	0.0068
10	2.4215	0.0048
11	2.3887	0.0082
12	2.4107	0.0032
13	2.4009	0.0077
14	2.3992	0.0107
15	2.3889	0.0025
16	2.4107	0.0138
17	2.4109	0.0037
18	2.3944	0.0052
19	2.3951	0.0038
20	2.4015	0.0017

17.3 Suponga para el ejercicio de repaso 17.2 que el comprador fija especificaciones para la parte. Las especificaciones requieren que el diámetro caiga en el rango cubierto por 2.40000 ± 0.0100 mm. ¿Qué proporción de unidades producidas por este proceso no cumplirán con las especificaciones?

17.4 Para la situación del ejercicio de repaso 17.2, proporcione estimaciones numéricas de la media y de la desviación estándar del diámetro para la parte que se fabrica en el proceso.

17.5 Considere los datos de la tabla 17.1. Suponga que se toman muestras adicionales de tamaño 5 y se registra la resistencia de rotura. El muestreo produce los siguientes resultados (en libras por pulgada cuadrada).

Muestra	$\bar{X}$	R
1	1511	22
2	1508	14
3	1522	11
4	1488	18
5	1519	6
6	1524	11
7	1519	8
8	1504	7
9	1500	8
10	1519	14

a) Grafique los datos, utilice las gráficas $\bar{X}$ y R para los datos preliminares de la tabla 17.1.

b) ¿El proceso parece estar bajo control? Si no, explique por qué.

17.6 Considere un proceso bajo control con media $\mu = 25$ y $\sigma = 1.0$. Suponga que se usan subgrupos de tamaño 5 con límites de control, $\mu \pm 3\sigma/\sqrt{n}$, y línea central en μ. Suponga que ocurre un corrimiento en la media, y por ello la nueva media es $\mu = 26.5$.

a) ¿Cuál es el número promedio de muestras que se requiere (después del corrimiento) para detectar la situación fuera de control?

b) ¿Cuál es la desviación estándar del número de corridas que se requiere?

17.7 Considere la situación del ejemplo 17.2. Se toman los siguientes datos de muestras adicionales de tamaño 5. Grafique los valores $\bar{X}$ y S sobre la gráfica $\bar{X}$ y S que producen los datos en la muestra preliminar. ¿El proceso parece estar bajo control? ¿Por qué?

Muestra	$\bar{X}$	S_i
1	62.280	0.062
2	62.319	0.049
3	62.297	0.077
4	62.318	0.042
5	62.315	0.038
6	62.389	0.052
7	62.401	0.059
8	62.315	0.042
9	62.298	0.036
10	62.337	0.068

17.8 Se toman muestras de tamaño 50 cada hora de un proceso que produce cierto tipo de artículo que se considera defectuoso o no defectuoso. Se toman 20 muestras.

Muestra	Número de artículos defectuosos	Muestra	Número de artículos defectuosos
1	4	11	2
2	3	12	4
3	5	13	1
4	3	14	2
5	2	15	3
6	2	16	1
7	2	17	1
8	1	18	2
9	4	19	3
10	3	20	1

a) Construya una gráfica de control para controlar la proporción de defectuosos.

b) ¿El proceso parece estar bajo control? Explique.

17.9 Para la situación del ejercicio de repaso 17.8, suponga que se colectan datos adicionales como se muestra a continuación:

Muestra	Número de artículos defectuosos
1	3
2	4
3	2
4	2
5	3
6	1
7	3
8	5
9	7
10	7

¿El proceso parece estar bajo control? Explique.

17.10 Se intenta un esfuerzo de control de calidad para un proceso, donde se fabrican grandes placas de acero e interesan los defectos superficiales. El objetivo es establecer una gráfica de control de calidad para el número de defectos por placa. Los datos se indican en la siguiente página. Establezca la gráfica de control apropiada; utlice está información muestral. ¿El proceso parece estar bajo control?

Gráfica muestral	Número de artículos defectuosos	Gráfica muestral	Número de artículos defectuosos
1	4	11	1
2	2	12	2
3	1	13	2
4	3	14	3
5	0	15	1
6	4	16	4
7	5	17	3
8	3	18	2
9	2	19	1
10	2	20	3

Capítulo 18

Estadística bayesiana (opcional)

18.1 Conceptos bayesianos

Los métodos clásicos de estimación que hemos estudiado hasta ahora se basan únicamente en información que brinda la muestra aleatoria. Estos métodos interpretan esencialmente probabilidades como frecuencias relativas. Por ejemplo, para obtener un intervalo de confianza de 95% para μ, interpretamos el planteamiento

$$P(-1.96 < Z < 1.96) = 0.95$$

como que 95% de las veces en experimentos repetidos Z caerá entre -1.96 y 1.96. Como

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

para una muestra normal con varianza conocida, el enunciado de probabilidad aquí significa que 95% de los intervalos aleatorios $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ contienen la media μ verdadera. Otro enfoque para los métodos estadísticos de estimación se denomina **metodología bayesiana**. La princial idea del método proviene de la regla de Bayes que examinamos en la sección 2.8. La diferencia fundamental entre los enfoques bayesiano y clasico (es decir, el que hemos estudiado en este libro hasta aquí) es que en los conceptos bayesianos, los parámetros se consideran variables aleatorias.

Probabilidad subjetiva

La probabilidad subjetiva es el fundamento de los conceptos bayesianos. En el capítulo 2 analizamos dos enfoques posibles de la probabilidad; a saber, la indiferencia y la frecuencia relativa. El primero decide una probabilidad como consecuencia de experimentos repetidos. Por ejemplo, para decidir el porcentaje de tiros libres de un jugador de baloncesto, podemos registrar el número de tiros que hace y el número total de intentos que tal jugador tiene hasta el momento. La probabilidad de acertar un tiro libre para este jugador puede calcularse como el cociente de estos dos números. Por otro lado, si no sabemos acerca del posible sesgo de un dado, la probabilidad de que un 3 aparezca en el siguiente lanzamiento será de 1/6. Dicho enfoque en la interpretación de la probabilidad se basa en la regla de la indiferencia.

Sin embargo, en muchas situaciones, no es posible aplicar las interpretaciones de probabilidad anteriores. Por ejemplo, considere las siguientes preguntas: "¿Cuál es la probabilidad de que llueva mañana?" "¿Qué tan probable es que este inventario aumente a fin de mes?" Y ¿cuál es la probabilidad de que dos compañías se fusionen?" Estas preguntas difícilmente podrían interpretarse mediante los enfoques anteriores, y las respuestas podrían ser diferentes para distintas personas. No obstante, este tipo de preguntas se plantean constantemente en la vida diaria y el enfoque utilizado para explicar esas probabilidades se llama *probabilidad subjetiva*, ya que refleja opiniones subjetivas.

Perspectiva condicional

Recuerde que en los capítulos 9 a 17, todas las inferencias estadísticas estuvieron basadas en el hecho de que los parámetros se desconocían, pero eran cantidades fijas, a excepción de en la sección 9.14, donde los parámetros se trataron como variables y las estimaciones de probabilidad máxima se calcularon usando el condicionamiento en los datos. En la estadística bayesiana, los parámetros se consideran aleatorios y desconocidos para el investigador.

Puesto que los datos observados son los únicos resultados experimentales para el practicante, la inferencia estadística se basa en los datos reales observados a partir de un experimento dado. Tal visión se llama *perspectiva condicional*. Más aún, en los conceptos bayesianos, en tanto que el parámetro se considera como aleatorio, es factible especificar una distribución de probabilidad generalmente utilizando la *probabilidad subjetiva* para el parámetro. Tal distribución se denomina *distribución a priori* y comúnmente refleja la creencia previa del experimentador acerca del parámetro. En la perspectiva bayesiana, una vez que se realiza un experimento y se observan los datos, todo el conocimiento acerca de un parámetro está contenido en los datos reales observados, así como en la información previa.

Aplicaciones bayesianas

Aunque la regla de Bayes se atribuye a Thomas Bayes, en realidad fue el científico francés Pierre Simon Laplace quien introdujo primero las aplicaciones bayesianas. Laplace publicó un artículo sobre el uso de la inferencia bayesiana en los parámetros binomiales desconocidos. Sin embargo, a causa de su complicado enfoque de modelamiento y las objeciones de muchos en torno al uso de la distribución a priori *subjetiva*, los investigadores y científicos no aceptaron ampliamente las aplicaciones bayesianas, sino hasta principios de la década de 1990, cuando se lograron avances en los métodos computacionales bayesianos. Desde entonces, los métodos bayesianos se han aplicado con éxito en muchos campos, como la ingeniería, la agricultura, la ciencia biomédica y la ecología, entre otros.

18.2 Inferencias bayesianas

Considere el problema de encontrar una estimación puntual del parámetro θ para la población con distribución $f(x\,|\,\theta)$, dado θ. Denote con $\pi(\theta)$ la distribución previa de θ. Suponga que se observa la muestra aleatoria de tamaño n, denotada con $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Definición 18.1: La distribución de θ, dado el dato de $\boldsymbol{x}$, que se denomina distribución a posteriori, está dada por

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{g(\boldsymbol{x})},$$

donde $g(\boldsymbol{x})$ es la distribución marginal de $\boldsymbol{x}$.

La distribución marginal de $\boldsymbol{x}$ puede calcularse como

$$g(\boldsymbol{x}) = \begin{cases} \sum\limits_{\theta} f(\boldsymbol{x}|\theta)\pi(\theta), & \theta \text{ es discreta,} \\ \int_{-\infty}^{\infty} f(\boldsymbol{x}|\theta)\pi(\theta)\ d\theta, & \theta \text{ es continua.} \end{cases}$$

Ejemplo 18.1: Suponga que la distribución previa para la proporción de artículos defectuosos que produce una máquina es

p	0.1	0.2
$\pi(p)$	0.6	0.4

Denote con x el número de defectuosos entre una muestra aleatoria de tamaño 2. Encuentre la distribución de probabilidad a posteriori de p, dado que se conoce x.

Solución: La variable aleatoria X sigue una distribución binomial

$$f(x|p) = b(x; 2, p) = \binom{2}{x} p^x q^{2-x}, \quad x = 0, 1, 2.$$

La distribución marginal de x se puede calcular como

$$\begin{aligned} g(x) &= f(x|\ 0.1)\pi(0.1) + f(x|\ 0.2)\pi(0.2) \\ &= \binom{2}{x}[(0.1)^x(0.9)^{2-x}(0.6) + (0.2)^x(0.8)^{2-x}(0.4)]. \end{aligned}$$

Por lo tanto, la probabilidad a posteriori de $p = 0.1$, dado x, es

$$\pi(0.1|\ x) = \frac{f(x|\ 0.1)\pi(0.1)}{g(x)} = \frac{(0.1)^x(0.9)^{2-x}(0.6)}{(0.1)^x(0.9)^{2-x}(0.6) + (0.2)^x(0.8)^{2-x}(0.4)},$$

y $\pi(0.2|\ x) = 1 - \pi(0.1|\ x)$.

Suponga que se conoce $x = 0$.

$$\pi(0.1|0) = \frac{(0.1)^0(0.9)^{2-0}(0.6)}{(0.1)^0(0.9)^{2-0}(0.6) + (0.2)^0(0.8)^{2-0}(0.4)} = 0.6550,$$

y $\pi(0.2\,|\,0) = 0.3450$. Si se conoce $x = 1$, $\pi(0.1\,|\,1) = 0.4576$ y $\pi(0.2\,|\,1) = 0.5424$. Por último, $\pi(0.1\,|\,2) = 0.2727$ y $\pi(0.2\,|\,2) = 0.7273$.

La distribución a priori del ejemplo 18.1 es discreta, aunque el rango natural de p es de 0 a 1. Considere el siguiente ejemplo donde tenemos una distribución a priori que cubre el espacio completo de p.

Ejemplo 18.2: Suponga que la distribución a priori de p es uniforme (es decir, $\pi(p) = 1$, para $0 < p < 1$). Use la misma variable aleatoria X como en el ejemplo 18.1, para encontrar la distribución a posteriori de p.

Solución: Como en el ejemplo 18.1, tenemos

$$f(x|p) = b(x; 2, p) = \binom{2}{x} p^x q^{2-x}, \quad x = 0, 1, 2.$$

La distribución marginal de x puede calcularse como

$$g(x) = \int_0^1 f(x|p)\pi(p)\ dp = \binom{2}{x} \int_0^1 p^x (1-p)^{2-x}\ dp.$$

La integral anterior puede evaluarse en cada x directamente como $g(0) = 1/3$, $g(1) = 1/3$ y $g(2) = 1/3$. Por lo tanto, la distribución a posteriori de p, dada x, es

$$\pi(p|x) = \frac{\binom{2}{x} p^x (1-p)^{2-x}}{1/3} = 3\binom{2}{x} p^x (1-p)^{2-x}, \quad 0 < p < 1.$$

Usando la distribución a posteriori, podemos estimar directamente los parámetros en una población.

Estimación usando la distribución a posteriori

Una vez que se deriva la distribución a posteriori, fácilmente podemos usar el resumen de la distribución a posteriori para realizar inferencias sobre los parámetros de la población. Por ejemplo, la media, la mediana y la moda a posteriori son útiles para estimar el parámetro.

Ejemplo 18.3: Suponga que $x = 1$ se observa en el ejemplo 18.2. Determine la media y la moda a posteriori.

Solución: Cuando $x = 1$, la distribución a posteriori de p puede expresarse como

$$\pi(p|1) = 6p(1-p), \quad \text{para} \quad 0 < p < 1.$$

Si deseamos calcular la media de esta distribución, necesitamos encontrar

$$\int_0^1 6p^2(1-p)\ dp = 6\left(\frac{1}{3} - \frac{1}{4}\right) = \frac{1}{2}.$$

Para determinar la moda a posteriori, se requiere obtener el valor de p tal que se maximice la distribución a posteriori. Tomando la derivada de $\pi(p)$ con respecto a p, obtenemos $6 - 12p$. Al despejar p de $0 = 6 - 12p$, nos queda $p = 1/2$. La segunda derivada es -12, la cual implica que la moda a posteriori llegue a $p = 1/2$.

Los métodos bayesianos de estimación con respecto a la media μ de una población normal se basan en el siguiente ejemplo.

Ejemplo 18.4: Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n tomada de una población normal con varianza conocida σ^2, y la distribución a priori de la media poblacional

es una distribución normal con media conocida μ_0 y varianza conocida σ_0^2, entonces la distribución a posteriori de la media poblacional es también una distribución normal con media μ^* y desviación estándar σ^*, donde

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0 \qquad \text{y} \qquad \sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}}.$$

Solución: Al multiplicar la densidad de nuestra muestra

$$f(x_1, x_2, \ldots, x_n|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right],$$

por $-\infty < x_i < \infty$ e $i = 1, 2, \ldots, n$ por nuestra distribución a priori

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right], \quad -\infty < \mu < \infty,$$

obtenemos la densidad conjunta de la muestra aleatoria y la media de la población a partir de la cual se seleccionó la muestra. Es decir,

$$f(x_1, x_2, \ldots, x_n, \mu) = \frac{1}{(2\pi)^{(n+1)/2}\sigma^n\sigma_0} \times \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]\right\}.$$

En la sección 8.6 establecimos la identidad

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2,$$

que nos permite escribir

$$f(x_1, x_2, \ldots, x_n, \mu) = \frac{1}{(2\pi)^{(n+1)/2}\sigma^n\sigma_0} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2\right] \times \exp\left\{-\frac{1}{2}\left[\frac{n(\bar{x} - \mu)^2}{\sigma^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2}\right]\right\}.$$

Al completar los cuadrados del segundo exponente, escribimos la densidad conjunta de la muestra aleatoria y la media poblacional de la forma

$$f(x_1, x_2, \ldots, x_n, \mu) = K \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu^*}{\sigma^*}\right)^2\right],$$

donde

$$\mu^* = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}, \qquad \sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}},$$

y K es una función de los valores muestrales y los parámetros conocidos. Entonces, la distribución marginal de la muestra es

$$\begin{aligned} g(x_1, x_2, \ldots, x_n) &= K\sqrt{2\pi}\sigma^* \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma^*} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu^*}{\sigma^*}\right)^2\right] d\mu \\ &= K\sqrt{2\pi}\sigma^*, \end{aligned}$$

y la distribución a posteriori es

$$\begin{aligned} \pi(\mu|x_1, x_2, \ldots, x_n) &= \frac{f(x_1, x_2, \ldots, x_n, \mu)}{g(x_1, x_2, \ldots, x_n)} \\ &= \frac{1}{\sqrt{2\pi}\sigma^*} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu^*}{\sigma^*}\right)^2\right], \quad -\infty < \mu < \infty, \end{aligned}$$

que se identifica como una distribución normal con media μ^* y desviación estándar σ^*, donde μ^* y σ^* se definieron anteriormente.

El teorema del límite central nos permite utilizar el ejemplo 18.4 también cuando seleccionamos muestras aleatorias ($n \geq 30$ para muchos casos de experimentación en ingeniería), a partir de poblaciones no normales (la distribución no está muy alejada de la simétrica), y cuando la distribución a priori de la media es aproximadamente normal.

Resulta pertinente hacer algunos comentarios acerca del ejemplo 18.4. La media a posteriori μ^* también se puede escribir como

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0,$$

que es el promedio ponderado de la media muestral $\bar{x}$ y la media previa μ_0. Como ambos coeficientes están entre 0 y 1 y se suman a 1, la media a posteriori μ^* siempre está entre $\bar{x}$ y μ_0. Esto significa que la estimación a posteriori de la localización de μ se ve influida tanto por $\bar{x}$ como por μ_0. Más aún, la ponderación de $\bar{x}$ depende de la varianza previa, así como de la varianza de la media muestral. Para un problema con una muestra grande ($n \to \infty$), la media a posteriori $\mu^* \to \bar{x}$. Esto significa que la media a priori no desempeña ninguna función en la estimación de la media poblacional μ utilizando la distribución a posteriori. Esto es muy razonable puesto que indica que cuando una cantidad de datos es sustancial, la información a partir de los datos dominará la información de μ que brinda la a priori. Por otro lado, cuando la varianza previa es grande ($\sigma_0^2 \to \infty$), la media a posteriori μ^* también va a $\bar{x}$. Note que para una distribución normal, cuanto mayor es la varianza, más plana será la función de densidad. El carácter plano de la distribución normal, en este caso, significa que casi no hay información previa subjetiva disponible para el parámetro μ. Por lo tanto, es razonable que la estimación a posteriori μ^* sólo dependa del valor de $\bar{x}$.

Ahora considere la desviación estándar a posteriori σ^*. Este valor también se escribe como

$$\sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2/n}{\sigma_0^2 + \sigma^2/n}}.$$

Es evidente que el valor σ^* es menor que σ_0 y que $\sigma/\sqrt{n}$, la desviación estándar a proiri y la desviación estándar de $\bar{x}$, respectivamente. Esto sugiere que la estimación

a posteriori es más precisa que la a priori y que los datos muestrales. De ahí que la incorporación tanto de los datos como de la información previa dé como resultado una mejor información posterior que si se utiliza cualquiera de los datos o la información previa por sí solos. Esto es un fenómeno común en la inferencia bayesiana. Además, para calcular μ^* y σ^* mediante las fórmulas del ejemplo 18.4, suponemos que se conoce σ^2. Como, por lo general, éste no es el caso, deberemos reemplazar σ^2 por la varianza de la muestra s^2 siempre que $n \geq 30$.

Estimación del intervalo bayesiano

De manera similar al intervalo de confianza clásico, en el análisis bayesiano, podemos calcular un intervalo bayesiano $(1 - \alpha)100\%$ empleando la distribución a posteriori.

Definición 18.2: El intervalo $a < \theta < b$ se llamará un intervalo de Bayes $(1 - \alpha)100\%$ para θ si

$$\int_{-\infty}^{a} \pi(\theta|x)\ d\theta = \int_{b}^{\infty} \pi(\theta|x)\ d\theta = \frac{\alpha}{2}.$$

Recuerde que según el enfoque de la frecuencia, la probabilidad de un intervalo de confianza, digamos del 95%, se interpreta como una probabilidad de cobertura, lo cual significa que si un experimento se repite una y otra vez (con considerables datos no observados), la probabilidad de que los intervalos calculados, de acuerdo con la regla, cubran el parámetro verdadero de 95%. Sin embargo, en la interpretación del intervalo bayesiano, por ejemplo para un intervalo de 95%, simplemente podemos decir que la probabilidad de que el parámetro desconocido caiga dentro del intervalo calculado (que sólo depende de los datos observados) es del 95%.

Ejemplo 18.5: Suponga que $X \sim b(x; n, p)$ con $n = 2$; la distribución a priori de p es uniforme $\pi(p) = 1$, para $0 < p < 1$. Encuentre el intervalo de Bayes de 95% para p.

Solución: Como en el ejemplo 18.2, cuando $x = 0$, la distribución a posteriori es $\pi(p\,|\,0) = 3(1 - p)^2$, para $0 < p < 1$. Así que necesitamos despejar a y b utilizando la definición 18.2, lo que da como resultado lo siguiente:

$$0.025 = \int_{0}^{a} 3(1 - p)^2\ dp = 1 - (1 - a)^3,$$

y

$$0.025 = \int_{b}^{1} 3(1 - p)^2\ dp = (1 - b)^3.$$

Las soluciones a las ecuaciones de arriba arrojan como resultado $a = 0.0084$ y $b = 0.7076$. Por lo tanto, la probabilidad de que p caiga dentro de (0.0084, 0.7076) es de 95%.

Para la (población) normal y el caso normal (a priori) descrito en el ejemplo 18.4, la media a posteriori μ^* es el estimado de Bayes de la media poblacional μ, y puede construirse un **intervalo bayesiano** $(1 - \alpha)100\%$ calculando el intervalo

$$\mu^* - z_{\alpha/2}\sigma^* < \mu < \mu^* + z_{\alpha/2}\sigma^*,$$

que se centra en la media a posteriori y contiene $(1 - \alpha)100\%$ de la probabilidad posterior.

Ejemplo 18.6: Una empresa de equipo eléctrico fabrica bombillas de luz que tienen una duración que está distribuida de forma aproximadamente normal con una desviación estándar de 100 horas. Experiencia anterior nos indica que μ es un valor de una variable aleatoria normal con una media $\mu_0 = 800$ horas y una desviación estándar $\sigma_0 = 10$ horas. Si una muestra aleatoria de 25 bombillas tiene una duración promedio de 780 horas, encuentre un intervalo bayesiano de 95% para μ.

Solución: De acuerdo con el ejemplo 18.4, la distribución a posteriori de la media también es una distribución normal con media

$$\mu^* = \frac{(25)(780)(10)^2 + (800)(100)^2}{(25)(10)^2 + (100)^2} = 796,$$

y desviación estándar

$$\sigma^* = \sqrt{\frac{(10)^2(100)^2}{(25)(10)^2 + (100)^2}} = \sqrt{80}.$$

El intervalo bayesiano de 95% para μ está dado entonces por

$$796 - 1.96\sqrt{80} < \mu < 796 + 1.96\sqrt{80},$$

o

$$778.5 < \mu < 813.5.$$

De esta manera, estamos un 95% seguros de que μ estará entre 778.5 y 813.5.

Por otro lado, si se ignora la información previa acerca de μ, procedemos como en la sección 9.4 para construir el intervalo de confianza clásico de 95%.

$$780 - (1.96)\left(\frac{100}{\sqrt{25}}\right) < \mu < 780 + (1.96)\left(\frac{100}{\sqrt{25}}\right),$$

o $740.8 < \mu < 819.2$, que se observa que es más amplio que el intervalo bayesiano correspondiente.

18.3 Estimación bayesiana utilizando el contexto de la teoría de decisión

Con la metodología bayesiana, se puede obtener la distribución a posteriori del parámetro. La estimación bayesiana también se deriva usando la distribución a posteriori cuando se incurre en una función de pérdida. Por ejemplo, el estimador bayesiano más común utilizado está bajo la **función de pérdida del error cuadrático**, que es similar a la estimación por mínimos cuadrados que presentamos en el capítulo 11 al estudiar el análisis de regresión.

Definición 18.3: La media de la distribución a posteriori $\pi(\theta|x)$, denotada con θ^*, se llama **estimación de Bayes** de θ, bajo la función de pérdida del error cuadrático.

Ejemplo 18.7: Encuentre la estimación de Bayes de p para todos los valores de x, en el ejemplo 18.1.

Solución: Cuando $x = 0$, $p^* = (0.1)(0.6550) + (0.2)(03450) = 0.1345$.
Cuando $x = 1$, $p^* = (0.1)(0.4576) + (0.2)(0.5424) = 0.1542$.

Cuando $x = 2$, $p^* = (0.1)(0.2727) + (0.2)(0.7273) = 0.1727$.

Note que la estimación clásica de p es $\hat{p} = x/n = 0$, $1/2$ y 1, respectivamente, para los valores de x en 0, 1 y 2. Tales estimaciones clásicas son muy diferentes de las estimaciones de Bayes correspondientes.

Ejemplo 18.8: Repita el ejemplo 18.7 en la situación del ejemplo 18.2.

Solución: Puesto que la distribución a posteriori de p se puede expresar como

$$\pi(p|x) = \frac{\binom{2}{x}p^x(1-p)^{2-x}}{1/3} = 3\binom{2}{x}p^x(1-p)^{2-x}, \qquad 0 < p < 1,$$

la estimación de Bayes de p es

$$p^* = E(p|x) = 3\binom{2}{x}\int_0^1 p^{x+1}(1-p)^{2-x}\, dp,$$

la cual da $p^* = 1/4$ para $x = 0$, $p^* = 1/2$ para $x = 1$ y $p^* = 3/4$ para $x = 2$, respectivamente. Entonces cuando se observa $x = 1$, la estimación de Bayes y la estimación clásica de $\hat{p}$ son equivalentes.

Para la situación normal que se describe en el ejemplo 18.4, la estimación de Bayes de μ bajo la pérdida del error cuadrático será la media a posteriori μ^*.

Ejemplo 18.9: Suponga que la distribución muestral de una variable aleatoria X es de Poisson con parámetro λ. Suponga que la distribución a priori de λ sigue una distribución gamma con parámetros (α, β). Encuentre la estimación de Bayes de λ bajo la función de pérdida del error cuadrático.

Solución: La función de densidad de X es

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad \text{para} \quad x = 0, 1, \ldots,$$

y la distribución a priori de λ es

$$\pi(\lambda) = \frac{1}{\beta^\alpha\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda/\beta}, \qquad \text{para} \quad \lambda > 0.$$

Por lo tanto, la distribución a posteriori de λ se puede expresar como

$$\pi(\lambda|x) = \frac{\frac{\lambda^x e^{-\lambda}\lambda^{\alpha-1}e^{-\lambda/\beta}}{x!\beta^\alpha\Gamma(\alpha)}}{\int_0^\infty \frac{\lambda^x e^{-\lambda}\lambda^{\alpha-1}e^{-\lambda/\beta}}{x!\beta^\alpha\Gamma(\alpha)}d\lambda} = \frac{\lambda^{x+\alpha-1}e^{-(1+1/\beta)\lambda}}{\int_0^\infty \lambda^{x+\alpha-1}e^{-(1+1/\beta)\lambda}\, d\lambda}$$
$$= \frac{1}{(1+1/\beta)^{-(x+\alpha)}\Gamma(x+\alpha)}\lambda^{x+\alpha-1}e^{-(1+1/\beta)\lambda},$$

la cual sigue otra distribución gamma con los parámetros $(x + \alpha, (1 + 1/\beta)^{-1}$. Usando el teorema 6.3, obtenemos la media a posteriori

$$\hat{\lambda} = \frac{x+\alpha}{1+1/\beta}.$$

Como la media a posteriori es la estimación de Bayes bajo la pérdida del error cuadrático, $\hat{\lambda}$ será nuestra estimación de Bayes.

Ejercicios

18.1 Estime la proporción de defectuosos que produce la máquina del ejemplo 18.1, si la muestra aleatoria de tamaño 2 produce 2 defectuosos.

18.2 Supongamos que la distribución a priori para la proporción p de bebidas de una máquina despachadora que se derraman al servirse es

p	0.05	0.10	0.15
$\pi(p)$	0.3	0.5	0.2

Si 2 de las siguientes 9 bebidas de esta máquina se derraman, encuentre

a) la distribución a posterior para la proporción p;

b) la estimación de Bayes de p.

18.3 Repita el ejercicio 18.2 cuando 1 de las siguientes 4 bebidas se derrama y la distribución uniforme a priori es

$$\pi(p) = 10, \quad 0.05 < p < 0.15.$$

18.4 El constructor de un nuevo complejo de condominios afirma que 3 de 5 compradores preferirá un departamento de dos recámaras; mientras que su banquero afirma que sería más correcto decir que 7 de 10 compradores preferirán uno de dos recámaras. En las predicciones a priori de este tipo, el banquero ha sido dos veces más confiable que el constructor. Si 12 de los siguientes 15 condominios que se venden en este complejo son de dos recámaras, encuentre

a) las probabilidades a posteriori que se asocian con las afirmaciones del constructor y del banquero;

b) una estimación puntual de la proporción de compradores que prefieren un departamento de dos recámaras.

18.5 El tiempo en que se consume la primera etapa de un cohete es una variable aleatoria normal, con una desviación estándar de 0.8 minutos. Suponga una distribución a priori normal para μ con una media de ocho minutos y una desviación estándar de 0.2 minutos. Si se lanzan 10 de estos cohetes y la primera etapa tiene un tiempo de consumo promedio de 9 minutos, encuentre un intervalo bayesiano de 95% para μ.

18.6 La utilidad diaria de una máquina despachadora de jugos que se coloca en un edificio de oficinas es un valor de una variable aleatoria normal, con media μ y varianza σ^2 desconocidas. Desde luego, la media variará algo de un edificio a otro, y el distribuidor considera que estas utilidades promedio diarias se pueden describir mejor usando una distribución normal con media $\mu_0 = \$30.00$ y desviación estándar $\sigma_0 = \$1.75$. Si una de estas máquinas despachadoras de jugo, que se coloca en cierto edificio, muestra una utilidad promedio diaria de $\bar{x} = \$24.90$, durante los primeros 30 días con una desviación estándar de $s = \$2.10$, encuentre

a) una estimación de Bayes de la utilidad promedio diaria real para este edificio;

b) un intervalo bayesiano de 95% de μ para este edificio;

c) la probabilidad de que la utilidad promedio diaria de la máquina en este edificio esté entre \$24.00 y \$26.00.

18.7 El departamento de matemáticas de una universidad grande diseña un examen de colocación para aplicarlo a los grupos de nuevo ingreso a primer año. Los miembros del departamento consideran que la calificación promedio para este examen variará de un grupo de primer año a otro. Esta variación de la calificación promedio del grupo se expresa de manera subjetiva mediante una distribución normal, con una media $\mu_0 = 72$ y varianza $\sigma_0^2 = 5.76$.

a) ¿Qué probabilidad a priori de que la calificación promedio real, que asigna el departamento para los alumnos de nuevo ingreso del siguiente año, caiga entre 71.8 y 73.4?

b) Si el examen se aplica a una muestra aleatoria de 100 estudiantes de primer grado del siguiente grupo de nuevo ingreso que tiene como resultado una calificación promedio de 70 con una varianza de 64, construya un intervalo bayesiano de 95% para μ.

c) ¿Qué probabilidad a posteriori debería asignar el departamento al evento del inciso *a*)?

18.8 Suponga que en el ejemplo 18.6 1a empresa de equipo eléctrico no tiene suficiente información a priori con respecto a la duración media poblacional, para ser capaz de suponer una distribución normal para μ. La empresa cree, sin embargo, que μ seguramente estará entre 770 y 830 horas y considera que una aproximación bayesiana más realista sería suponer la distribución a priori

$$\pi(\mu) = \frac{1}{60}, \quad 770 < \mu < 830.$$

Si una muestra aleatoria de 25 bombillas da una vida promedio de 780 horas, siga los pasos de la demostra-

ción del ejemplo 18.4 para encontrar la distribución a posteriori

$$\pi(\mu|x_1, x_2, \ldots, x_{25}).$$

18.9 Suponga que el tiempo de falla T para cierta bisagra es una variable aleatoria exponencial con densidad de probabilidad

$$f(t) = \theta e^{-\theta t}, \quad t > 0.$$

De cierta experiencia anterior nos inclinamos a pensar que θ es un valor de una variable aleatoria exponencial con densidad de probabilidad

$$\pi(\theta) = 2e^{-2\theta}, \quad \theta > 0.$$

Si tenemos una muestra de n observaciones de T, muestre que la distribución a posteriori de Θ es una distribución gamma con parámetros

$$\alpha = n + 1, \quad \text{y} \quad \beta = \left(\sum_{i=1}^{n} t_i + 2 \right)^{-1}.$$

18.10 Suponga que una muestra consta de 5, 6, 6, 7, 5, 6, 4, 9, 3, 6 y proviene de una población de Poisson con media λ. Suponga que el parámetro λ sigue una distribución gamma con parámetros (3, 2). Bajo la pérdida del error cuadrático, encuentre la estimación de Bayes para λ.

18.11 Una variable aleatoria X sigue una distribución binomial negativa con parámetros $k = 5$ y p (esto es, $b^*(x; 5, p)$). Además, se sabe que p sigue una distribución uniforme en el intervalo (0, 1). Encuentre el estimado de Bayes de p bajo la pérdida del error cuadrático. [Sugerencia: Le resultará útil la función de densidad en el ejercicio 6.50. Además, la media de la distribución beta con parámetros (α, β) es $\alpha/(\alpha + \beta)$.]

Bibliografía

[1] Bartlett, M. S. y Kendall, D. G. (1946). "The Statistical Analysis of Variante Heterogeneity and Logarithmic Transformation", *Journal of the Royal Statistical Society,* Ser. B. **8**, 128-138.

[2] Bowker, A. H. y Lieberman, G. J. (1972). *Engineering Statistics,* 2a. ed. Upper Saddle River, N.J.: Prentice Hall.

[3] Box, G. E. P. (1988). "Signal to Noise Ratios, Performance Criteria and Transformations (with discussion);" *Technometrics,* **30***,* 1-17.

[4] Box, G. E. P. Fung, C. A. (1986). "Studies in Quality Improvement: Minimizing Transmitted Variation by Parameter Design", informe 8. University of Wisconsin-Madison, Center for Quality and Productivity Improvement.

[5] Box, G. E. P., Hunter, W. C. y Hunter, T. S. (1978). *Statistics for Experimenters.* Nueva York: John Wiley & Sons.

[6] Brownlee, K. A. (1984). *Statistical Theory and Methodology: in Science and Engineering,* 2a. ed. Nueva York: John Wiles & Sons.

[7] Carroll, R. J. y Ruppert, D. (1988). *Transformation and Weighting in Regression.* Nueva York: Chapman and Hall.

[8] Chatterjee, S. Hadi. A. S. y Price, B. (1999). *Regression Analysis by Example,* 3a. ed. Nueva York: John Wiles & Sons.

[9] Cook, R. D. y Weisberg, S. (1982). *Residuals and Influence in Regression.* Nueva York: Chapman and Hall.

[10] Daniel, C. y Wood, F. S. (1999). *Fitting Equations to Data: Competer Analysis of Multifactor Data,* 2a. ed. Nueva York: John Wiley & Sons.

[11] Daniel, W. W. (1989). *Applied Nonparametric Statistics,* 2a. ed. Belmont, California: Wadsworth Publishing Company.

[12] Devore, J. L. (2003). *Probability and Statistics for Engineering and the Sciences,* 6a. ed. Belmont, Calif: Duxbury Press.

[13] Dixon, W. J. (1983). *Introduction to Statistical Analysis,* 4a. ed. Nueva York: McGraw-Hill.

[14] Draper, N. R. y Smith, H. (1998). *Applied Regression Analysis,* 3a. ed. Nueva York: John Wiley & Sons.

[15] Duncan, A. (1986). *Quality Control and Industrial Statistics,* 5a. ed. Homewood, Illinois: Irwin.

[16] Dyer, D. D. y Keating, J. P. (1980). "On the Determination of Critical Values for Bartlett's Test", *J. Am. Stat. Assoc.*, **75**, 313-319.

[17] Ewan, W. D. y Kemp, K. W. (1960). "Sampling Inspection of Continuous Processes with No Autocorrelation between Successive Results", *Biometrika*, vol. **47**, 363-380.

[18] Gunst, R. F. y Mason, R. L. (1980). *Regression Analysis and Its Application: A Data-Oriented Approach.* Nueva York: Marcel Dekker.

[19] Guttman, I. Wilks, S. S. y Hunter, J. S. (1971). *Introductory Engineering Statistics.* Nueva York: Wiley & Sons.

[20] Hicks, C. R. y Turner, K. V. (1999). *Fundamental Concepts in the Design of Experiments.* 5a. ed. Oxford: Oxford University Press.

[21] Hoaglin, D. C., Mosteller, F. y Tukey, J. W. (1991). *Fundamentals of Exploratory Analysis of Variance.* Nueva York: Wiley & Sons.

[22] Hocking, R.R., (1976). "The Analysis and Selection of Variables in Linear Regression", *Biometrics,* **32**, 1-49.

[23] Hoerl, A. E. y Wennard. R. W. (1970). "Ridge Regression: Applications to Nonorthogonal Problems", *Technometrics,* **12**, 55-67.

[24] Hogg, R. V., Craig, A. y McKean, J. W. (2004). *Introduction to Mathematical Statistics;* 6a. ed. Upper Saddle River, N.J.: Prentice Hall.

[25] Hogg, R. V. y Ledolter, J. (1992). *Applied Statistics for Engineers and Physical Scientists,* 2a. ed. Upper Saddle River: N.J.: Prentice Hall.

[26] Hollander, M. y Wolfe, D. (1999). *Nonparametric Statistical Methods.* Nueva York: John Wiley & Sons.

[27] Johnson, N. L. y Leone, F. C. (1977). *Statistics and Experimental Design: In Engineering and the Physical Sciences,* vols. I y II, 2a. ed. Nueva York: John Wiley & Sons.

[28] Kackar, R. (1985). "Off-Line Quality Control, Parameter Design, and the Taguchi Methods", *Journal of Quality Technology*, **17**, 176-188.

[29] Koopmans, L. H. (1987). *An Introduction to Contemporary Statistics,* 2a. ed. Boston: Duxbury Press.

[30] Larsen, R. J. y Morris, M. L. (2000). *An Introduction to Mathematical Statistics and Its Applications,* 3a. ed. Upper Saddle River. N.J.: Prentice Hall.

[31] Lehmann, E. L. y D'Abrera, H. J. M. (1998). *Nonparametrics: Statistical Methods Based on Ranks,* ed. rev. Upper Saddle River: N.J.: Prentice Hall.

[32] Lentner, M. y Bishop, T. (1986). *Design and Analysis of Experiments,* 2a. ed. Blacksburg. VA: Valley Book Co.

[33] Mallows, C. L. (1973). "Some comments of C_p", *Technometrics,* **15**, 661-675.

[34] McClave. J. T., Dietrich, F. H. y Sincich, T. (1997). *Statistics,* 7a. ed. Upper Saddle River, N.J.: Prentice Hall.

[35] Montgomery. D. C. (2000). *Introduction to Statistical Quality Control,* 4a. ed. Nueva York: John Wiley & Sons.

[36] Montgomery. D. C. (2001). *Design and Analysis of Experiments,* 5a. ed. Nueva York: John Wiley & Sons.

[37] Mosteller, F. y Tukey, J. (1977). *Data Analysis and Regression.* Reading; MA: Addison-Wesley Publishing Co.

[38] Myers. R. H. (1990). *Classical and Modern Regression with Applications,* 2a. ed. Boston: Duxbury Press.

[39] Myers, R. H., Khuri, A. I. y Vining, G. G. (1992). "Response Surface Alternatives to the Taguchi Robust Parameter Design Approach", *The American Statistician,* **46**, 131-139.

[40] Myers, R. H. y Montgomery, D. C. (2002). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments,* 2a. ed. Nueva York: John Wiley & Sons.

[41] Neter, J., Wassermann, W., y Kutner, M. H. (1989). *Applied Linear Regression Models,* 2a. ed. Burr Ridge, Illinois: Irwin.

[42] Noether, G. E. (1976). *Introduction to Statistics: A Nonparametric Approach,* 2a. ed. Boston: Houghton Mifflin Company.

[43] Olkin, I., Gleser, L. J. y Derman, C. (1994). *Probability Models and Applications,* 2a. ed. Nueva York: Prentice Hall.

[44] Ott. R. L. y Longnecker, M. T. (2000). *An Introduction to Statistical Methods and Data Analysis.* 5a. ed. Boston: Duxbury Press.

[45] Plackett, R. L. y Burman, J. P. (1946). "The Design of Multifactor Experiments", *Biometrika,* 33, 305-325.

[46] Ross, S. M. (2002). *Introduction to Applied Probability Models,* 8a. ed. Nueva York: Academic Press, Inc.

[47] Satterthwaite, F. E. (1946). "An approximate distribution of estimates of variance components", *Biometrics,* **2***,* 110-114.

[48] Schilling. E. G. y Nelson, P. R. (1976). "The Effect of Nonnormality on the Control Limits of $\bar{X}$ Charts", *J. Quality Tech.,* **8**, 347-373.

[49] Schmidt. S. R. y Launsby, R. G. (1991). *Understanding Industrial Designed Experiments.* Colorado Springs, CO: Air Academy Press.

[50] Shoemaker, A. C., Tsui, K.-L., y Wu, C. F. J. (1991). "Economical Experimentation Methods for Robust Parameter Design", *Technometrics,* **33**, 415-428.

[51] Snedecor, G. W. y Cochran, W. G. (1989). *Statistical Methods,* 8a. ed. Allies, Iowa: The Iowa State University Press.

[52] Steel, R. G. D., Torrie, J. H. y Dickey, D. A. (1996). *Principles and Procedures of Statistics: A Biometrical Approach,* 3a. ed. Nueva York: McGraw-Hill.

[53] Taguchi, G. (1991). *Introduction to Quality Engineering.* White Plains, N.Y.: Unipub/Kraus International.

[54] Taguchi, G. y Wu, Y. (1985). *Introduction to Off-Line Quality Control.* Nagoya, Japan: Central Japan Quality Control Association.

[55] Thompson, W. O. y Cady, F. B. (1973). *Proceedings of the University of Kentucky Conference en Regression with a Larqe Number of Predictor Variables.* Lexington, Kentucky: University of Kentucky Press.

[56] Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley Publishing Co.

[57] Vining, G. G. y Myers, R. H. (1990). "Combining Taguchi and Response Surface Philosophies: A Dual Response Approach", *Journal of Quality Technology,* **22**, 38- 45.

[58] Welch, W. J., Yu, T. K., Kang, S. M. y Sacks, J. (1990). "Computer Experiments for Quality Control by Parameter Design", *Journal of Quality Technology*, **22**, 15-22.

[59] Winer, B. J. (1991). *Statistical Principles In Experimental Design,* 3a. ed. Nueva York: McGraw-Hill.

Apéndice A

Tablas y pruebas estadísticas

Tabla A.1 Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	r	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
1	**0**	0.9000	0.8000	0.7500	0.7000	0.6000	0.5000	0.4000	0.3000	0.2000	0.1000
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	**0**	0.8100	0.6400	0.5625	0.4900	0.3600	0.2500	0.1600	0.0900	0.0400	0.0100
	1	0.9900	0.9600	0.9375	0.9100	0.8400	0.7500	0.6400	0.5100	0.3600	0.1900
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	**0**	0.7290	0.5120	0.4219	0.3430	0.2160	0.1250	0.0640	0.0270	0.0080	0.0010
	1	0.9720	0.8960	0.8438	0.7840	0.6480	0.5000	0.3520	0.2160	0.1040	0.0280
	2	0.9990	0.9920	0.9844	0.9730	0.9360	0.8750	0.7840	0.6570	0.4880	0.2710
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	**0**	0.6561	0.4096	0.3164	0.2401	0.1296	0.0625	0.0256	0.0081	0.0016	0.0001
	1	0.9477	0.8192	0.7383	0.6517	0.4752	0.3125	0.1792	0.0837	0.0272	0.0037
	2	0.9963	0.9728	0.9492	0.9163	0.8208	0.6875	0.5248	0.3483	0.1808	0.0523
	3	0.9999	0.9984	0.9961	0.9919	0.9744	0.9375	0.8704	0.7599	0.5904	0.3439
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	**0**	0.5905	0.3277	0.2373	0.1681	0.0778	0.0313	0.0102	0.0024	0.0003	0.0000
	1	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0067	0.0005
	2	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.0579	0.0086
	3	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.2627	0.0815
	4	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.6723	0.4095
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	**0**	0.5314	0.2621	0.1780	0.1176	0.0467	0.0156	0.0041	0.0007	0.0001	0.0000
	1	0.8857	0.6554	0.5339	0.4202	0.2333	0.1094	0.0410	0.0109	0.0016	0.0001
	2	0.9842	0.9011	0.8306	0.7443	0.5443	0.3438	0.1792	0.0705	0.0170	0.0013
	3	0.9987	0.9830	0.9624	0.9295	0.8208	0.6563	0.4557	0.2557	0.0989	0.0159
	4	0.9999	0.9984	0.9954	0.9891	0.9590	0.8906	0.7667	0.5798	0.3446	0.1143
	5	1.0000	0.9999	0.9998	0.9993	0.9959	0.9844	0.9533	0.8824	0.7379	0.4686
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	**0**	0.4783	0.2097	0.1335	0.0824	0.0280	0.0078	0.0016	0.0002	0.0000	
	1	0.8503	0.5767	0.4449	0.3294	0.1586	0.0625	0.0188	0.0038	0.0004	0.0000
	2	0.9743	0.8520	0.7564	0.6471	0.4199	0.2266	0.0963	0.0288	0.0047	0.0002
	3	0.9973	0.9667	0.9294	0.8740	0.7102	0.5000	0.2898	0.1260	0.0333	0.0027
	4	0.9998	0.9953	0.9871	0.9712	0.9037	0.7734	0.5801	0.3529	0.1480	0.0257
	5	1.0000	0.9996	0.9987	0.9962	0.9812	0.9375	0.8414	0.6706	0.4233	0.1497
	6		1.0000	0.9999	0.9998	0.9984	0.9922	0.9720	0.9176	0.7903	0.5217
	7			1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1 (continuación) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
8	**0**	0.4305	0.1678	0.1001	0.0576	0.0168	0.0039	0.0007	0.0001	0.0000	
	1	0.8131	0.5033	0.3671	0.2553	0.1064	0.0352	0.0085	0.0013	0.0001	
	2	0.9619	0.7969	0.6785	0.5518	0.3154	0.1445	0.0498	0.0113	0.0012	0.0000
	3	0.9950	0.9437	0.8862	0.8059	0.5941	0.3633	0.1737	0.0580	0.0104	0.0004
	4	0.9996	0.9896	0.9727	0.9420	0.8263	0.6367	0.4059	0.1941	0.0563	0.0050
	5	1.0000	0.9988	0.9958	0.9887	0.9502	0.8555	0.6846	0.4482	0.2031	0.0381
	6		0.9999	0.9996	0.9987	0.9915	0.9648	0.8936	0.7447	0.4967	0.1869
	7		1.0000	1.0000	0.9999	0.9993	0.9961	0.9832	0.9424	0.8322	0.5695
	8				1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	**0**	0.3874	0.1342	0.0751	0.0404	0.0101	0.0020	0.0003	0.0000		
	1	0.7748	0.4362	0.3003	0.1960	0.0705	0.0195	0.0038	0.0004	0.0000	
	2	0.9470	0.7382	0.6007	0.4628	0.2318	0.0898	0.0250	0.0043	0.0003	0.0000
	3	0.9917	0.9144	0.8343	0.7297	0.4826	0.2539	0.0994	0.0253	0.0031	0.0001
	4	0.9991	0.9804	0.9511	0.9012	0.7334	0.5000	0.2666	0.0988	0.0196	0.0009
	5	0.9999	0.9969	0.9900	0.9747	0.9006	0.7461	0.5174	0.2703	0.0856	0.0083
	6	1.0000	0.9997	0.9987	0.9957	0.9750	0.9102	0.7682	0.5372	0.2618	0.0530
	7		1.0000	0.9999	0.9996	0.9962	0.9805	0.9295	0.8040	0.5638	0.2252
	8			1.0000	1.0000	0.9997	0.9980	0.9899	0.9596	0.8658	0.6126
	9					1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	**0**	0.3487	0.1074	0.0563	0.0282	0.0060	0.0010	0.0001	0.0000		
	1	0.7361	0.3758	0.2440	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	
	2	0.9298	0.6778	0.5256	0.3828	0.1673	0.0547	0.0123	0.0016	0.0001	
	3	0.9872	0.8791	0.7759	0.6496	0.3823	0.1719	0.0548	0.0106	0.0009	0.0000
	4	0.9984	0.9672	0.9219	0.8497	0.6331	0.3770	0.1662	0.0473	0.0064	0.0001
	5	0.9999	0.9936	0.9803	0.9527	0.8338	0.6230	0.3669	0.1503	0.0328	0.0016
	6	1.0000	0.9991	0.9965	0.9894	0.9452	0.8281	0.6177	0.3504	0.1209	0.0128
	7		0.9999	0.9996	0.9984	0.9877	0.9453	0.8327	0.6172	0.3222	0.0702
	8		1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.6242	0.2639
	9				1.0000	0.9999	0.9990	0.9940	0.9718	0.8926	0.6513
	10					1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	**0**	0.3138	0.0859	0.0422	0.0198	0.0036	0.0005	0.0000			
	1	0.6974	0.3221	0.1971	0.1130	0.0302	0.0059	0.0007	0.0000		
	2	0.9104	0.6174	0.4552	0.3127	0.1189	0.0327	0.0059	0.0006	0.0000	
	3	0.9815	0.8389	0.7133	0.5696	0.2963	0.1133	0.0293	0.0043	0.0002	
	4	0.9972	0.9496	0.8854	0.7897	0.5328	0.2744	0.0994	0.0216	0.0020	0.0000
	5	0.9997	0.9883	0.9657	0.9218	0.7535	0.5000	0.2465	0.0782	0.0117	0.0003
	6	1.0000	0.9980	0.9924	0.9784	0.9006	0.7256	0.4672	0.2103	0.0504	0.0028
	7		0.9998	0.9988	0.9957	0.9707	0.8867	0.7037	0.4304	0.1611	0.0185
	8		1.0000	0.9999	0.9994	0.9941	0.9673	0.8811	0.6873	0.3826	0.0896
	9			1.0000	1.0000	0.9993	0.9941	0.9698	0.8870	0.6779	0.3026
	10					1.0000	0.9995	0.9964	0.9802	0.9141	0.6862
	11						1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1 (continuación) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
12	**0**	0.2824	0.0687	0.0317	0.0138	0.0022	0.0002	0.0000			
	1	0.6590	0.2749	0.1584	0.0850	0.0196	0.0032	0.0003	0.0000		
	2	0.8891	0.5583	0.3907	0.2528	0.0834	0.0193	0.0028	0.0002	0.0000	
	3	0.9744	0.7946	0.6488	0.4925	0.2253	0.0730	0.0153	0.0017	0.0001	
	4	0.9957	0.9274	0.8424	0.7237	0.4382	0.1938	0.0573	0.0095	0.0006	0.0000
	5	0.9995	0.9806	0.9456	0.8822	0.6652	0.3872	0.1582	0.0386	0.0039	0.0001
	6	0.9999	0.9961	0.9857	0.9614	0.8418	0.6128	0.3348	0.1178	0.0194	0.0005
	7	1.0000	0.9994	0.9972	0.9905	0.9427	0.8062	0.5618	0.2763	0.0726	0.0043
	8		0.9999	0.9996	0.9983	0.9847	0.9270	0.7747	0.5075	0.2054	0.0256
	9		1.0000	1.0000	0.9998	0.9972	0.9807	0.9166	0.7472	0.4417	0.1109
	10				1.0000	0.9997	0.9968	0.9804	0.9150	0.7251	0.3410
	11					1.0000	0.9998	0.9978	0.9862	0.9313	0.7176
	12						1.0000	1.0000	1.0000	1.0000	1.0000
13	**0**	0.2542	0.0550	0.0238	0.0097	0.0013	0.0001	0.0000			
	1	0.6213	0.2336	0.1267	0.0637	0.0126	0.0017	0.0001	0.0000		
	2	0.8661	0.5017	0.3326	0.2025	0.0579	0.0112	0.0013	0.0001		
	3	0.9658	0.7473	0.5843	0.4206	0.1686	0.0461	0.0078	0.0007	0.0000	
	4	0.9935	0.9009	0.7940	0.6543	0.3530	0.1334	0.0321	0.0040	0.0002	
	5	0.9991	0.9700	0.9198	0.8346	0.5744	0.2905	0.0977	0.0182	0.0012	0.0000
	6	0.9999	0.9930	0.9757	0.9376	0.7712	0.5000	0.2288	0.0624	0.0070	0.0001
	7	1.0000	0.9988	0.9944	0.9818	0.9023	0.7095	0.4256	0.1654	0.0300	0.0009
	8		0.9998	0.9990	0.9960	0.9679	0.8666	0.6470	0.3457	0.0991	0.0065
	9		1.0000	0.9999	0.9993	0.9922	0.9539	0.8314	0.5794	0.2527	0.0342
	10			1.0000	0.9999	0.9987	0.9888	0.9421	0.7975	0.4983	0.1339
	11				1.0000	0.9999	0.9983	0.9874	0.9363	0.7664	0.3787
	12					1.0000	0.9999	0.9987	0.9903	0.9450	0.7458
	13						1.0000	1.0000	1.0000	1.0000	1.0000
14	**0**	0.2288	0.0440	0.0178	0.0068	0.0008	0.0001	0.0000			
	1	0.5846	0.1979	0.1010	0.0475	0.0081	0.0009	0.0001			
	2	0.8416	0.4481	0.2811	0.1608	0.0398	0.0065	0.0006	0.0000		
	3	0.9559	0.6982	0.5213	0.3552	0.1243	0.0287	0.0039	0.0002		
	4	0.9908	0.8702	0.7415	0.5842	0.2793	0.0898	0.0175	0.0017	0.0000	
	5	0.9985	0.9561	0.8883	0.7805	0.4859	0.2120	0.0583	0.0083	0.0004	
	6	0.9998	0.9884	0.9617	0.9067	0.6925	0.3953	0.1501	0.0315	0.0024	0.0000
	7	1.0000	0.9976	0.9897	0.9685	0.8499	0.6047	0.3075	0.0933	0.0116	0.0002
	8		0.9996	0.9978	0.9917	0.9417	0.7880	0.5141	0.2195	0.0439	0.0015
	9		1.0000	0.9997	0.9983	0.9825	0.9102	0.7207	0.4158	0.1298	0.0092
	10			1.0000	0.9998	0.9961	0.9713	0.8757	0.6448	0.3018	0.0441
	11				1.0000	0.9994	0.9935	0.9602	0.8392	0.5519	0.1584
	12					0.9999	0.9991	0.9919	0.9525	0.8021	0.4154
	13					1.0000	0.9999	0.9992	0.9932	0.9560	0.7712
	14						1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1 (continuación) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
15	**0**	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000				
	1	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000			
	2	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000		
	3	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001		
	4	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0000	
	5	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0001	
	6	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0008	
	7	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0042	0.0000
	8		0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0181	0.0003
	9		0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.0611	0.0022
	10		1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.1642	0.0127
	11			1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.3518	0.0556
	12				1.0000	0.9997	0.9963	0.9729	0.8732	0.6020	0.1841
	13					1.0000	0.9995	0.9948	0.9647	0.8329	0.4510
	14						1.0000	0.9995	0.9953	0.9648	0.7941
	15							1.0000	1.0000	1.0000	1.0000
16	**0**	0.1853	0.0281	0.0100	0.0033	0.0003	0.0000				
	1	0.5147	0.1407	0.0635	0.0261	0.0033	0.0003	0.0000			
	2	0.7892	0.3518	0.1971	0.0994	0.0183	0.0021	0.0001			
	3	0.9316	0.5981	0.4050	0.2459	0.0651	0.0106	0.0009	0.0000		
	4	0.9830	0.7982	0.6302	0.4499	0.1666	0.0384	0.0049	0.0003		
	5	0.9967	0.9183	0.8103	0.6598	0.3288	0.1051	0.0191	0.0016	0.0000	
	6	0.9995	0.9733	0.9204	0.8247	0.5272	0.2272	0.0583	0.0071	0.0002	
	7	0.9999	0.9930	0.9729	0.9256	0.7161	0.4018	0.1423	0.0257	0.0015	0.0000
	8	1.0000	0.9985	0.9925	0.9743	0.8577	0.5982	0.2839	0.0744	0.0070	0.0001
	9		0.9998	0.9984	0.9929	0.9417	0.7728	0.4728	0.1753	0.0267	0.0005
	10		1.0000	0.9997	0.9984	0.9809	0.8949	0.6712	0.3402	0.0817	0.0033
	11			1.0000	0.9997	0.9951	0.9616	0.8334	0.5501	0.2018	0.0170
	12				1.0000	0.9991	0.9894	0.9349	0.7541	0.4019	0.0684
	13					0.9999	0.9979	0.9817	0.9006	0.6482	0.2108
	14					1.0000	0.9997	0.9967	0.9739	0.8593	0.4853
	15						1.0000	0.9997	0.9967	0.9719	0.8147
	16							1.0000	1.0000	1.0000	1.0000

Tabla A.1 (continuación) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
17	**0**	0.1668	0.0225	0.0075	0.0023	0.0002	0.0000				
	1	0.4818	0.1182	0.0501	0.0193	0.0021	0.0001	0.0000			
	2	0.7618	0.3096	0.1637	0.0774	0.0123	0.0012	0.0001			
	3	0.9174	0.5489	0.3530	0.2019	0.0464	0.0064	0.0005	0.0000		
	4	0.9779	0.7582	0.5739	0.3887	0.1260	0.0245	0.0025	0.0001		
	5	0.9953	0.8943	0.7653	0.5968	0.2639	0.0717	0.0106	0.0007	0.0000	
	6	0.9992	0.9623	0.8929	0.7752	0.4478	0.1662	0.0348	0.0032	0.0001	
	7	0.9999	0.9891	0.9598	0.8954	0.6405	0.3145	0.0919	0.0127	0.0005	
	8	1.0000	0.9974	0.9876	0.9597	0.8011	0.5000	0.1989	0.0403	0.0026	0.0000
	9		0.9995	0.9969	0.9873	0.9081	0.6855	0.3595	0.1046	0.0109	0.0001
	10		0.9999	0.9994	0.9968	0.9652	0.8338	0.5522	0.2248	0.0377	0.0008
	11		1.0000	0.9999	0.9993	0.9894	0.9283	0.7361	0.4032	0.1057	0.0047
	12			1.0000	0.9999	0.9975	0.9755	0.8740	0.6113	0.2418	0.0221
	13				1.0000	0.9995	0.9936	0.9536	0.7981	0.4511	0.0826
	14					0.9999	0.9988	0.9877	0.9226	0.6904	0.2382
	15					1.0000	0.9999	0.9979	0.9807	0.8818	0.5182
	16						1.0000	0.9998	0.9977	0.9775	0.8332
	17							1.0000	1.0000	1.0000	1.0000
18	**0**	0.1501	0.0180	0.0056	0.0016	0.0001	0.0000				
	1	0.4503	0.0991	0.0395	0.0142	0.0013	0.0001				
	2	0.7338	0.2713	0.1353	0.0600	0.0082	0.0007	0.0000			
	3	0.9018	0.5010	0.3057	0.1646	0.0328	0.0038	0.0002			
	4	0.9718	0.7164	0.5187	0.3327	0.0942	0.0154	0.0013	0.0000		
	5	0.9936	0.8671	0.7175	0.5344	0.2088	0.0481	0.0058	0.0003		
	6	0.9988	0.9487	0.8610	0.7217	0.3743	0.1189	0.0203	0.0014	0.0000	
	7	0.9998	0.9837	0.9431	0.8593	0.5634	0.2403	0.0576	0.0061	0.0002	
	8	1.0000	0.9957	0.9807	0.9404	0.7368	0.4073	0.1347	0.0210	0.0009	
	9		0.9991	0.9946	0.9790	0.8653	0.5927	0.2632	0.0596	0.0043	0.0000
	10		0.9998	0.9988	0.9939	0.9424	0.7597	0.4366	0.1407	0.0163	0.0002
	11		1.0000	0.9998	0.9986	0.9797	0.8811	0.6257	0.2783	0.0513	0.0012
	12			1.0000	0.9997	0.9942	0.9519	0.7912	0.4656	0.1329	0.0064
	13				1.0000	0.9987	0.9846	0.9058	0.6673	0.2836	0.0282
	14					0.9998	0.9962	0.9672	0.8354	0.4990	0.0982
	15					1.0000	0.9993	0.9918	0.9400	0.7287	0.2662
	16						0.9999	0.9987	0.9858	0.9009	0.5497
	17						1.0000	0.9999	0.9984	0.9820	0.8499
	18							1.0000	1.0000	1.0000	1.0000

Tabla A.1 (continuación) Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	***r***	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
19	**0**	0.1351	0.0144	0.0042	0.0011	0.0001					
	1	0.4203	0.0829	0.0310	0.0104	0.0008	0.0000				
	2	0.7054	0.2369	0.1113	0.0462	0.0055	0.0004	0.0000			
	3	0.8850	0.4551	0.2631	0.1332	0.0230	0.0022	0.0001			
	4	0.9648	0.6733	0.4654	0.2822	0.0696	0.0096	0.0006	0.0000		
	5	0.9914	0.8369	0.6678	0.4739	0.1629	0.0318	0.0031	0.0001		
	6	0.9983	0.9324	0.8251	0.6655	0.3081	0.0835	0.0116	0.0006		
	7	0.9997	0.9767	0.9225	0.8180	0.4878	0.1796	0.0352	0.0028	0.0000	
	8	1.0000	0.9933	0.9713	0.9161	0.6675	0.3238	0.0885	0.0105	0.0003	
	9		0.9984	0.9911	0.9674	0.8139	0.5000	0.1861	0.0326	0.0016	
	10		0.9997	0.9977	0.9895	0.9115	0.6762	0.3325	0.0839	0.0067	0.0000
	11		1.0000	0.9995	0.9972	0.9648	0.8204	0.5122	0.1820	0.0233	0.0003
	12			0.9999	0.9994	0.9884	0.9165	0.6919	0.3345	0.0676	0.0017
	13			1.0000	0.9999	0.9969	0.9682	0.8371	0.5261	0.1631	0.0086
	14				1.0000	0.9994	0.9904	0.9304	0.7178	0.3267	0.0352
	15					0.9999	0.9978	0.9770	0.8668	0.5449	0.1150
	16					1.0000	0.9996	0.9945	0.9538	0.7631	0.2946
	17						1.0000	0.9992	0.9896	0.9171	0.5797
	18							0.9999	0.9989	0.9856	0.8649
	19							1.0000	1.0000	1.0000	1.0000
20	**0**	0.1216	0.0115	0.0032	0.0008	0.0000					
	1	0.3917	0.0692	0.0243	0.0076	0.0005	0.0000				
	2	0.6769	0.2061	0.0913	0.0355	0.0036	0.0002				
	3	0.8670	0.4114	0.2252	0.1071	0.0160	0.0013	0.0000			
	4	0.9568	0.6296	0.4148	0.2375	0.0510	0.0059	0.0003			
	5	0.9887	0.8042	0.6172	0.4164	0.1256	0.0207	0.0016	0.0000		
	6	0.9976	0.9133	0.7858	0.6080	0.2500	0.0577	0.0065	0.0003		
	7	0.9996	0.9679	0.8982	0.7723	0.4159	0.1316	0.0210	0.0013	0.0000	
	8	0.9999	0.9900	0.9591	0.8867	0.5956	0.2517	0.0565	0.0051	0.0001	
	9	1.0000	0.9974	0.9861	0.9520	0.7553	0.4119	0.1275	0.0171	0.0006	
	10		0.9994	0.9961	0.9829	0.8725	0.5881	0.2447	0.0480	0.0026	0.0000
	11		0.9999	0.9991	0.9949	0.9435	0.7483	0.4044	0.1133	0.0100	0.0001
	12		1.0000	0.9998	0.9987	0.9790	0.8684	0.5841	0.2277	0.0321	0.0004
	13			1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.0867	0.0024
	14				1.0000	0.9984	0.9793	0.8744	0.5836	0.1958	0.0113
	15					0.9997	0.9941	0.9490	0.7625	0.3704	0.0432
	16					1.0000	0.9987	0.9840	0.8929	0.5886	0.1330
	17						0.9998	0.9964	0.9645	0.7939	0.3231
	18						1.0000	0.9995	0.9924	0.9308	0.6083
	19							1.0000	0.9992	0.9885	0.8784
	20								1.0000	1.0000	1.0000

Tabla A.2 Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x;\mu)$

	μ								
r	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865
4		1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977
5				1.0000	1.0000	1.0000	0.9999	0.9998	0.9997
6							1.0000	1.0000	1.0000

	μ								
r	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**4.0**	**4.5**	**5.0**
0	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067
1	0.7358	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404
2	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247
3	0.9810	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650
4	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405
5	0.9994	0.9955	0.9834	0.9580	0.9161	0.8576	0.7851	0.7029	0.6160
6	0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622
7	1.0000	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666
8		1.0000	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319
9			1.0000	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682
10				0.9999	0.9997	0.9990	0.9972	0.9933	0.9863
11				1.0000	0.9999	0.9997	0.9991	0.9976	0.9945
12					1.0000	0.9999	0.9997	0.9992	0.9980
13						1.0000	0.9999	0.9997	0.9993
14							1.0000	0.9999	0.9998
15								1.0000	0.9999
16									1.0000

Tabla A.2 (continuación) Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x;\mu)$

	μ								
r	**5.5**	**6.0**	**6.5**	**7.0**	**7.5**	**8.0**	**8.5**	**9.0**	**9.5**
0	0.0041	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001
1	0.0266	0.0174	0.0113	0.0073	0.0047	0.0030	0.0019	0.0012	0.0008
2	0.0884	0.0620	0.0430	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042
3	0.2017	0.1512	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149
4	0.3575	0.2851	0.2237	0.1730	0.1321	0.0996	0.0744	0.0550	0.0403
5	0.5289	0.4457	0.3690	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885
6	0.6860	0.6063	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649
7	0.8095	0.7440	0.6728	0.5987	0.5246	0.4530	0.3856	0.3239	0.2687
8	0.8944	0.8472	0.7916	0.7291	0.6620	0.5925	0.5231	0.4557	0.3918
9	0.9462	0.9161	0.8774	0.8305	0.7764	0.7166	0.6530	0.5874	0.5218
10	0.9747	0.9574	0.9332	0.9015	0.8622	0.8159	0.7634	0.7060	0.6453
11	0.9890	0.9799	0.9661	0.9467	0.9208	0.8881	0.8487	0.8030	0.7520
12	0.9955	0.9912	0.9840	0.9730	0.9573	0.9362	0.9091	0.8758	0.8364
13	0.9983	0.9964	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981
14	0.9994	0.9986	0.9970	0.9943	0.9897	0.9827	0.9726	0.9585	0.9400
15	0.9998	0.9995	0.9988	0.9976	0.9954	0.9918	0.9862	0.9780	0.9665
16	0.9999	0.9998	0.9996	0.9990	0.9980	0.9963	0.9934	0.9889	0.9823
17	1.0000	0.9999	0.9998	0.9996	0.9992	0.9984	0.9970	0.9947	0.9911
18		1.0000	0.9999	0.9999	0.9997	0.9993	0.9987	0.9976	0.9957
19			1.0000	1.0000	0.9999	0.9997	0.9995	0.9989	0.9980
20						0.9999	0.9998	0.9996	0.9991
21						1.0000	0.9999	0.9998	0.9996
22							1.0000	0.9999	0.9999
23								1.0000	0.9999
24									1.0000

Tabla A.2 (continuación) Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x;\mu)$

	μ								
r	**10.0**	**11.0**	**12.0**	**13.0**	**14.0**	**15.0**	**16.0**	**17.0**	**18.8**
0	0.0000	0.0000	0.0000						
1	0.0005	0.0002	0.0001	0.0000	0.0000				
2	0.0028	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000		
3	0.0103	0.0049	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000
4	0.0293	0.0151	0.0076	0.0037	0.0018	0.0009	0.0004	0.0002	0.0001
5	0.0671	0.0375	0.0203	0.0107	0.0055	0.0028	0.0014	0.0007	0.0003
6	0.1301	0.0786	0.0458	0.0259	0.0142	0.0076	0.0040	0.0021	0.0010
7	0.2202	0.1432	0.0895	0.0540	0.0316	0.0180	0.0100	0.0054	0.0029
8	0.3328	0.2320	0.1550	0.0998	0.0621	0.0374	0.0220	0.0126	0.0071
9	0.4579	0.3405	0.2424	0.1658	0.1094	0.0699	0.0433	0.0261	0.0154
10	0.5830	0.4599	0.3472	0.2517	0.1757	0.1185	0.0774	0.0491	0.0304
11	0.6968	0.5793	0.4616	0.3532	0.2600	0.1848	0.1270	0.0847	0.0549
12	0.7916	0.6887	0.5760	0.4631	0.3585	0.2676	0.1931	0.1350	0.0917
13	0.8645	0.7813	0.6815	0.5730	0.4644	0.3632	0.2745	0.2009	0.1426
14	0.9165	0.8540	0.7720	0.6751	0.5704	0.4657	0.3675	0.2808	0.2081
15	0.9513	0.9074	0.8444	0.7636	0.6694	0.5681	0.4667	0.3715	0.2867
16	0.9730	0.9441	0.8987	0.8355	0.7559	0.6641	0.5660	0.4677	0.3751
17	0.9857	0.9678	0.9370	0.8905	0.8272	0.7489	0.6593	0.5640	0.4686
18	0.9928	0.9823	0.9626	0.9302	0.8826	0.8195	0.7423	0.6550	0.5622
19	0.9965	0.9907	0.9787	0.9573	0.9235	0.8752	0.8122	0.7363	0.6509
20	0.9984	0.9953	0.9884	0.9750	0.9521	0.9170	0.8682	0.8055	0.7307
21	0.9993	0.9977	0.9939	0.9859	0.9712	0.9469	0.9108	0.8615	0.7991
22	0.9997	0.9990	0.9970	0.9924	0.9833	0.9673	0.9418	0.9047	0.8551
23	0.9999	0.9995	0.9985	0.9960	0.9907	0.9805	0.9633	0.9367	0.8989
24	1.0000	0.9998	0.9993	0.9980	0.9950	0.9888	0.9777	0.9594	0.9317
25		0.9999	0.9997	0.9990	0.9974	0.9938	0.9869	0.9748	0.9554
26		1.0000	0.9999	0.9995	0.9987	0.9967	0.9925	0.9848	0.9718
27			0.9999	0.9998	0.9994	0.9983	0.9959	0.9912	0.9827
28			1.0000	0.9999	0.9997	0.9991	0.9978	0.9950	0.9897
29				1.0000	0.9999	0.9996	0.9989	0.9973	0.9941
30					0.9999	0.9998	0.9994	0.9986	0.9967
31					1.0000	0.9999	0.9997	0.9993	0.9982
32						1.0000	0.9999	0.9996	0.9990
33							0.9999	0.9998	0.9995
34							1.0000	0.9999	0.9998
35								1.0000	0.9999
36									0.9999
37									1.0000

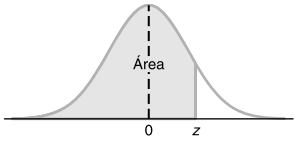


Tabla A.3 Áreas bajo la curva normal

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Tabla A.3 (continuación) Áreas bajo la curva normal

z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

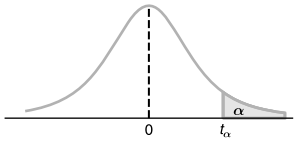


Tabla A.4 Valores críticos de la distribución t

	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Tabla A.4 (continuación) Valores críticos de la distribución t

	α						
v	**0.02**	**0.015**	**0.01**	**0.0075**	**0.005**	**0.0025**	**0.0005**
1	15.894	21.205	31.821	42.433	63.656	127.321	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.600
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.689
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.660
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.290

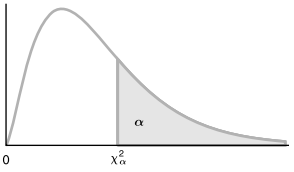


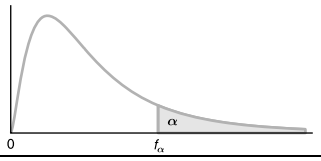
Tabla A.5 Valores críticos de la distribución chi cuadrada

	α									
v	**0.995**	**0.99**	**0.98**	**0.975**	**0.95**	**0.90**	**0.80**	**0.75**	**0.70**	**0.50**
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.647	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	3.565	4.107	4.765	5.009	5.892	7.041	8.634	9.299	9.926	12.340
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.037	11.721	14.339
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	9.260	10.196	11.293	11.689	13.091	14.848	17.187	18.137	19.021	22.337
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867	24.337
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	11.808	12.878	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.336
30	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336
40	20.707	22.164	23.838	24.433	26.509	29.051	32.345	33.66	34.872	39.335
50	27.991	29.707	31.664	32.357	34.764	37.689	41.449	42.942	44.313	49.335
60	35.534	37.485	39.699	40.482	43.188	46.459	50.641	52.294	53.809	59.335

Tabla A.5 (continuación) Valores críticos de la distribución chi cuadrada

	α									
v	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.466
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.515
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.321
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.471	27.688	29.819	34.527
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.124
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.698
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.791
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.819
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.314
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.796
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.619
26	29.246	30.435	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.051
27	30.319	31.528	32.912	36.741	40.113	43.195	44.140	46.963	49.645	55.475
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.994	56.892
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.335	58.301
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.702
40	44.165	45.616	47.269	51.805	55.758	59.342	60.436	63.691	66.766	73.403
50	54.723	56.334	58.164	63.167	67.505	71.420	72.613	76.154	79.490	86.660
60	65.226	66.981	68.972	74.397	79.082	83.298	84.58	88.379	91.952	99.608

Tabla A.6* Valores críticos de la distribución F



	$f_{0.05}(v_1, v_2)$								
v_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

*Reproducida de la tabla 18 de *Biometrika Tables for Statisticians*, vol. 1, con autorización de E. S. Pearson y Biometrika Trustees.

Tabla A.6 (continuación) Valores críticos de la distribución F

	$f_{0.05}(v_1, v_2)$									
v_2	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Tabla A.6 (continuación) Valores críticos de la distribución F

	$f_{0.01}(v_1, v_2)$								
v_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Tabla A.6 (continuación) Valores críticos de la distribución F

	$f_{0.01}(v_1, v_2)$									
v_2	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	6055.85	6106.32	6157.28	6208.73	6234.63	6260.65	6286.78	6313.03	6339.39	6365.86
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Tabla A.7* Factores de tolerancia para distribuciones normales

	Intervalos bilaterales						**Intervalos unilaterales**					
	$\gamma = 0.05$			$\gamma = 0.01$			$\gamma = 0.05$			$\gamma = 0.01$		
	$1 - \alpha$			$1 - \alpha$			$1 - \alpha$			$1 - \alpha$		
n	**0.90**	**0.95**	**0.99**	**0.90**	**0.95**	**0.99**	**0.90**	**0.95**	**0.99**	**0.90**	**0.95**	**0.99**
2	32.019	37.674	48.430	160.193	188.491	242.300	20.581	26.260	37.094	103.029	131.426	185.617
3	8.380	9.916	12.861	18.930	22.401	29.055	6.156	7.656	10.553	13.995	17.170	23.896
4	5.369	6.370	8.299	9.398	11.150	14.527	4.162	5.144	7.042	7.380	9.083	12.387
5	4.275	5.079	6.634	6.612	7.855	10.260	3.407	4.203	5.741	5.362	6.578	8.939
6	3.712	4.414	5.775	5.337	6.345	8.301	3.006	3.708	5.062	4.411	5.406	7.335
7	3.369	4.007	5.248	4.613	5.488	7.187	2.756	3.400	4.642	3.859	4.728	6.412
8	3.136	3.732	4.891	4.147	4.936	6.468	2.582	3.187	4.354	3.497	4.285	5.812
9	2.967	3.532	4.631	3.822	4.550	5.966	2.454	3.031	4.143	3.241	3.972	5.389
10	2.839	3.379	4.433	3.582	4.265	5.594	2.355	2.911	3.981	3.048	3.738	5.074
11	2.737	3.259	4.277	3.397	4.045	5.308	2.275	2.815	3.852	2.898	3.556	4.829
12	2.655	3.162	4.150	3.250	3.870	5.079	2.210	2.736	3.747	2.777	3.410	4.633
13	2.587	3.081	4.044	3.130	3.727	4.893	2.155	2.671	3.659	2.677	3.290	4.472
14	2.529	3.012	3.955	3.029	3.608	4.737	2.109	2.615	3.585	2.593	1.189	4.337
15	2.480	2.954	3.878	2.945	3.507	4.605	2.068	2.566	3.520	2.522	3.102	4.222
16	2.437	2.903	3.812	2.872	3.421	4.492	2.033	2.524	3.464	2.460	3.028	4.123
17	2.400	2.858	3.754	2.808	3.345	4.393	2.002	2.486	3.414	2.405	2.963	4.037
18	2.366	2.819	3.702	2.753	3.279	4.307	1.974	2.453	3.370	2.357	2.905	3.960
19	2.337	2.784	3.656	2.703	3.221	4.230	1.949	2.423	3.331	2.314	2.854	3.892
20	2.310	2.752	3.615	2.659	3.168	4.161	1.926	2.396	3.295	2.276	2.808	1.832
25	2.208	2.631	3.457	2.494	2.972	3.904	1.838	2.292	3.158	2.129	2.633	3.001
30	2.140	2.549	3.350	2.385	2.841	3.733	1.777	2.220	3.064	2.030	2.516	3.447
35	2.090	2.490	3.272	2.306	2.748	3.611	1.732	2.167	2.995	1.957	2.430	3.334
40	2.052	2.445	3.213	2.247	2.677	3.518	1.697	2.126	2.941	1.902	2.364	3.249
45	2.021	2.408	3.165	2.200	2.621	3.444	1.669	2.092	2.898	1.857	2.312	3.180
50	1.996	2.379	3.126	2.162	2.576	3.385	1.646	2.065	2.863	1.821	2.269	3.125
60	1.958	2.333	3.066	2.103	2.506	3.293	1.609	2.022	2.807	1.764	2.202	3.038
70	1.929	2.299	3.021	2.060	2.454	3.225	1.581	1.990	2.765	1.722	2.153	2.974
80	1.907	2.272	2.986	2.026	2.414	3.173	1.559	1.965	2.733	1.688	2.114	2.924
90	1.889	2.251	2.958	1.999	2.382	3.130	1.542	1.944	2.706	1.661	2.082	2.883
100	1.874	2.233	2.934	1.977	2.355	3.096	1.527	1.927	2.684	1.639	2.056	2.850
150	1.825	2.175	2.859	1.905	2.270	2.983	1.478	1.870	2.611	1.566	1.971	2.741
200	1.798	2.143	2.816	1.865	2.222	2.921	1.450	1.837	2.570	1.524	1.923	2.679
250	1.780	2.121	2.788	1.839	2.191	2.880	1.431	1.815	2.542	1.496	1.891	2.638
300	1.767	2.106	2.767	1.820	2.169	2.850	1.417	1.800	2.522	1.476	1.868	2.608
∞	1.645	1.960	2.576	1.645	1.960	2.576	1.282	1.645	2.326	1.282	1.645	2.326

*Adaptada de C. Eisenhart, M. W. Hastay y W. A. Wallis, *Techniques of Statistical Analysis*, capítulo 2, McGraw-Hill Book Company, Nueva York, 1947. Se utiliza con autorización de McGraw-Hill Book Company.

Tabla A.8* Tamaño muestral para la prueba t de la media

		Nivel de la prueba t																			
Prueba unilateral		**$\alpha = 0.005$**					**$\alpha = 0.01$**					**$\alpha = 0.025$**					**$\alpha = 0.05$**				
Prueba bilateral		**$\alpha = 0.01$**					**$\alpha = 0.02$**					**$\alpha = 0.05$**					**$\alpha = 0.1$**				
	$\beta = 0.1$	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**
	0.05																				
	0.10																				
	0.15																				122
	0.20										139					99					70
	0.25					110					90				128	64			139	101	45
	0.30				134	78				115	63			119	90	45		122	97	71	32
	0.35			125	99	58			109	85	47		109	88	67	34		90	72	52	24
	0.40		115	97	77	45		101	85	66	37	117	84	68	51	26	101	70	55	40	19
	0.45		92	77	62	37	110	81	68	53	30	93	67	54	41	21	80	55	44	33	15
	0.50	100	75	63	51	30	90	66	55	43	25	76	54	44	34	18	65	45	36	27	13
	0.55	83	63	53	42	26	75	55	46	36	21	63	45	37	28	15	54	38	30	22	11
	0.60	71	53	45	36	22	63	47	39	31	18	53	38	32	24	13	46	32	26	19	9
	0.65	61	46	39	31	20	55	41	34	27	16	46	33	27	21	12	39	28	22	17	8
	0.70	53	40	34	28	17	47	35	30	24	14	40	29	24	19	10	34	24	19	15	8
	0.75	47	36	30	25	16	42	31	27	21	13	35	26	21	16	9	30	21	17	13	7
	0.80	41	32	27	22	14	37	28	24	19	12	31	22	19	15	9	27	19	15	12	6
	0.85	37	29	24	20	13	33	25	21	17	11	28	21	17	13	8	24	17	14	11	6
	0.90	34	26	22	18	12	29	23	19	16	10	25	19	16	12	7	21	15	13	10	5
Valor de	**0.95**	31	24	20	17	11	27	21	18	14	9	23	17	14	11	7	19	14	11	9	5
$\Delta = \lvert\delta\rvert/\sigma$	**1.00**	28	22	19	16	10	25	19	16	13	9	21	16	13	10	6	18	13	11	8	5
	1.1	24	19	16	14	9	21	16	14	12	8	18	13	11	9	6		15	11	9	7
	1.2	21	16	14	12	8	18	14	12	10	7	15	12	10	8	5		13	10	8	6
	1.3	18	15	13	11	8	16	13	11	9	6		14	10	9	7		11	8	7	6
	1.4	16	13	12	10	7	14	11	10	9	6	12	9	8	7		10	8	7	5	
	1.5	15	12	11	9	7	13	10	9	8	6	11	8	7	6			9	7	6	
	1.6	13	11	10	8	6	12	10	9	7	5		10	8	7	6			8	6	6
	1.7	12	10	9	8	6		11	9	8	7		9	7	6	5			8	6	5
	1.8	12	10	9	8	6		10	8	7	7			8	7	6				7	6
	1.9	11	9	8	7	6		10	8	7	6			8	6	6				7	5
	2.0	10	8	8	7	5		9	7	7	6			7	6	5					6
	2.1		10	8	7	7		8	7	6	6				7	6					6
	2.2		9	8	7	6		8	7	6	5				7	6					6
	2.3		9	7	7	6			8	6	6				6	5					5
	2.4		8	7	7	6			7	6	6					6					
	2.5		8	7	6	6			7	6	6					6					
	3.0		7	6	6	5			6	5	5					5					
	3.5			6	5	5					5										
	4.0					6															

*Reproducida con autorización de O. L. Davies, ed., *Design and Analysis of Industrial Experiments*, Oliver & Boyd, Edimburgo, 1956.

Tabla A.9* Tamaño muestral para la prueba t de la diferencia entre dos medias

		Nivel de la prueba t																			
Prueba unilateral		**$\alpha = 0.005$**					**$\alpha = 0.01$**					**$\alpha = 0.025$**					**$\alpha = 0.05$**				
Prueba bilateral		**$\alpha = 0.01$**					**$\alpha = 0.02$**					**$\alpha = 0.05$**					**$\alpha = 0.1$**				
	$\beta = 0$ 1	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**	**.01**	**.05**	**.1**	**.2**	**.5**
	0.05																				137
	0.10															124					88
	0.15										123					87					61
	0.20					110					90					64				102	45
	0.25					85					70				100	50			108	78	35
	0.30				118	68				101	55			105	79	39		108	86	62	28
	0.35				96	55			106	82	45		106	86	64	32		88	70	51	23
	0.40			101	79	46		106	88	68	38		87	71	53	27	112	73	58	42	19
	0.45		101	85	67	39		90	74	58	32	104	74	60	45	23	89	61	49	36	16
	0.50		87	73	57	34	104	77	64	49	27	88	63	51	39	20	76	52	42	30	14
	0.55	100	75	63	50	29	90	66	55	43	24	76	55	44	34	17	66	45	36	26	12
	0.60	88	66	55	44	26	79	58	48	38	21	67	48	39	29	15	57	40	32	23	11
	0.65	77	58	49	39	23	70	51	43	33	19	59	42	34	26	14	50	35	28	21	10
	0.70	69	51	43	35	21	62	46	38	30	17	52	37	31	23	12	45	31	25	18	9
	0.75	62	46	39	31	19	55	41	34	27	15	47	34	27	21	11	40	28	22	16	8
	0.80	55	42	35	28	17	50	37	31	24	14	42	30	25	19	10	36	25	20	15	7
	0.85	50	38	32	26	15	45	33	28	22	13	38	27	23	17	9	33	23	18	14	7
	0.90	42	32	27	22	13	38	28	23	19	11	32	23	19	14	8	27	19	15	12	6
	0.95	36	27	23	18	11	32	24	20	16	9	27	20	16	12	7	23	16	13	10	5
	1.00	31	23	20	16	10	28	21	17	14	8	23	17	14	11	6	20	14	11	9	5
	1.1	27	20	17	14	9	24	18	15	12	8	20	15	12	10	6	17	12	10	8	4
	1.2	24	18	15	13	8	21	16	14	11	7	18	13	11	9	5	15	11	9	7	4
	1.3	21	16	14	11	7	19	14	12	10	6	16	12	10	8	5	14	10	8	6	4
	1.4	19	15	13	10	7	17	13	11	9	6	14	11	9	7	4	12	9	7	6	3
	1.5	17	13	71	10	6	15	12	10	8	5	13	10	8	6	4	11	8	7	5	
	1.6	16	12	11	9	6	14	11	9	8	5	12	9	7	6	4	10	7	6	5	
	1.7	14	11	10	8	6	13	10	9	7	5	11	8	7	6	4	9	7	6	4	
	1.8	13	10	9	8	5	12	9	8	7	5	10	8	6	5	3	8	6	5	4	
	1.9	12	10	8	7	5	11	9	7	6	4	9	7	6	5		8	6	5	4	
	2.0	11	9	8	7	5	10	8	7	6	4	9	7	6	5		7	5	5	4	
	2.1	11	9	8	6	5	10	8	7	6	48	6	5	4		7	5	4	4		
	2.2	10	8	7	6	4	9	7	6	5	4	8	6	5	4		6	5	4	3	
	2.3	8	6	6	5	4	7	6	5	4	3	6	5	4	4		5	4	3		
	2.4	6	5	5	4	3	6	5	4	4	5	4	4	3		4	3				
	2.5	6	5	4	4		5	4	4	3	4	4	3			4					
	3.0																				
	3.5																				
	4.0																				

Valor de $\Delta = |\delta|/\sigma$

*Reproducida con autorización de O. L. Davies, ed., *Design and Analysis of Industrial Experiments*, Oliver & Boyd, Edimburgo, 1956.

Tabla A.10* Valores críticos para la prueba de Bartlett

	$b_k(0.01; n)$								
	Número de poblaciones, k								
n	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	0.1411	0.1672							
4	0.2843	0.3165	0.3475	0.3729	0.3937	0.4110			
5	0.3984	0.4304	0.4607	0.4850	0.5046	0.5207	0.5343	0.5458	0.5558
6	0.4850	0.5149	0.5430	0.5653	0.5832	0.5978	0.6100	0.6204	0.6 293
7	0.5512	0.5787	0.6045	0.6248	0.6410	0.6542	0.6652	0.6744	0.6824
8	0.6031	0.6282	0.6518	0.6704	0.6851	0.6970	0.7069	0.7153	0.7225
9	0.6445	0.6676	0.6892	0.7062	0.7197	0.7305	0.7395	0.7471	0.7536
10	0.6783	0.6996	0.7195	0.7352	0.7475	0.7575	0.7657	0.7726	0.7786
11	0.7063	0.7260	0.7445	0.7590	0.7703	0.7795	0.7871	0.7935	0.7990
12	0.7299	0.7483	0.7654	0.7789	0.7894	0.7980	0.8050	0.8109	0.8160
13	0.7501	0.7672	0.7832	0.7958	0.8056	0.8135	0.8201	0.8256	0.8303
14	0.7674	0.7835	0.7985	0.8103	0.8195	0.8269	0.8330	0.8382	0.8426
15	0.7825	0.7977	0.8118	0.8229	0.8315	0.8385	0.8443	0.8491	0.8532
16	0.7958	0.8101	0.8235	0.8339	0.8421	0.8486	0.8541	0.8586	0.8625
17	0.8076	0.8211	0.8338	0.8436	0.8514	0.8576	0.8627	0.8670	0.8707
18	0.8181	0.8309	0.8429	0.8523	0.8596	0.8655	0.8704	0.8745	0.8780
19	0.8275	0.8397	0.8512	0.8601	0.8670	0.8727	0.8773	0.8811	0.8845
20	0.8360	0.8476	0.8586	0.8671	0.8737	0.8791	0.8835	0.8871	0.8903
21	0.8437	0.8548	0.8653	0.8734	0.8797	0.8848	0.8890	0.8926	0.8956
22	0.8507	0.8614	0.8714	0.8791	0.8852	0.8901	0.8941	0.8975	0.9004
23	0.8571	0.8673	0.8769	0.8844	0.8902	0.8949	0.8988	0.9020	0.9047
24	0.8630	0.8728	0.8820	0.8892	0.8948	0.8993	0.9030	0.9061	0.9087
25	0.8684	0.8779	0.8867	0.8936	0.8990	0.9034	0.9069	0.9099	0.9124
26	0.8734	0.8825	0.8911	0.8977	0.9029	0.9071	0.9105	0.9134	0.9158
27	0.8781	0.8869	0.8951	0.9015	0.9065	0.9105	0.9138	0.9166	0.9190
28	0.8824	0.8909	0.8988	0.9050	0.9099	0.9138	0.9169	0.9196	0.9219
29	0.8864	0.8946	0.9023	0.9083	0.9130	0.9167	0.9198	0.9224	0.9246
30	0.8902	0.8981	0.9056	0.9114	0.9159	0.9195	0.9225	0.9250	0.9271
40	0.9175	0.9235	0.9291	0.9335	0.9370	0.9397	0.9420	0.9439	0.9455
50	0.9339	0.9387	0.9433	0.9468	0.9496	0.9518	0.9536	0.9551	0.9564
60	0.9449	0.9489	0.9527	0.9557	0.9580	0.9599	0.9614	0.9626	0.9637
80	0.9586	0.9617	0.9646	0.9668	0.9685	0.9699	0.9711	0.9720	0.9728
100	0.9669	0.9693	0.9716	0.9734	0.9748	0.9759	0.9769	0.9776	0.9783

*Reproducida de D. D. Dyer y J. P. Keating, "On the Determination of Critical Values for Bartlett's Test", *J. Am. Stat. Assoc.*, **75**, 1980, con autorización del consejo de directores.

Tabla A.10 (continuación) Valores críticos para la prueba de Bartlett

	$b_k(0.05; n)$								
	Número de poblaciones, k								
n	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	0.3123	0.3058	0.3173	0.3299					
4	0.4780	0.4699	0.4803	0.4921	0.5028	0.5122	0.5204	0.5277	0.5341
5	0.5845	0.5762	0.5850	0.5952	0.6045	0.6126	0.6197	0.6260	0.6315
6	0.6563	0.6483	0.6559	0.6646	0.6727	0.6798	0.6860	0.6914	0.6961
7	0.7075	0.7000	0.7065	0.7142	0.7213	0.7275	0.7329	0.7376	0.7418
8	0.7456	0.7387	0.7444	0.7512	0.7574	0.7629	0.7677	0.7719	0.7757
9	0.7751	0.7686	0.7737	0.7798	0.7854	0.7903	0.7946	0.7984	0.8017
10	0.7984	0.7924	0.7970	0.8025	0.8076	0.8121	0.8160	0.8194	0.8224
11	0.8175	0.8118	0.8160	0.8210	0.8257	0.8298	0.8333	0.8365	0.8392
12	0.8332	0.8280	0.8317	0.8364	0.8407	0.8444	0.8477	0.8506	0.8531
13	0.8465	0.8415	0.8450	0.8493	0.8533	0.8568	0.8598	0.8625	0.8648
14	0.8578	0.8532	0.8564	0.8604	0.8641	0.8673	0.8701	0.8726	0.8748
15	0.8676	0.8632	0.8662	0.8699	0.8734	0.8764	0.8790	0.8814	0.8834
16	0.8761	0.8719	0.8747	0.8782	0.8815	0.8843	0.8868	0.8890	0.8909
17	0.8836	0.8796	0.8823	0.8856	0.8886	0.8913	0.8936	0.8957	0.8975
18	0.8902	0.8865	0.8890	0.8921	0.8949	0.8975	0.8997	0.9016	0.9033
19	0.8961	0.8926	0.8949	0.8979	0.9006	0.9030	0.9051	0.9069	0.9086
20	0.9015	0.8980	0.9003	0.9031	0.9057	0.9080	0.9100	0.9117	0.9132
21	0.9063	0.9030	0.9051	0.9078	0.9103	0.9124	0.9143	0.9160	0.9175
22	0.9106	0.9075	0.9095	0.9120	0.9144	0.9165	0.9183	0.9199	0.9213
23	0.9146	0.9116	0.9135	0.9159	0.9182	0.9202	0.9219	0.9235	0.9248
24	0.9182	0.9153	0.9172	0.9195	0.9217	0.9236	0.9253	0.9267	0.9280
25	0.9216	0.9187	0.9205	0.9228	0.9249	0.9267	0.9283	0.9297	0.9309
26	0.9246	0.9219	0.9236	0.9258	0.9278	0.9296	0.9311	0.9325	0.9336
27	0.9275	0.9249	0.9265	0.9286	0.9305	0.9322	0.9337	0.9350	0.9361
28	0.9301	0.9276	0.9292	0.9312	0.9330	0.9347	0.9361	0.9374	0.9385
29	0.9326	0.9301	0.9316	0.9336	0.9354	0.9370	0.9383	0.9396	0.9406
30	0.9348	0.9325	0.9340	0.9358	0.9376	0.9391	0.9404	0.9416	0.9426
40	0.9513	0.9495	0.9506	0.9520	0.9533	0.9545	0.9555	0.9564	0.9572
50	0.9612	0.9597	0.9606	0.9617	0.9628	0.9637	0.9645	0.9652	0.9658
60	0.9677	0.9665	0.9672	0.9681	0.9690	0.9698	0.9705	0.9710	0.9716
80	0.9758	0.9749	0.9754	0.9761	0.9768	0.9774	0.9779	0.9783	0.9787
100	0.9807	0.9799	0.9804	0.9809	0.9815	0.9819	0.9823	0.9827	0.9830

Tabla A.11* Valores críticos para la prueba de Cochran

$\alpha = 0.01$														
	n													
k	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**17**	**37**	**145**	∞
2	0.9999	0.9950	0.9794	0.9586	0.9373	0.9172	0.8988	0.8823	0.8674	0.8539	0.7949	0.7067	0.6062	0.5000
3	0.9933	0.9423	0.8831	0.8335	0.7933	0.7606	0.7335	0.7107	0.6912	0.6743	0.6059	0.5153	0.4230	0.3333
4	0.9676	0.8643	0.7814	0.7212	0.6761	0.6410	0.6129	0.5897	0.5702	0.5536	0.4884	0.4057	0.3251	0.2500
5	0.9279	0.7885	0.6957	0.6329	0.5875	0.5531	0.5259	0.5037	0.4854	0.4697	0.4094	0.3351	0.2644	0.2000
6	0.8828	0.7218	0.6258	0.5635	0.5195	0.4866	0.4608	0.4401	0.4229	0.4084	0.3529	0.2858	0.2229	0.1667
7	0.8376	0.6644	0.5685	0.5080	0.4659	0.4347	0.4105	0.3911	0.3751	0.3616	0.3105	0.2494	0.1929	0.1429
8	0.7945	0.6152	0.5209	0.4627	0.4226	0.3932	0.3704	0.3522	0.3373	0.3248	0.2779	0.2214	0.1700	0.1250
9	0.7544	0.5727	0.4810	0.4251	0.3870	0.3592	0.3378	0.3207	0.3067	0.2950	0.2514	0.1992	0.1521	0.1111
10	0.7175	0.5358	0.4469	0.3934	0.3572	0.3308	0.3106	0.2945	0.2813	0.2704	0.2297	0.1811	0.1376	0.1000
12	0.6528	0.4751	0.3919	0.3428	0.3099	0.2861	0.2680	0.2535	0.2419	0.2320	0.1961	0.1535	0.1157	0.0833
15	0.5747	0.4069	0.3317	0.2882	0.2593	0.2386	0.2228	0.2104	0.2002	0.1918	0.1612	0.1251	0.0934	0.0667
20	0.4799	0.3297	0.2654	0.2288	0.2048	0.1877	0.1748	0.1646	0.1567	0.1501	0.1248	0.0960	0.0709	0.0500
24	0.4247	0.2871	0.2295	0.1970	0.1759	0.1608	0.1495	0.1406	0.1338	0.1283	0.1060	0.0810	0.0595	0.0417
30	0.3632	0.2412	0.1913	0.1635	0.1454	0.1327	0.1232	0.1157	0.1100	0.1054	0.0867	0.0658	0.0480	0.0333
40	0.2940	0.1915	0.1508	0.1281	0.1135	0.1033	0.0957	0.0898	0.0853	0.0816	0.0668	0.0503	0.0363	0.0250
60	0.2151	0.1371	0.1069	0.0902	0.0796	0.0722	0.0668	0.0625	0.0594	0.0567	0.0461	0.0344	0.0245	0.0167
120	0.1225	0.0759	0.0585	0.0489	0.0429	0.0387	0.0357	0.0334	0.0316	0.0302	0.0242	0.0178	0.0125	0.0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

*Reproducida de C. Eisenhart, M. W. Hastay y W. A. Wallis, *Techniques of Statistical Analysis*, capítulo 15, McGraw-Hill Book Company, Nueva York, 1947. Utilizada con autorización de McGraw-Hill Book Company.

Tabla A.11 (continuación) Valores críticos para la prueba de Cochran

$\alpha = 0.05$														
	n													
k	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**17**	**37**	**145**	∞
2	0.9985	0.9750	0.9392	0.9057	0.8772	0.8534	0.8332	0.8159	0.8010	0.7880	0.7341	0.6602	0.5813	0.5000
3	0.9669	0.8709	0.7977	0.7457	0.7071	0.6771	0.6530	0.6333	0.6167	0.6025	0.5466	0.4748	0.4031	0.3333
4	0.9065	0.7679	0.6841	0.6287	0.5895	0.5598	0.5365	0.5175	0.5017	0.4884	0.4366	0.3720	0.3093	0.2500
5	0.8412	0.6838	0.5981	0.5441	0.5065	0.4783	0.4564	0.4387	0.4241	0.4118	0.3645	0.3066	0.2513	0.2000
6	0.7808	0.6161	0.5321	0.4803	0.4447	0.4184	0.3980	0.3817	0.3682	0.3568	0.3135	0.2612	0.2119	0.1667
7	0.7271	0.5612	0.4800	0.4307	0.3974	0.3726	0.3535	0.3384	0.3259	0.3154	0.2756	0.2278	0.1833	0.1429
8	0.6798	0.5157	0.4377	0.3910	0.3595	0.3362	0.3185	0.3043	0.2926	0.2829	0.2462	0.2022	0.1616	0.1250
9	0.6385	0.4775	0.4027	0.3584	0.3286	0.3067	0.2901	0.2768	0.2659	0.2568	0.2226	0.1820	0.1446	0.1111
10	6.6020	0.4450	0.3733	0.3311	0.3029	0.2823	0.2666	0.2541	0.2439	0.2353	0.2032	0.1655	0.1308	0.1000
12	0.5410	0.3924	0.3264	0.2880	0.2624	0.2439	0.2299	0.2187	0.2098	0.2020	0.1737	0.1403	0.1100	0.0833
15	0.4709	0.3346	0.2758	0.2419	0.2195	0.2034	0.1911	0.1815	0.1736	0.1671	0.1429	0.1144	0.0889	0.0667
20	0.3894	0.2705	0.2205	0.1921	0.1735	0.1602	0.1501	0.1422	0.1357	0.1303	0.1108	0.0879	0.0675	0.0500
24	0.3434	0.2354	0.1907	0.1656	0.1493	0.1374	0.1286	0.1216	0.1160	0.1113	0.0942	0.0743	0.0567	0.0417
30	0.2929	0.1980	0.1593	0.1377	0.1237	0.1137	0.1061	0.1002	0.0958	0.0921	0.0771	0.0604	0.0457	0.0333
40	0.2370	0.1576	0.1259	0.1082	0.0968	0.0887	0.0827	0.0780	0.0745	0.0713	0.0595	0.0462	0.0347	0.0250
60	0.1737	0.1131	0.0895	0.0765	0.0682	0.0623	0.0583	0.0552	0.0520	0.0497	0.0411	0.0316	0.0234	0.0167
120	0.0998	0.0632	0.0495	0.0419	0.0371	0.0337	0.0312	0.0292	0.0279	0.0266	0.0218	0.0165	0.0120	0.0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla A.12 Puntos porcentuales superiores de la distribución de rangos studentizados: Valores de $q(0.05;\, k,\, v)$

Grados de	**Número de tratamientos,** k								
libertad, v	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	18.0	27.0	32.8	37.2	40.5	43.1	15.1	47.1	49.1
2	6.09	5.33	9.80	10.89	11.73	12.43	13.03	13.54	13.99
3	4.50	5.91	6.83	7.51	8.04	8.47	8.85	9.18	9.46
4	3.93	5.04	5.76	6.29	6.71	7.06	7.35	7.60	7.83
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.35	5.59	5.80	5.99	6.15
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74
10	3.15	3.88	4.33	4.66	4.91	5.12	5.30	5.46	5.60
11	3.11	3.82	4.26	4.58	4.82	5.03	5.20	5.35	5.49
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40
13	3.06	3.73	4.15	4.46	4.69	4.88	5.05	5.19	5.32
14	3.03	3.70	4.11	4.41	4.65	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
16	3.00	3.65	4.05	4.34	4.56	4.74	4.90	5.03	5.05
17	2.98	3.62	4.02	4.31	4.52	4.70	4.86	4.99	5.11
18	2.97	3.61	4.00	4.28	4.49	4.67	4.83	4.96	5.07
19	2.96	3.59	3.98	4.26	4.47	4.64	4.79	4.92	5.04
20	2.95	3.58	3.96	4.24	4.45	4.62	4.77	4.90	5.01
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92
30	2.89	3.48	3.84	4.11	4.30	4.46	4.60	4.72	4.83
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
120	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.47	4.56
∞	2.77	3.32	3.63	3.86	4.03	4.17	4.29	4.39	4.47

Tabla A.13* Rangos studentizados significativos mínimos $r_p(0.05; p, v)$

$\alpha = 0.05$									
	p								
v	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.927	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.797	3.814	3.814	3.814	3.814	3.814	3.814
6	3.461	3.587	3.649	3.68	3.694	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.626	3.626	3.626
8	3.261	3.399	3.475	3.521	3.549	3.566	3.575	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522
11	3.113	3.256	3.342	3.397	3.435	3.462	3.48	3.493	3.501
12	3.082	3.225	3.313	3.370	3.410	3.439	3.459	3.474	3.484
13	3.055	3.200	3.289	3.348	3.389	3.419	3.442	3.458	3.470
14	3.033	3.178	3.268	3.329	3.372	3.403	3.426	3.444	3.457
15	3.014	3.160	3.25	3.312	3.356	3.389	3.413	3.432	3.446
16	2.998	3.144	3.235	3.298	3.343	3.376	3.402	3.422	3.437
17	2.984	3.130	3.222	3.285	3.331	3.366	3.392	3.412	3.429
18	2.971	3.118	3.210	3.274	3.321	3.356	3.383	3.405	3.421
19	2.960	3.107	3.199	3.264	3.311	3.347	3.375	3.397	3.415
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.391	3.409
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371
40	2.858	3.006	3.102	3.171	3.224	3.266	3.300	3.328	3.352
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.287	3.314
∞	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294

*Condensada de H. Leon Harter, "Critical Values for Duncan's New Multiple Range Test", *Biometrics*, **16**, núm. 4, 1960, con autorización del autor y del editor

Tabla A.13 (continuación) Rangos studentizados significativos mínimos $r_p(0.01; p, v)$

$\alpha = 0.01$									
	p								
v	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03
2	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04
3	8.261	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321
4	6.512	6.677	6.740	6.756	6.756	6.756	6.756	6.756	6.756
5	5.702	5.893	5.989	6.040	6.065	6.074	6.074	6.074	6.074
6	5.243	5.439	5.549	5.614	5.655	5.680	5.694	5.701	5.703
7	4.949	5.145	5.260	5.334	5.383	5.416	5.439	5.454	5.464
8	4.746	4.939	5.057	5.135	5.189	5.227	5.256	5.276	5.291
9	4.596	4.787	4.906	4.986	5.043	5.086	5.118	5.142	5.160
10	4.482	4.671	4.790	4.871	4.931	4.975	5.010	5.037	5.058
11	4.392	4.579	4.697	4.780	4.841	4.887	4.924	4.952	4.975
12	4.320	4.504	4.622	4.706	4.767	4.815	4.852	4.883	4.907
13	4.260	4.442	4.560	4.644	4.706	4.755	4.793	4.824	4.850
14	4.210	4.391	4.508	4.591	4.654	4.704	4.743	4.775	4.802
15	4.168	4.347	4.463	4.547	4.610	4.660	4.700	4.733	4.760
16	4.131	4.309	4.425	4.509	4.572	4.622	4.663	4.696	4.724
17	4.099	4.275	4.391	4.475	4.539	4.589	4.630	4.664	4.693
18	4.071	4.246	4.362	4.445	4.509	4.560	4.601	4.635	4.664
19	4.046	4.220	4.335	4.419	4.483	4.534	4.575	4.610	4.639
20	4.024	4.197	4.312	4.395	4.459	4.510	4.552	4.587	4.617
24	3.956	4.126	4.239	4.322	4.386	4.437	4.480	4.516	4.546
30	3.889	4.056	4.168	4.250	4.314	4.366	4.409	4.445	4.477
40	3.825	3.988	4.098	4.180	4.244	4.296	4.339	4.376	4.408
60	3.762	3.922	4.031	4.111	4.174	4.226	4.270	4.307	4.340
120	3.702	3.858	3.965	4.044	4.107	4.158	4.202	4.239	4.272
∞	3.643	3.796	3.900	3.978	4.040	4.091	4.135	4.172	4.205

Tabla A.14* Valores de $d_{\alpha/2}(k,v)$ para comparaciones bilaterales entre k tratamientos y un control

$\alpha = 0.05$									
	k = número de medias de tratamiento (excluye el control)								
v	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5	2.57	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97
6	2.45	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71
7	2.36	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53
8	2.31	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41
9	2.26	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32
10	2.23	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24
11	2.20	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19
12	2.18	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14
13	2.16	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10
14	2.14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07
15	2.13	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04
16	2.12	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02
17	2.11	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00
18	2.10	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98
19	2.09	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96
20	2.09	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95
24	2.06	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90
30	2.04	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86
40	2.02	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81
60	2.00	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77
120	1.98	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73
∞	1.96	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69

*Reproducida de Charles W. Dunnett, "New Tables for Multiple Comparison with a Control", *Biometrics*, **20**, núm. 3, 1964, con autorización del autor y del editor.

Tabla A.14 (continuación) Valores de $d_{\alpha/2}(k,v)$ para comparaciones bilaterales entre k tratamientos y un control

$\alpha = 0.01$									
	k = número de medias de tratamiento (excluye el control)								
v	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5	4.03	4.63	4.98	5.22	5.41	5.56	5.69	5.80	5.89
6	3.71	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28
7	3.50	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89
8	3.36	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62
9	3.25	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43
10	3.17	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28
11	3.11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16
12	3.05	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07
13	3.01	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99
14	2.98	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93
15	2.95	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88
16	2.92	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83
17	2.90	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79
18	2.88	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75
19	2.86	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72
20	2.85	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69
24	2.80	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61
30	2.75	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52
40	2.70	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44
60	2.66	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37
120	2.62	2.85	2.97	3.06	3.12	3.18	3.22	3.26	3.29
∞	2.58	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22

Tabla A.15* Valores de $d_{\alpha/2}(k, v)$ para comparaciones unilaterales entre k tratamientos y un control

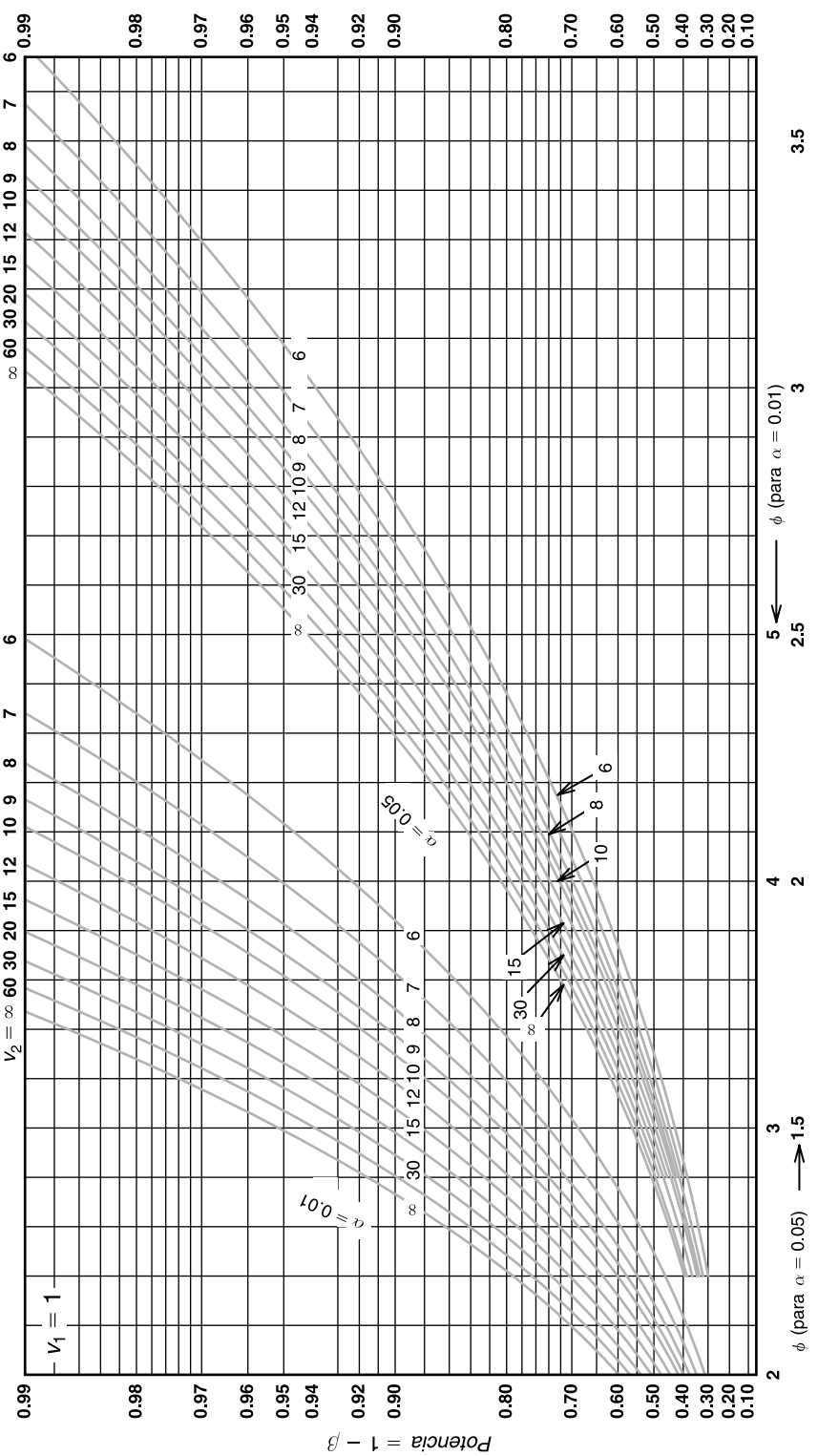
	k = número de medias de tratamiento (excluye el control)								
v	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5	2.02	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30
6	1.94	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12
7	1.89	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01
8	1.86	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92
9	1.83	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86
10	1.81	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81
11	1.80	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77
12	1.78	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74
13	1.77	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71
14	1.76	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69
15	1.75	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67
16	1.75	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65
17	1.74	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64
18	1.73	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62
19	1.73	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61
20	1.72	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60
24	1.71	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57
30	1.70	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54
40	1.68	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51
60	1.67	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48
120	1.66	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45
∞	1.64	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42

*Reproducida de Charles W. Dunnett, "A Multiple Comparison Procedure for Comparing Several Treatments with a Control", *J. Am. Stat. Assoc.*, **50**, 1955, 1096-1121, con autorización del autor y del editor.

Tabla A.15 (continuación) Valores de $d_{\alpha/2}(k, v)$ para comparaciones unilaterales entre k tratamientos y un control

	k = número de medias de tratamiento (excluye el control)								
v	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5	3.37	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03
6	3.14	3.61	3.88	4.07	4.21	4.33	4.43	4.51	4.59
7	3.00	3.42	3.66	3.83	3.96	4.07	4.15	4.23	4.30
8	2.90	3.29	3.51	3.67	3.79	3.88	3.96	4.03	4.09
9	2.82	3.19	3.40	3.55	3.66	3.75	3.82	3.89	3.94
10	2.76	3.11	3.31	3.45	3.56	3.64	3.71	3.78	3.83
11	2.72	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74
12	2.68	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67
13	2.65	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61
14	2.62	2.94	3.11	3.23	3.32	3.40	3.46	3.51	3.56
15	2.60	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52
16	2.58	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48
17	2.57	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45
18	2.55	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42
19	2.54	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40
20	2.53	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38
24	2.49	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31
30	2.46	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.24
40	2.42	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18
60	2.39	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12
120	2.36	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06
∞	2.33	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00

Tabla A.16* Potencia de la prueba de análisis de varianza



*Reproducida de E. S. Pearson y H. O. Hartley, "Charts of the Power Function for Analysis-of-Variance Tests, Derived from the Non-central F-Distribution", *Biometrika*, **38**, 1951, 112-130, con autorización del editor.

Tabla A.16 (continuación) Potencia de la prueba de análisis de varianza

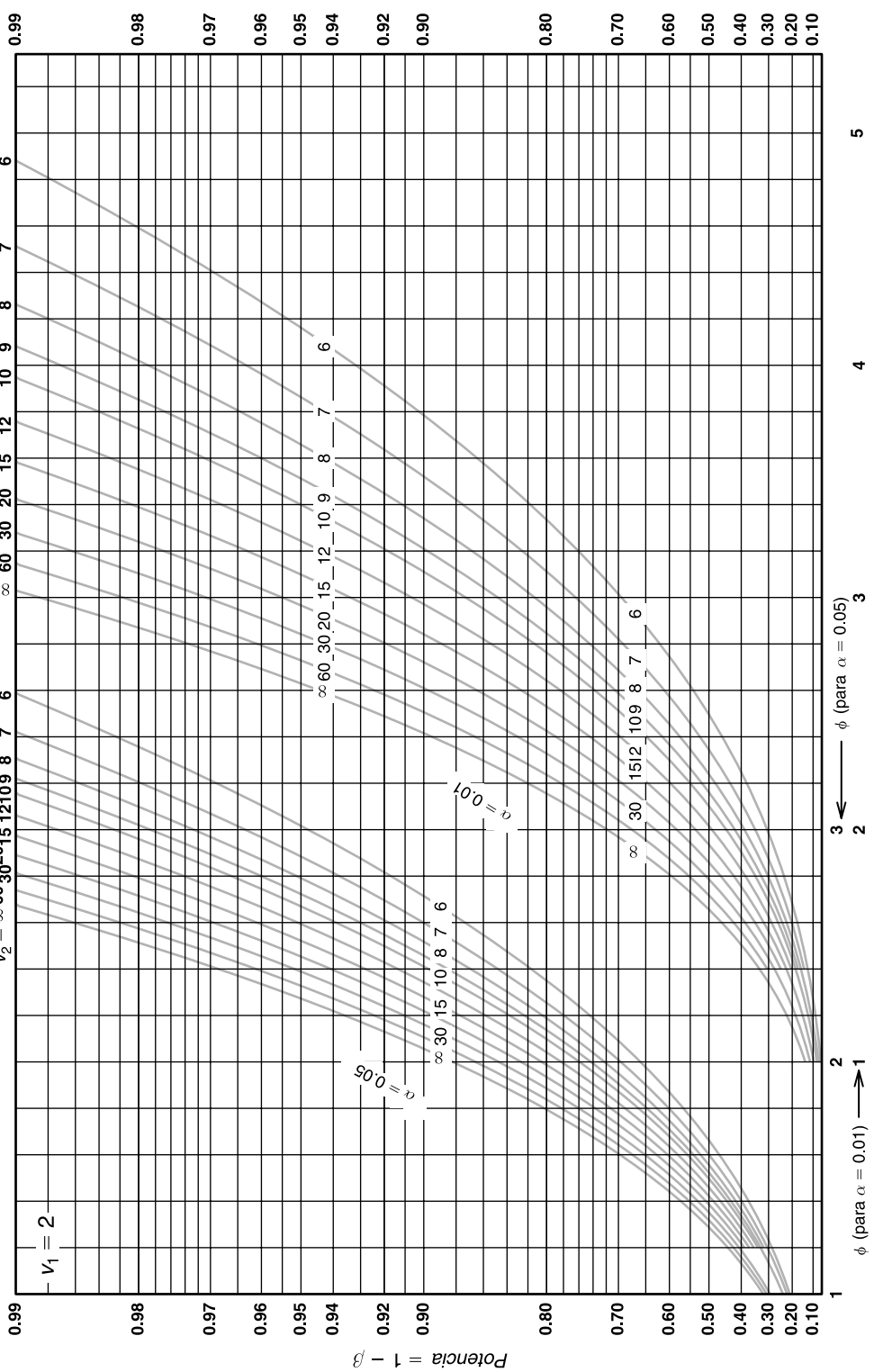


Tabla A.16 (continuación) Potencia de la prueba de análisis de varianza

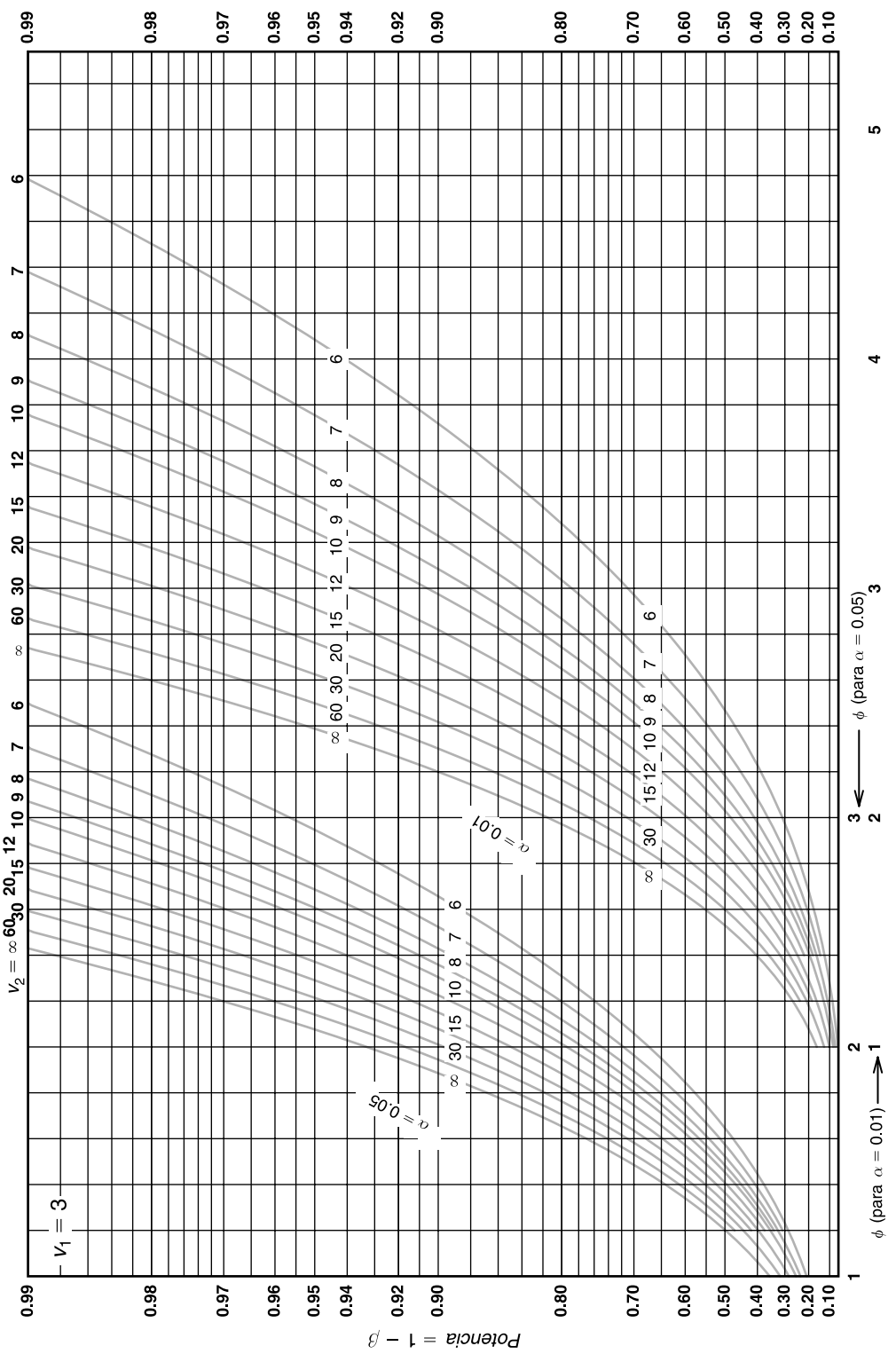


Tabla A.16 (continuación) Potencia de la prueba de análisis de varianza

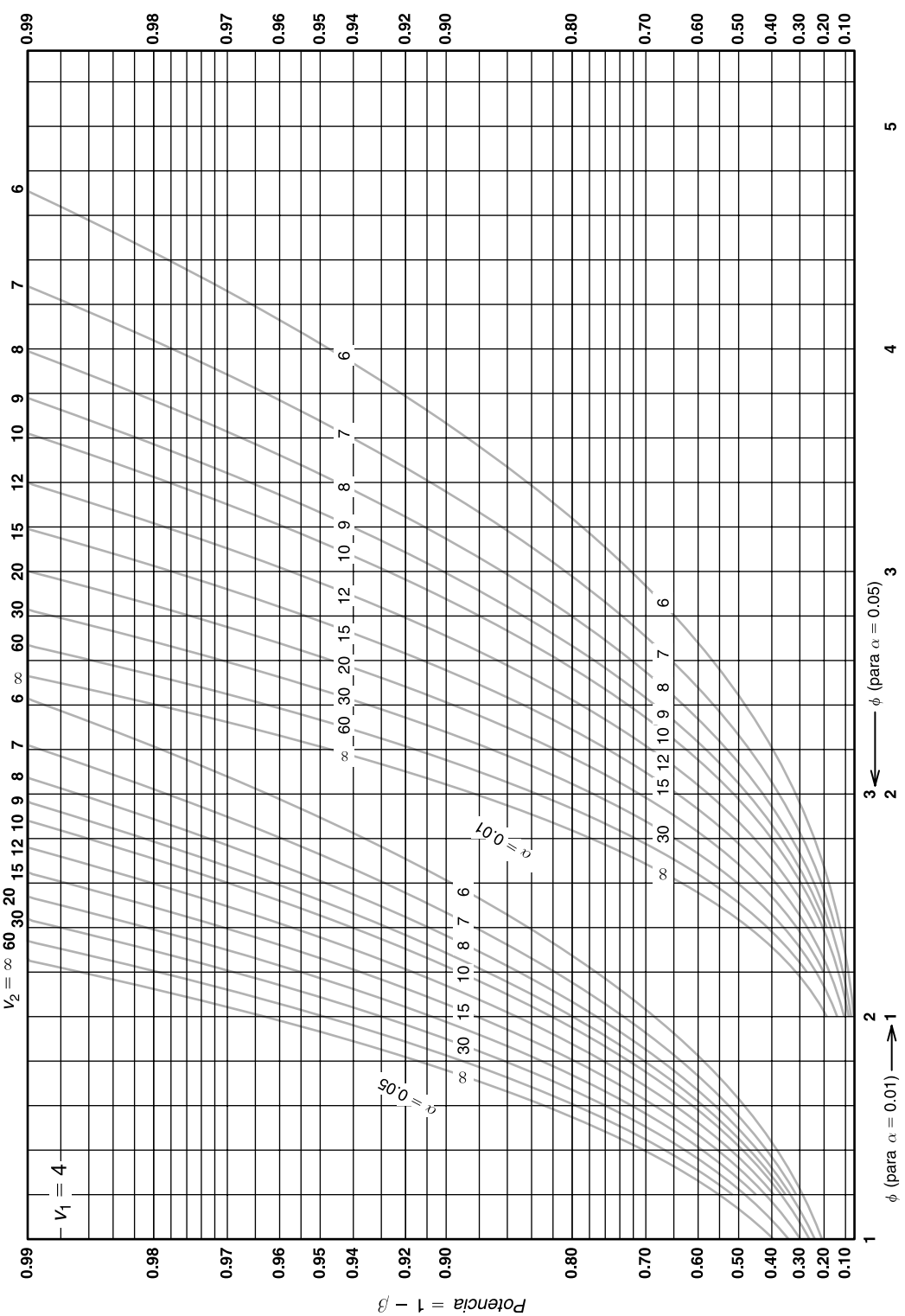


Tabla A.16 (continuación) Potencia de la prueba de análisis de varianza

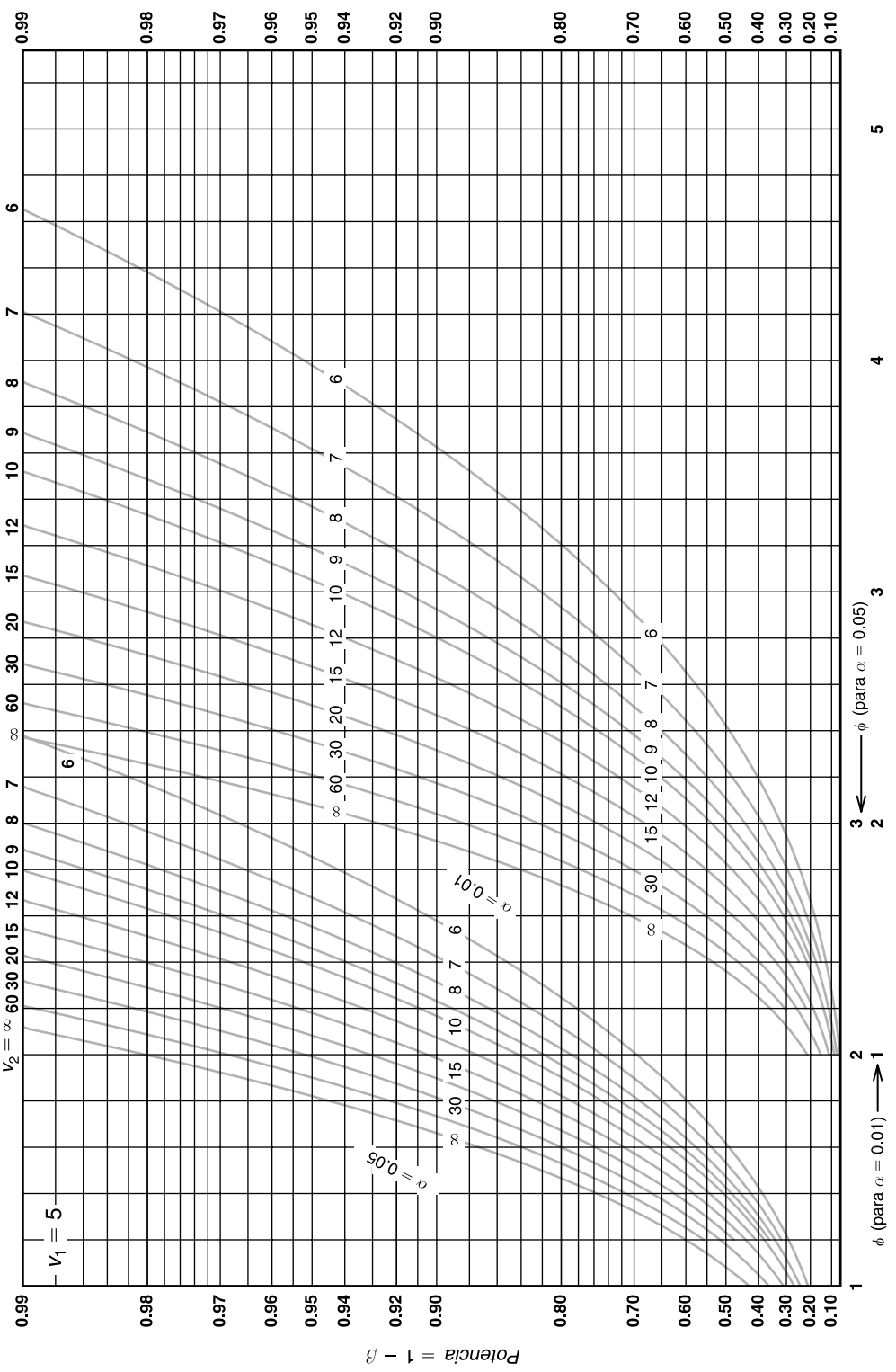


Tabla A.16 (continuación) Potencia de la prueba de análisis de varianza

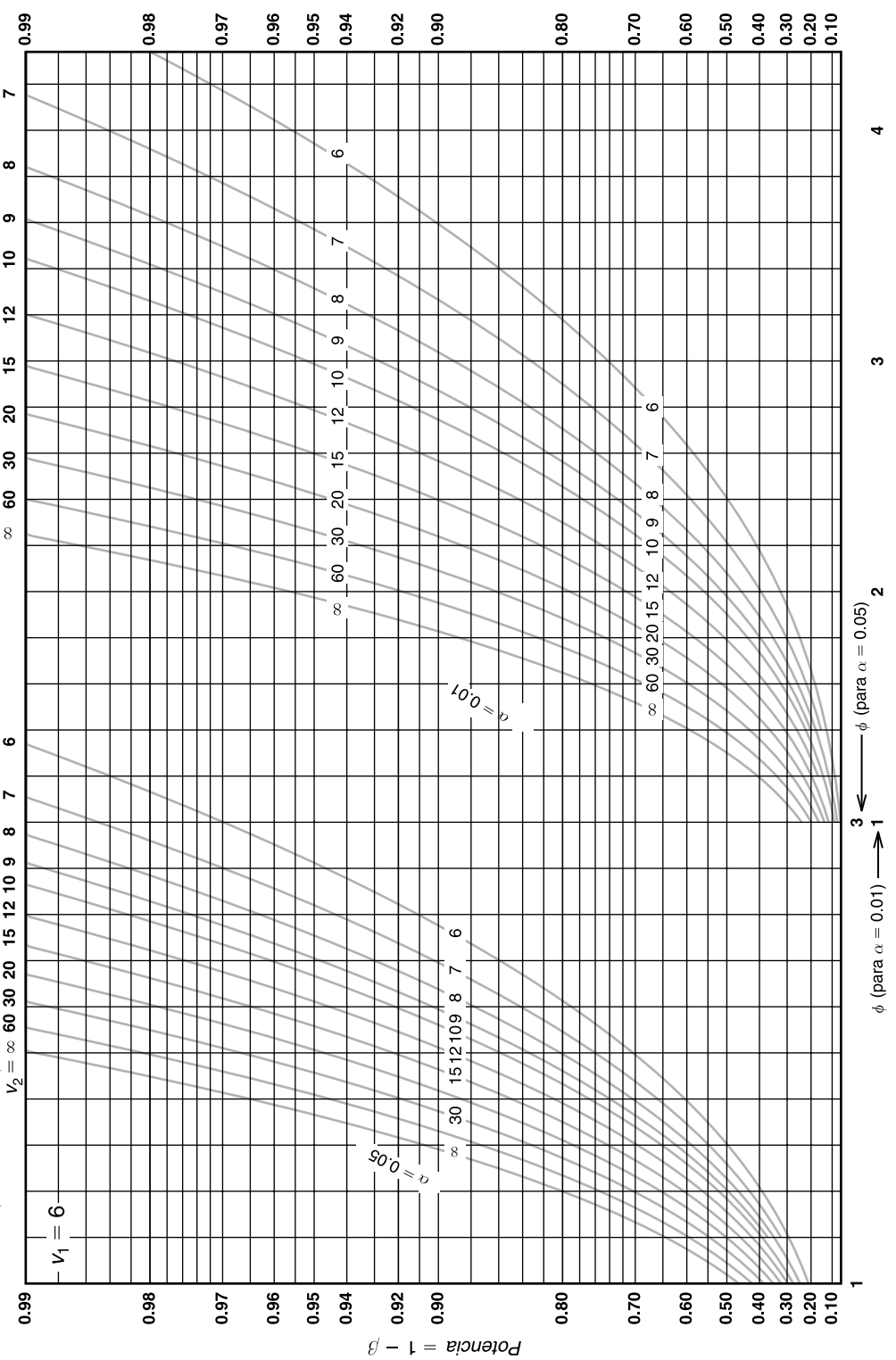


Tabla A.16 (continuación) Potencia de la prueba de análisis de varianza

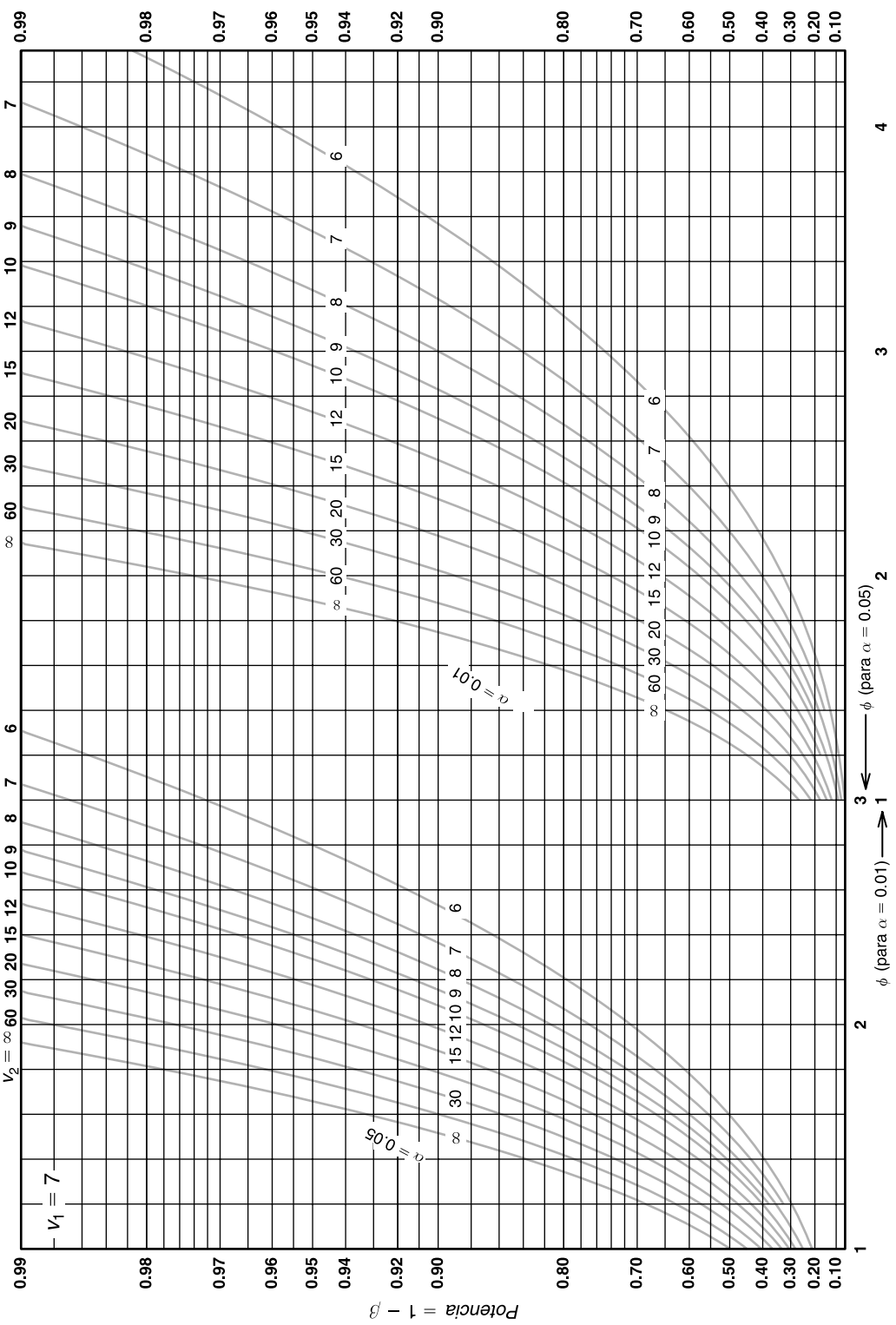


Tabla A.16 (continuación) Potencia de la prueba de análisis de varianza

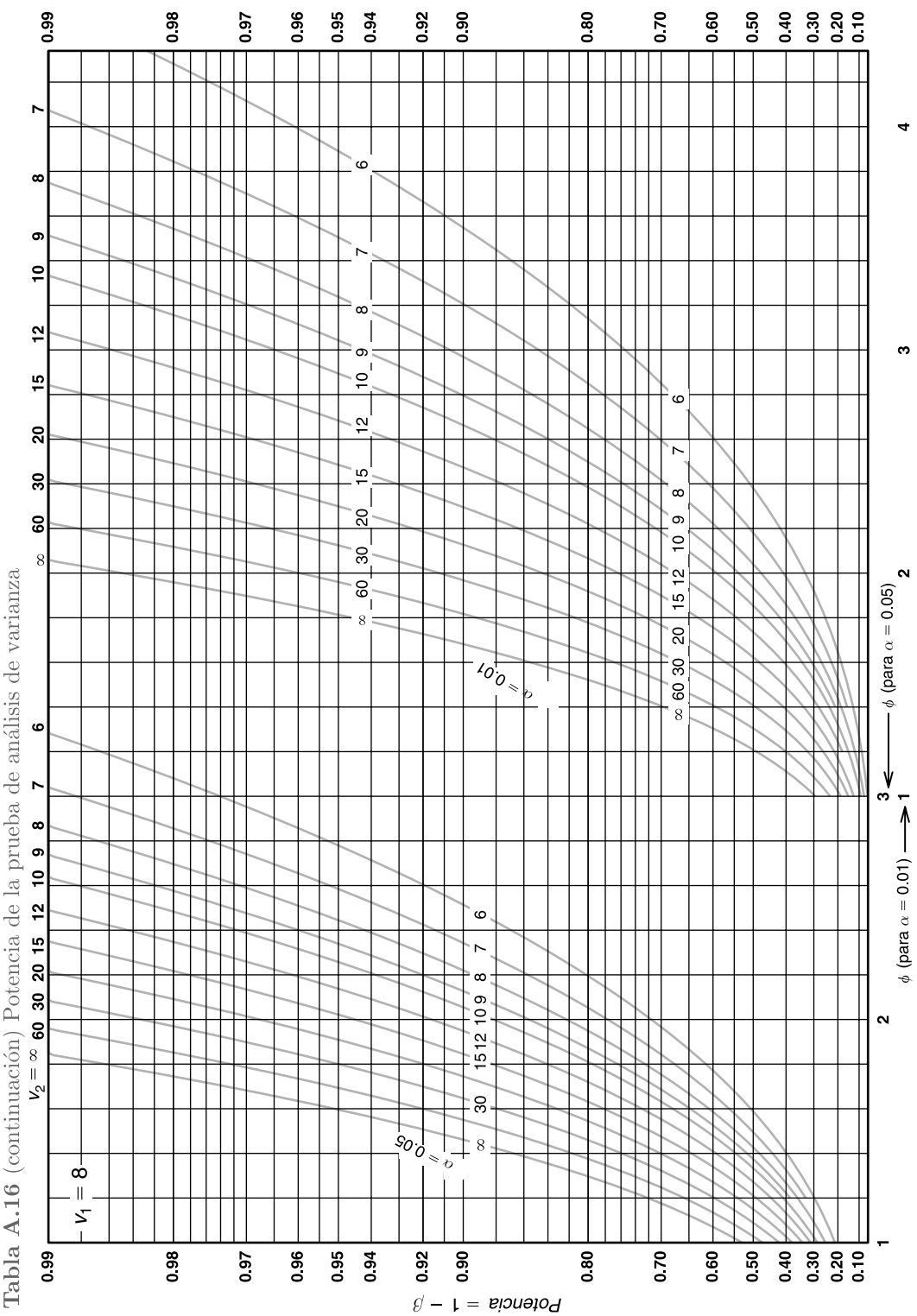


Tabla A.17* Valores críticos para la prueba de rangos con signo

n	Unilateral $\alpha = 0.01$ Bilateral $\alpha = 0.02$	Unilateral $\alpha = 0.025$ Bilateral $\alpha = 0.05$	Unilateral $\alpha = 0.05$ Bilateral $\alpha = 0.1$
5			1
6		1	2
7	0	2	4
8	2	4	6
9	3	6	8
10	5	8	11
11	7	11	14
12	10	14	17
13	13	17	21
14	16	21	26
15	20	25	30
16	24	30	36
17	28	35	41
18	33	40	47
19	38	46	54
20	43	52	60
21	49	59	68
22	56	66	75
23	62	73	83
24	69	81	92
25	77	90	101
26	85	98	110
27	93	107	120
28	102	117	130
29	111	127	141
30	120	137	152

*Reproducida de F. Wilcoxon y R. A. Wilcox, *Some Rapid Approximate Statistical Procedures, American Cyanamid Company*, Pearl River, N. Y., 1964, con autorización de la American Cyanamid Company.

Tabla A.18* Valores críticos para la prueba de suma de rangos de Wilcoxon

Prueba de una cola con $\alpha = 0.001$ o prueba de dos colas con $\alpha = 0.002$

	n_2														
n_1	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1															
2															
3												0	0	0	0
4					0	0	0	1	1	1	2	2	3	3	3
5		0	0	1	1	2	2	3	3	4	5	5	6	7	7
6	0	1	2	2	3	4	4	5	6	7	8	9	10	11	12
7		2	3	3	5	6	7	8	9	10	11	13	14	15	16
8			5	5	6	8	9	11	12	14	15	17	18	20	21
9				7	8	10	12	14	15	17	19	21	23	25	26
10					10	12	14	17	19	21	23	25	27	29	32
11						15	17	20	22	24	27	29	32	34	37
12							20	23	25	28	31	34	37	40	42
13								26	29	32	35	38	42	45	48
14									32	36	39	43	46	50	54
15										40	43	47	51	55	59
16											48	52	56	60	65
17												57	61	66	70
18													66	71	76
19														77	82
20															88

Prueba de una cola con $\alpha = 0.01$ o prueba de dos colas con $\alpha = 0.02$

	n_2														
n_1	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1																
2									0	0	0	0	0	0	1	1
3			0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6		3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7			6	8	9	11	12	14	16	17	19	21	23	24	26	28
8				10	11	13	15	17	20	22	24	26	28	30	32	34
9					14	16	18	21	23	26	28	31	33	36	38	40
10						19	22	24	27	30	33	36	38	41	44	47
11							25	28	31	34	37	41	44	47	50	53
12								31	35	38	42	46	49	53	56	60
13									39	43	47	51	55	59	63	67
14										47	51	56	60	65	69	73
15											56	61	66	70	75	80
16												66	71	76	82	87
17													77	82	88	93
18														88	94	100
19															101	107
20																114

*Basada en parte de las tablas 1, 3, 5 y 7 de D. Auble, "Extended Tables for the Mann-Whitney Statistic", *Bulletin of the Institute of Educational Research at Indiana University,* **1**, núm. 2, 1953, con autorización del director.

Tabla A.18 (continuación) Valores críticos para la prueba de suma de rangos de Wilcoxon

Prueba de una cola con $\alpha = 0.025$ o prueba de dos colas con $\alpha = 0.05$

	n_2																
n_1	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1																	
2					0	0	0	0	1	1	1	1	1	2	2	2	2
3		0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5		2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7				8	10	12	14	16	18	20	22	24	26	28	30	32	34
8					13	15	17	19	22	24	26	29	31	34	36	38	41
9						17	20	23	26	28	31	34	37	39	42	45	48
10							23	26	29	33	36	39	42	45	48	52	55
11								30	33	37	40	44	47	51	55	58	62
12									37	41	45	49	53	57	61	65	69
13										45	50	54	59	63	67	72	76
14											55	59	64	67	74	78	83
15												64	70	75	80	85	90
16													75	81	86	92	98
17														87	93	99	105
18															99	106	112
19																113	119
20																	127

Prueba de una cola con $\alpha = 0.05$ o prueba de dos colas con $\alpha = 0.1$

	n_2																	
n_1	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1																	0	0
2			0	0	0	1	1	1	1	2	2	3	3	3	3	4	4	4
3	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
4		1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5			4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6				7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7					11	13	15	17	19	21	24	26	28	30	33	35	37	39
8						15	18	20	23	26	28	31	33	36	39	41	44	47
9							21	24	27	30	33	36	39	42	45	48	51	54
10								27	31	34	37	41	44	48	51	55	58	62
11									34	38	42	46	50	54	57	61	65	69
12										42	47	51	55	60	64	68	72	77
13											51	56	61	65	70	75	80	84
14												61	66	71	77	82	87	92
15													72	77	83	88	94	100
16														83	89	95	101	107
17															96	102	109	115
18																109	116	123
19																	123	130
20																		138

Tabla A.19* $P(V \leq v^*$ cuando H_0 es verdadera) en la prueba de corridas

	v^*								
(n_1, n_2)	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
(2,3)	0.200	0.500	0.900	1.000					
(2,4)	0.133	0.400	0.800	1.000					
(2,5)	0.095	0.333	0.714	1.000					
(2,6)	0.071	0.286	0.643	1.000					
(2,7)	0.056	0.250	0.583	1.000					
(2,8)	0.044	0.222	0.533	1.000					
(2,9)	0.036	0.200	0.491	1.000					
(2,10)	0.030	0.182	0.455	1.000					
(3,3)	0.100	0.300	0.700	0.900	1.000				
(3,4)	0.057	0.200	0.543	0.800	0.971	1.000			
(3,5)	0.036	0.143	0.429	0.714	0.929	1.000			
(3,6)	0.024	0.107	0.345	0.643	0.881	1.000			
(3,7)	0.017	0.083	0.283	0.583	0.833	1.000			
(3,8)	0.012	0.067	0.236	0.533	0.788	1.000			
(3,9)	0.009	0.055	0.200	0.491	0.745	1.000			
(3,10)	0.007	0.045	0.171	0.455	0.706	1.000			
(4,4)	0.029	0.114	0.371	0.629	0.886	0.971	1.000		
(4,5)	0.016	0.071	0.262	0.500	0.786	0.929	0.992	1.000	
(4,6)	0.010	0.048	0.190	0.405	0.690	0.881	0.976	1.000	
(4,7)	0.006	0.033	0.142	0.333	0.606	0.833	0.954	1.000	
(4,8)	0.004	0.024	0.109	0.279	0.533	0.788	0.929	1.000	
(4,9)	0.003	0.018	0.085	0.236	0.471	0.745	0.902	1.000	
(4,10)	0.002	0.014	0.068	0.203	0.419	0.706	0.874	1.000	
(5,5)	0.008	0.040	0.167	0.357	0.643	0.833	0.960	0.992	1.000
(5,6)	0.004	0.024	0.110	0.262	0.522	0.738	0.911	0.976	0.998
(5,7)	0.003	0.015	0.076	0.197	0.424	0.652	0.854	0.955	0.992
(5,8)	0.002	0.010	0.054	0.152	0.347	0.576	0.793	0.929	0.984
(5,9)	0.001	0.007	0.039	0.119	0.287	0.510	0.734	0.902	0.972
(5,10)	0.001	0.005	0.029	0.095	0.239	0.455	0.678	0.874	0.958
(6,6)	0.002	0.013	0.067	0.175	0.392	0.608	0.825	0.933	0.987
(6,7)	0.001	0.008	0.043	0.121	0.296	0.500	0.733	0.879	0.966
(6,8)	0.001	0.005	0.028	0.086	0.226	0.413	0.646	0.821	0.937
(6,9)	0.000	0.003	0.019	0.063	0.175	0.343	0.566	0.762	0.902
(6,10)	0.000	0.002	0.013	0.047	0.137	0.288	0.497	0.706	0.864
(7,7)	0.001	0.004	0.025	0.078	0.209	0.383	0.617	0.791	0.922
(7,8)	0.000	0.002	0.015	0.051	0.149	0.296	0.514	0.704	0.867
(7,9)	0.000	0.001	0.010	0.035	0.108	0.231	0.427	0.622	0.806
(7,10)	0.000	0.001	0.006	0.024	0.080	0.182	0.355	0.549	0.743
(8,8)	0.000	0.001	0.009	0.032	0.100	0.214	0.405	0.595	0.786
(8,9)	0.000	0.001	0.005	0.020	0.069	0.157	0.319	0.500	0.702
(8,10)	0.000	0.000	0.003	0.013	0.048	0.117	0.251	0.419	0.621
(9,9)	0.000	0.000	0.003	0.012	0.044	0.109	0.238	0.399	0.601
(9,10)	0.000	0.000	0.002	0.008	0.029	0.077	0.179	0.319	0.510
(10,10)	0.000	0.000	0.001	0.004	0.019	0.051	0.128	0.242	0.414

*Reproducida de C. Eisenhart y F. Swed, "Tables for Testing Randomness of Grouping in a Sequence of Alternatives", *Ann. Math. Stat.*, **14**, 1943, con autorización del editor.

Tabla A.19 (continuación) $P(V \leq v^*$ cuando H_0 es verdadera) en la prueba de corridas

	v^*									
(n_1, n_2)	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
(2,3)										
(2,4)										
(2,5)										
(2,6)										
(2,7)										
(2,8)										
(2,9)										
(2,10)										
(3,3)										
(3,4)										
(3,5)										
(3,6)										
(3,7)										
(3,8)										
(3,9)										
(3,10)										
(4,4)										
(4,5)										
(4,6)										
(4,7)										
(4,8)										
(4,9)										
(4,10)										
(5,5)										
(5,6)	1.000									
(5,7)	1.000									
(5,8)	1.000									
(5,9)	1.000									
(5,10)	1.000									
(6,6)	0.998	1.000								
(6,7)	0.992	0.999	1.000							
(6,8)	0.984	0.998	1.000							
(6,9)	0.972	0.994	1.000							
(6,10)	0.958	0.990	1.000							
(7,7)	0.975	0.996	0.999	1.000						
(7,8)	0.949	0.988	0.998	1.000	1.000					
(7,9)	0.916	0.975	0.994	0.999	1.000					
(7,10)	0.879	0.957	0.990	0.998	1.000					
(8,8)	0.900	0.968	0.991	0.999	1.000	1.000				
(8,9)	0.843	0.939	0.980	0.996	0.999	1.000	1.000			
(8,10)	0.782	0.903	0.964	0.990	0.998	1.000	1.000			
(9,9)	0.762	0.891	0.956	0.988	0.997	1.000	1.000	1.000		
(9,10)	0.681	0.834	0.923	0.974	0.992	0.999	1.000	1.000	1.000	
(10,10)	0.586	0.758	0.872	0.949	0.981	0.996	0.999	1.000	1.000	1.000

Tabla A.20* Tamaño muestral para límites de tolerancia no paramétricos bilaterales

	$1-\gamma$					
$1-\alpha$	**0.50**	**0.70**	**0.90**	**0.95**	**0.99**	**0.995**
0.995	336	488	777	947	1,325	1,483
0.99	168	244	388	473	662	740
0.95	34	49	77	93	130	146
0.90	17	24	38	46	64	72
0.85	11	16	25	30	42	47
0.80	9	12	18	22	31	34
0.75	7	10	15	18	24	27
0.70	6	8	12	14	20	22
0.60	4	6	9	10	14	16
0.50	3	5	7	8	11	12

*Reproducida de la tabla A-25d de Wilfrid J. Dixon y Frank J. Massey, Jr., *Introduction to Statistical Analysis*, 3a. ed., McGraw-Hill, Nueva York, 1969. Utilizada con autorización de McGraw-Hill Book Company.

Tabla A.21* Tamaño muestral para límites de tolerancia no paramétricos unilaterales

	$1-\gamma$				
$1-\alpha$	**0.50**	**0.70**	**0.95**	**0.99**	**0.995**
0.995	139	241	598	919	1,379
0.99	69	120	299	459	688
0.95	14	24	59	90	135
0.90	7	12	29	44	66
0.85	5	8	19	29	43
0.80	4	6	14	21	31
0.75	3	5	11	7	25
0.70	2	4	9	13	20
0.60	2	3	6	10	14
0.50	1	2	5	7	10

*Reproducida de la tabla A-25e de Wilfrid J. Dixon y Frank J. Massey, Jr., *Introduction to Statistical Analysis*, 3a. ed., McGraw-Hill, Nueva York, 1969. Utilizada con autorización de McGraw-Hill Book Company.

Tabla A.22* Valores críticos del coeficiente de correlación de rangos de Spearman

n	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
5	0.900			
6	0.829	0.886	0.943	
7	0.714	0.786	0.893	
8	0.643	0.738	0.833	0.881
9	0.600	0.683	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.523	0.623	0.736	0.818
12	0.497	0.591	0.703	0.780
13	0.475	0.566	0.673	0.745
14	0.457	0.545	0.646	0.716
15	0.441	0.525	0.623	0.689
16	0.425	0.507	0.601	0.666
17	0.412	0.490	0.582	0.645
18	0.399	0.476	0.564	0.625
19	0.388	0.462	0.549	0.608
20	0.377	0.450	0.534	0.591
21	0.368	0.438	0.521	0.576
22	0.359	0.428	0.508	0.562
23	0.351	0.418	0.496	0.549
24	0.343	0.409	0.485	0.537
25	0.336	0.400	0.475	0.526
26	0.329	0.392	0.465	0.515
27	0.323	0.385	0.456	0.505
28	0.317	0.377	0.448	0.496
29	0.311	0.370	0.440	0.487
30	0.305	0.364	0.432	0.478

*Reproducida de E. G. Olds, "Distribution of Sums of Squares of Rank Differences for Small Samples", *Ann. Math. Stat.*, **9**, 1938, con autorización del editor.

Tabla A.23* Factores para la elaboración de gráficas de control

	Gráfica para promedios		Gráfica para desviaciones estándar						Gráfica para rangos				
	Factores para los límites de control		Factores para la línea central			Factores para los límites de control			Factores para la línea central		Factores para los límites de control		
Obs. en la muestra, n	A_2	A_3	c_4	$1/c_4$	B_3	B_4	B_5	B_6	d_2	$1/d_2$	d_3	D_3	D_4
2	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.267
3	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	2.574
4	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	2.282
5	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	2.114
6	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	2.004
7	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.076	1.924
8	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.136	1.864
9	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.184	1.816
10	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.223	1.777
11	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.256	1.744
12	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.283	1.717
13	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	0.307	1.693
14	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	0.328	1.672
15	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	0.347	1.653
16	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	0.363	1.637
17	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	0.378	1.622
18	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	0.391	1.608
19	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	0.403	1.597
20	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	0.415	1.585
21	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	0.425	1.575
22	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	0.434	1.566
23	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	0.443	1.557
24	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	0.451	1.548
25	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	0.459	4.541

Tabla A.24 La función gamma incompleta: $\boldsymbol{F(x;\alpha)} = \int_0^x \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, \boldsymbol{dy}$

	$\boldsymbol{\alpha}$									
$\boldsymbol{x}$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0.6320	0.2640	0.0800	0.0190	0.0040	0.0010	0.0000	0.0000	0.0000	0.0000
2	0.8650	0.5940	0.3230	0.1430	0.0530	0.0170	0.0050	0.0010	0.0000	0.0000
3	0.9500	0.8010	0.5770	0.3530	0.1850	0.0840	0.0340	0.0120	0.0040	0.0010
4	0.9820	0.9080	0.7620	0.5670	0.3710	0.2150	0.1110	0.0510	0.0210	0.0080
5	0.9930	0.9600	0.8750	0.7350	0.5600	0.3840	0.2380	0.1330	0.0680	0.0320
6	0.9980	0.9830	0.9380	0.8490	0.7150	0.5540	0.3940	0.2560	0.1530	0.0840
7	0.9990	0.9930	0.9700	0.9180	0.8270	0.6990	0.5500	0.4010	0.2710	0.1700
8	1.0000	0.9970	0.9860	0.9580	0.9000	0.8090	0.6870	0.5470	0.4070	0.2830
9		0.9990	0.9940	0.9790	0.9450	0.8840	0.7930	0.6760	0.5440	0.4130
10		1.0000	0.9970	0.9900	0.9710	0.9330	0.8700	0.7800	0.6670	0.5420
11			0.9990	0.9950	0.9850	0.9620	0.9210	0.8570	0.7680	0.6590
12			1.0000	0.9980	0.9920	0.9800	0.9540	0.9110	0.8450	0.7580
13				0.9990	0.9960	0.9890	0.9740	0.9460	0.9000	0.8340
14				1.0000	0.9980	0.9940	0.9860	0.9680	0.9380	0.8910
15					0.9990	0.9970	0.9920	0.9820	0.9630	0.9300

A.25 Prueba de la media de la distribución hipergeométrica

Para calcular la media de la distribución hipergeométrica, escribimos

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = k \sum_{x=1}^{n} \frac{(k-1)!}{(x-1)!(k-x)!} \cdot \frac{\binom{N-k}{n-x}}{\binom{N}{n}} \\ &= k \sum_{x=1}^{n} \frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}}. \end{aligned}$$

Puesto que

$$\binom{N-k}{n-1-y} = \binom{(N-1)-(k-1)}{n-1-y} \quad \text{y} \quad \binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N}{n}\binom{N-1}{n-1},$$

y con $y = x - 1$, obtenemos

$$\begin{aligned} E(X) &= k \sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{N-k}{n-1-y}}{\binom{N}{n}} \\ &= \frac{nk}{N} \sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{(N-1)-(k-1)}{n-1-y}}{\binom{N-1}{n-1}} = \frac{nk}{N}, \end{aligned}$$

ya que la sumatoria representa el total de todas las probabilidades en un experimento hipergeométrico cuando $N - 1$ artículos se seleccionan al azar de $N - 1$, de los cuales $k - 1$ son éxitos marcados.

A.26 Prueba de la media y la varianza de la distribución de Poisson

Sea $\mu = \lambda t$.

$$E(X) = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\mu}\mu^x}{x!} = \sum_{x=1}^{\infty} x \cdot \frac{e^{-\mu}\mu^x}{x!} = \mu \sum_{x=1}^{\infty} \frac{e^{-\mu}\mu^{x-1}}{(x-1)!}.$$

Puesto que la sumatoria en el último término de la expresión anterior es la probabilidad total de una variable aleatoria de Poisson con media μ, la cual puede verse con facilidad con $y = x - 1$, y es igual a 1. Por lo tanto, $E(X) = \mu$. Para calcular la varianza de X, note que

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\mu}\mu^x}{x!} = \mu^2 \sum_{x=2}^{\infty} \frac{e^{-\mu}\mu^{x-2}}{(x_2)!}.$$

De nuevo, sea $y = x - 2$, la sumatoria en el último término de la expresión anterior es la probabilidad total de una variable aleatoria de Poisson con media μ. Entonces, obtenemos

$$\sigma^2 = E(X^2) - [E(X)]^2 = E[X(X-1)] + E(X) - [E(X)]^2 = \mu^2 + \mu - \mu^2 = \mu = \lambda t.$$

A.27 Prueba de que la distribución de Poisson es una limitante de la binomial

La distribución binomial se puede escribir como

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = \frac{n(n-1)\cdots(n-x+1)}{x!} p^x (1-p)^{n-x}.$$

Al sustituir $p = \mu/n$,

$$\begin{aligned} b(x; n, p) &= \frac{n(n-1)\cdots(n-x+1)}{x!} \left(\frac{\mu}{n}\right) \left(1 - \frac{\mu}{n}\right)^{n-x} \\ &= 1 \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{x-1}{n}\right) \frac{\mu^x}{x!} \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-x}. \end{aligned}$$

Conforme $n \to \infty$, siempre que x y y permanezcan constantes,

$$\lim_{n\to\infty} 1 \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{x-1}{n}\right) = 1, \qquad \lim_{n\to\infty} \left(1 - \frac{\mu}{n}\right)^{-x} = 1,$$

y a partir de la definición de e,

$$\lim_{n\to\infty} \left(1 - \frac{\mu}{n}\right)^n = \lim_{n\to\infty} \left\{ \left[1 + \frac{1}{(-n)/\mu}\right]^{-n/\mu} \right\}^{-\mu} = e^{-\mu}.$$

Por lo tanto, bajo las condiciones limitantes dadas,

$$b(x; n, p) \to \frac{e^{-\mu}\mu^x}{x!}, \qquad x = 0, 1, 2, \ldots.$$

A.28 Prueba de la media y la varianza de la distribución gamma

Para calcular la media y la varianza de la distribución gamma, calculamos primero

$$E(X^k) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha+k-1} e^{-x/\beta}\, dx = \frac{\beta^{k+\alpha}\Gamma(\alpha+k)}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \frac{x^{\alpha+k-1} e^{-x/\beta}}{\beta^{k+\alpha}\Gamma(\alpha+k)}\, dx,$$

para $k = 0, 1, 2, \ldots$. Puesto que el integrando en el último término de la expresión anterior es una función de densidad gamma, con parámetros $\alpha + k$ y β, y es igual a 1. Entonces,

$$E(X^k) = \beta^k \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)}.$$

Usando la fórmula de recurrencia de la función gamma de la página 194, obtenemos

$$\mu = \beta\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha\beta \qquad \text{y} \qquad \sigma^2 = E(X^2) - \mu^2 = \beta^2\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} - \mu^2 = \beta^2\alpha(\alpha+1) - (\alpha\beta)^2 = \alpha\beta^2.$$

Apéndice B

Respuesta a los ejercicios de repaso impares

Capítulo 1

1.1 *a*) Tamaño de la muestra = 15
b) Media de la muestra = 3.787
c) Mediana de la muestra = 3.6
e) $\bar{x}_{\text{tr}(20)} = 3.678$

1.3 *b*) Sí, el proceso de envejecimiento redujo la resistencia a la tensión.
c) $\bar{x}_{\text{envejecimiento}} = 209.90$, $\bar{x}_{\text{No envejecimiento}} = 222.10$.
d) $\bar{x}_{\text{envejecimiento}} = 210.00$, $\bar{x}_{\text{No envejecimiento}} = 221.50$. Las medias y las medianas son similares en cada grupo.

1.5 *b*) Control: $\bar{x} = 5.60$, $\tilde{x} = 5.00$, $\bar{x}_{\text{tr}(10)} = 5.13$.
Tratamiento: $\bar{x} = 7.60$, $\tilde{x} = 4.50$, $\bar{x}_{\text{tr}(10)} = 5.63$.
c) El valor extremo de 37 en el grupo de tratamiento desempeña un papel significativo en el cálculo de la media.

1.7 Varianza de la muestra = 0.943
Desviación estándar de la muestra = 0.971

1.9 Sin envejecimiento: varianza de la muestra = 23.62, desviación estándar de la muestra = 4.86.
Con envejecimiento: varianza de la muestra = 42.12, desviación estándar de la muestra = 6.49.

1.11 Control: varianza de la muestra = 69.38, desviación estándar de la muestra = 8.33.
Tratamiento: varianza de la muestra = 128.04, desviación estándar de la muestra = 11.32.

1.13 *a*) Media = 124.3, mediana = 120;
b) 175 es un valor extremo.

1.15 Sí, el valor $P = 0.03125$; probabilidad de obtener $H\,H\,H\,H\,H$ con una moneda legal.

1.17 No fumadores *a*) 30.32, *b*) 7.13;
Fumadores *a*) 43.70, *b*) 16.93.
d) Parece que a los fumadores les toma más tiempo quedarse dormidos. Para los fumadores el tiempo para quedarse dormido es más variable.

1.19 *a*)

Tallo	Hojas	Frecuencia
0	22233457	8
1	023558	6
2	035	3
3	03	2
4	057	3
5	0569	4
6	0005	4

b)

Intervalo de clase	Punto medio de la clase	Frec.	Frec. rel.
0.0−0.9	0.45	8	0.267
1.0−1.9	1.45	6	0.200
2.0−2.9	2.45	3	0.100
3.0−3.9	3.45	2	0.067
4.0−4.9	4.45	3	0.100
5.0−5.9	5.45	4	0.133
6.0−6.9	6.45	4	0.133

c) Media de la muestra = 2.7967
Rango de la muestra = 6.3
Desviación estándar de la muestra = 2.2273

1.21 *a*) Media de la muestra = 1.7742
Mediana de la muestra = 1.77
b) Desviación estándar de la muestra = 0.3905.

1.23 *b*) $\bar{x}_{1990} = 160.15, \quad \bar{x}_{1980} = 395.10.$

c) Las emisiones medias cayeron entre 1980 y 1990, la variabilidad también disminuyó porque no hubo emisiones extremadamente más grandes.

1.25 *a*) Media de la muestra = 33.31

b) Media de la muestra = 26.35

d) $\bar{x}_{\text{tr}(10)} = 30.97$

Capítulo 2

2.1 *a*) $S = \{8, 16, 24, 32, 40,48\}$

b) $S = \{-5, 1\}$

c) $S = \{T, HT, HHT, HHH\}$

d) $S = \{$África, Antártica, Asia, Australia, Europa, Norteamérica, Sudamérica$\}$

e) $S = \phi$

2.3 $A = C$

2.5 $S = \{1HH, 1HT, 1TH, 1TT, 2H, 2T, 3HH, 3HT, 3TH, 3TT, 4H, 4T, 5HH, 5HT, 5TH, 5TT, 6H, 6T\}$

2.7 $S_1 = \{MMMM, MMMF, MMFM, MFMM, FMMM, MMFF, MFMF, MFFM, FMFM, FFMM, FMMF, MFFF, FMFF, FFMF, FFFM, FFFF\}$;
$S_2 = \{0, 1, 2, 3, 4\}$

2.9 *a*) $A = \{1HH, 1HT, 1TH, 1TT, 2H, 2T\}$

b) $B = \{1TT, 3TT, 5TT\}$

c) $A' = \{3HH, 3HT, 3TH, 3TT, 4H, 4T, 5HH, 5HT, 5TH, 5TT, 6H, 6T\}$

d) $A' \cap B = \{3TT, 5TT\}$

e) $A \cup B = \{1HH, 1HT, 1TH, 1TT, 2H, 2T, 3TT, 5TT\}$

2.11 *a*) $S = \{M_1M_2, M_1F_1, M_1F_2, M_2M_1, M_2F_1, M_2F_2, F_1M_1, F_1M_2, F_1F_2, F_2M_1, F_2M_2, F_2F_1\}$

b) $A = \{M_1M_2, M_1F_1, M_1F_2, M_2M_1, M_2F_1, M_2F_2\}$

c) $B = \{M_1F_1, M_1F_2, M_2F_1, M_2F_2, F_1M_1, F_1M_2, F_2M_1, F_2M_2\}$

d) $C = \{F_1F_2, F_2F_1\}$

e) $A \cap B = \{M_1F_1, M_1F_2, M_2F_1, M_2F_2\}$

f) $A \cup C = \{M_1M_2, M_1F_1, M_1F_2, M_2M_1, M_2F_1, M_2F_2, F_1F_2, F_2F_1\}$

2.15 *a*) {nitrógeno, potasio, uranio, oxígeno}

b) {cobre, sodio, zinc, oxígeno}

c) {cobre, sodio, nitrógeno, potasio, uranio, zinc}

d) {cobre, uranio, zinc}

e) ϕ

f) {oxígeno}

2.19 *a*) La familia experimentará fallas mecánicas, pero no recibirá una boleta de infracción por cometer una falta de tránsito, y no llegará a un lugar para acampar que esté lleno.

b) La familia recibirá una boleta por cometer una falta de tránsito, y llegará a un lugar para acampar que esté lleno, pero no experimentará fallas mecánicas.

c) La familia experimentará fallas mecánicas y llegará a un lugar para acampar que esté lleno.

d) La familia recibirá una boleta por cometer una falta de tránsito, pero no llegará a un lugar para acampar que esté lleno.

e) La familia no experimentará fallas mecánicas.

2.21 18

2.23 156

2.25 20

2.27 48

2.29 210

2.31 *a*) 1024; *b*) 243

2.33 72

2.35 362,880

2.37 2,880

2.39 *a*) 40,320; *b*) 336

2.41 360

2.43 24

2.45 3,360

2.47 7,920

2.49 56

2.51 *a*) La suma de las probabilidades excede 1.
b) La suma de las probabilidades es menor que 1.
c) Una probabilidad negativa.
d) La probabilidad de tanto un corazón como una carta negra es cero.

2.53 $S = \{\$10, \$25, \$100\}$; $P(10) = \frac{11}{20}$; $P(25) = \frac{3}{10}$, $P(100) = \frac{15}{100}$; $\frac{17}{20}$

2.55 *a*) 0.3; *b*) 0.2

2.57 *a*) 5/26; *b*) 9/26; c) 19/26

2.59 10/117

2.61 95/663

2.63 *a*) 94/54,145; *b*) 143/39,984

2.65 *a*) 22/25; *b*) 3/25; *c*) 17/50

2.67 *a*) 0.32; *b*) 0.68; *c*) oficina o estudio

2.69 *a*) 0.8; *b*) 0.45; *c*) 0.55

2.71 *a*) 0.31; *b*) 0.93; *c*) 0.31

2.73 *a*) 0.009; *b*) 0.999; *c*) 0.01

2.75 *a*) 0.048; *b*) \$50,000; *c*) \$12,500

2.77 *a*) La probabilidad de que un convicto promoviera el consumo de drogas y también cometiera robo a mano armada.
b) La probabilidad de que un convicto cometiera robo a mano armada y no promoviera el consumo de drogas.
c) La probabilidad de que un convicto no promoviera el consumo de drogas ni tampoco cometiera robo a mano armada.

2.79 *a*) 14/39; *b*) 95/112

2.81 *a*) 5/34; *b*) 3/8

2.83 *a*) 0.018; *b*) 0.614; *c*) 0.166; *d*) 0.479

2.85 *a*) 0.35; *b*) 0.875; *c*) 0.55

2.87 *a*) 9/28; *b*) 3/4; *c*) 0.91

2.89 0.27

2.91 5/8

2.93 *a*) 0.0016; *b*) 0.9984.

2.95 *a*) 1/5; *b*) 4/15; *c*) 3/5.

2.97 *a*) 91/323; *b*) 91/323.

2.93 *a*) 0.75112; *b*) 0.2045.

2.101 0.0960

2.103 0.40625

2.105 0.1124

2.107 *a*) 0.045; *b*) 0.564; *c*) 0.630; *d*) 0.1064

Capítulo 3

3.1 Discreta; continua; continua; discreta; discreta; continua.

3.3

Espacio muestral	w
HHH	3
HHT	1
HTH	1
THH	1
HTT	−1
THT	−1
TTH	−1
TTT	−3

3.5 *a*) 1/30; *b*) 1/10

3.7 *a*) 0.68; *b*) 0.375

3.9 *b*) 19/80

3.11

x	**0**	**1**	**2**
$f(x)$	$\frac{2}{7}$	$\frac{4}{7}$	$\frac{1}{7}$

3.13
$$F(x) = \begin{cases} 0, & \text{para } x < 0, \\ 0.41, & \text{para } 0 \le x < 1, \\ 0.78, & \text{para } 1 \le x < 2, \\ 0.94, & \text{para } 2 \le x < 3, \\ 0.99, & \text{para } 3 \le x < 4, \\ 1, & \text{para } x \ge 4 \end{cases}$$

3.15
$$F(x) = \begin{cases} 0, & \text{para } x < 0, \\ \frac{2}{7}, & \text{para } 0 \le x < 1, \\ \frac{6}{7}, & \text{para } 1 \le x < 2, \\ 1, & \text{para } x \ge 2 \end{cases}$$
a) 4/7; *b*) 5/7

3.17 *b*) 1/4; *c*) 0.3

3.19 $F(x) = (x-1)/2$, para $0 \le x < 3$; 1/4

3.21 *a*) 3/2; *b*) $F(x) = x^{3/2}$, para $0 \le x < 1$; 0.3004

3.23
$$F(w)=\begin{cases}0, & \text{para } w<-3,\\ \frac{1}{27}, & \text{para } -3\le w<-1,\\ \frac{7}{27}, & \text{para } -1\le w<1,\\ \frac{19}{27}, & \text{para } 1\le w<3,\\ 1, & \text{para } w\ge 3\end{cases}$$
a) 20/27; *b*) 2/3

3.25

t	20	25	30
$P(T=t)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$

3.27 *a*)
$$F(x)=\begin{cases}0, & x<0,\\ 1-\exp(-x/2000), & x\ge 0.\end{cases}$$
b) 0.6065; *c*) 0.6321

3.29 *b*)
$$F(x)=\begin{cases}0, & x<1,\\ 1-x^{-3}, & x\ge 1.\end{cases}$$
c) 0.0156

3.31 *a*) 0.2231; *b*) 0.2212

3.33 *a*) $k=280$; *b*) 0.3633; *c*) 0.0563

3.35 *a*) 0.1528; *b*) 0.0446

3.37 *a*) 1/36; *b*) 1/15

3.39 *a*)

		x			
	$f(x,y)$	0	1	2	3
y	0	0	$\frac{3}{70}$	$\frac{9}{70}$	$\frac{3}{70}$
	1	$\frac{2}{70}$	$\frac{18}{70}$	$\frac{18}{70}$	$\frac{2}{70}$
	2	$\frac{3}{70}$	$\frac{9}{70}$	$\frac{3}{70}$	

b) 1/2

3.41 *a*) 1/16; *b*) $g(x)=12x(1-x)^2$, para $0\le x\le 1$; *c*) 1/4

3.43 *a*) 3/64; *b*) 1/2

3.45 0.6534

3.47 *a*) Dependiente; *b*) 1/3

3.49 *a*)

x	1	2	3
$g(x)$	0.10	0.35	0.55

b)

y	1	2	3
$h(y)$	0.20	0.50	0.30

c) 0.5714

3.51 *a*)

		x		
	$f(x,y)$	0	1	2
y	0	$\frac{16}{36}$	$\frac{8}{36}$	$\frac{1}{36}$
	1	$\frac{8}{36}$	$\frac{2}{36}$	0
	2	$\frac{1}{36}$	0	0

b) 11/12

3.53 *a*)

		x			
	$f(x,y)$	0	1	2	3
y	0	$\frac{1}{55}$	$\frac{6}{55}$	$\frac{6}{55}$	$\frac{1}{55}$
	1	$\frac{6}{55}$	$\frac{16}{55}$	$\frac{6}{55}$	0
	2	$\frac{6}{55}$	$\frac{6}{55}$	0	0
	3	$\frac{1}{55}$	0	0	0

b) 42/55

3.55 5/8

3.57 Independiente

3.59 *a*) 3; *b*) 21/512

3.61 Dependiente

Capítulo 4

4.1 0

4.3 25 centavos

4.5 0.88

4.7 $500

4.9 $1.23

4.11 $6,900

4.13 $(\ln 4)/\pi$

4.15 100 horas

4.17 209

4.19 $1,855

4.21 $833.33

4.23 *a*) 35.2; *b*) $\mu_X=3.20$, $\mu_Y=3.00$

4.25 2

4.27 2,000 horas

4.29 *b*) 3/2

4.31 *a*) 1/6; *b*) $(5/6)^5$

4.33 $5,250,000

4.35 0.74

4.37 1/18; en ganancia real la varianza es $\frac{1}{18}(5000)^2$

4.39 1/6

4.41 118.9

4.43 $\mu_Y = 10$; $\sigma_Y^2 = 144$.

4.45 $\sigma_{XY} = 0.005$

4.47 -0.0062

4.49 $\sigma_X^2 = 0.8456$; $\sigma_X = 0.9196$

4.51 10.33; 6.66

4.53 80 centavos

4.55 209

4.57 $\mu = 7/2$; $\sigma^2 = 15/4$

4.59 3/14

4.61 0.03125

4.63 0.9340

4.65 52

4.67 *a*) Cuando más 4/9; *b*) al menos 5/9; *c*) al menos 21/25; *d*) 10.

4.69 *a*) 7; *b*) 0; *c*) 12.25

4.71 46/63

4.73 *a*) 2.5; 2.08

4.75 *a*) $E(X) = E(Y) = 1/3$ y $Var(X) = Var(Y) = 4/9$; *b*) $E(Z) = 2/3$ y $Var(Z) = 8/9$

4.77 *a*) 4; *b*) 32; 16

4.79 Mediante cálculo directo, $E(e^Y) = 1{,}884.32$. Con aproximación de ajuste de segundo orden, $E(e^Y) \approx 1{,}883.38$, que es muy cercano al valor real.

Capítulo 5

5.1 3/10

5.3 $\mu = 5.5$; $\sigma^2 = 8.25$

5.5 *a*) 0.0480; *b*) 0.2375; *c*) $P(X = 5|p = 0.3) = 0.1789$, $P = 0.3$ es razonable.

5.7 *a*) 0.0474; *b*) 0.0171

5.9 *a*) 0.7073; *b*) 0.4613; *c*) 0.1484

5.11 0.1240

5.13 0.8369

5.15 *a*) 0.0778; *b*) 0.3370; *c*) 0.0870

5.17 $\mu \pm 2\sigma = 3.5 \pm 2.05$

5.19 $f(x_1, x_2, x_3) = \binom{n}{x_1, x_2, x_3} 0.35^{x_1} 0.05^{x_2} 0.60^{x_3}$

5.21 0.0095

5.23 0.0077

5.25 0.8670

5.27 *a*) 0.2852; *b*) 0.9887; *c*) 0.6083

5.29 *a*) 0.3246; *b*) 0.4496

5.31 5/14

5.33 $h(x; 6, 3, 4) = \frac{\binom{4}{x}\binom{2}{3-x}}{\binom{6}{3}}$, para $x = 1, 2, 3$; $P(2 \leq X \leq 3) = 4/5$

5.35 0.9517

5.37 *a*) 0.6815; *b*) 0.1153

5.39 3.25; desde 0.52 hasta 5.98

5.41 0.9453

5.43 0.6077

5.45 *a*) 4/33; *b*) 8/165

5.47 0.2315

5.49 *a*) 0.3991; *b*) 0.1316

5.51 0.0515

5.53 *a*) 0.3840; *b*) 0.0067

5.55 63/64

5.57 *a*) 0.0630; *b*) 0.9730

5.59 *a*) 0.1429; *b*) 0.1353

5.61 *a*) 0.1638; *b*) 0.032

5.63 *a*) 0.3840; *b*) 0.1395; *c*) 0.0553

5.65 0.2657

5.67 *a*) $\mu = 4$; $\sigma^2 = 4$; *b*) Desde 0 hasta 8.

5.69 *a*) 0.2650; *b*) 0.9596

5.71 *a*) 0.8243; *b*) 14

5.73 4

5.75 5.53×10^{-4}; $\mu = 7.5$

5.77 *a*) 0.0137; *b*) 0.0830

5.79 0.4686

Capítulo 6

6.1 *a*) 0.9236; *b*) 0.8133; *c*) 0.2424; *d*) 0.0823; *e*) 0.0250; *f*) 0.6435

6.3 *a*) −1.72; *b*) 0.54; *c*) 1.28

6.5 *a*) 0.1151; *b*) 16.1; *c*) 20.275; *d*) 0.5403

6.7 *a*) 0.8980; *b*) 0.0287; *c*) 0.6080

6.9 *a*) 0.0548; *b*) 0.4514; *c*) 23; *d*) 189.95 mililitros

6.11 *a*) 0.0571; *b*) 99.11%; *c*) 0.3974; *d*) 27.952 minutos; *e*) 0.0092

6.13 6.24 años

6.15 *a*) 51%; *b*) \$18.37

6.17 *a*) 0.0401; *b*) 0.0244

6.19 26

6.21 *a*) 0.6; *b*) 0.7; *c*) 0.5

6.23 *a*) 0.8006; *b*) 0.7803

6.25 *a*) 0.3085; *b*) 0.0197

6.27 *a*) 0.9514; *b*) 0.0668

6.29 *a*) 0.1171; *b*) 0.2049

6.31 0.1357

6.33 *a*) 0.0778; *b*) 0.0571; *c*) 0.6811

6.35 *a*) 0.8749; *b*) 0.0059

6.37 *a*) 0.0228; *b*) 0.3974

6.39 $2.8e^{-1.8} - 3.4e^{-2.4} = 0.1545$

6.43 *a*) $\mu = 6$; $\sigma^2 = 18$; *b*) desde 0 hasta 14.485 millones de litros.

6.45 $\sum_{x=4}^{6} \binom{6}{x}(1 - e^{-3/4})^x (e^{-3/4})^{6-x} = 0.3968$

6.47 *a*) $\sqrt{\pi/2} = 1.2533$; *b*) e^{-2}

6.49 $e^{-4} = 0.0183$

6.51 *a*) $\mu = \alpha\beta = 50$; *b*) $\sigma^2 = \alpha\beta^2 = 500$; $\sigma = \sqrt{500}$; *c*) 0.815

6.53 *a*) 0.1889; *b*) 0.0357

6.55 Media=e^6, varianza=$e^{12}(e^4 - 1)$

6.57 *a*) e^{-10}; *b*) $\beta = 0.10$

Capítulo 7

7.1 $g(y) = 1/3$; para $y = 1, 3, 5$

7.3
$$g(y_1, y_2) = \left(\frac{2}{\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}, 2-y_1}\right) \times \left(\frac{1}{4}\right)^{(y_1+y_2)/2} \left(\frac{1}{3}\right)^{(y_1-y_2)/2} \left(\frac{5}{12}\right)^{2-y_1};$$
para $y_1 = 0, 1, 2$; $y_2 = -2, -1, 0, 1, 2$; $y_2 \leq y_1$; $y_1 + y_2 = 0, 2, 4$

7.7 Distribución gamma con $\alpha = 3/2$ y $\beta = m/2b$

7.9 *a*) $g(y) = 32/y^3$, para $y > 4$; *b*) 1/4

7.11 $h(z) = 2(1 - z)$, para $0 < z < 1$

7.13 $h(w) = 6 + 6w - 12w^{1/2}$, para $0 < w < 1$

7.15
$$g(y) = \begin{cases} \frac{2}{9\sqrt{y}}, & 0 < y < 1, \\ \frac{\sqrt{y}+1}{9\sqrt{y}}, & 1 < y < 4 \end{cases}$$

7.19 Ambas son iguales a μ

7.23 *a*) Gamma(2,1); *b*) Uniforme(0,1)

Capítulo 8

8.1 *a*) Las repuestas de todas las personas en Richmond que tienen un teléfono;

b) Resultados para un número grande o infinito de lanzamientos de una moneda;

c) Periodo de vida de tal calzado deportivo cuando es utilizado en el torneo profesional;

d) Todos los posibles intervalos de tiempo para esta abogada que maneja desde su casa hasta su oficina.

8.3 *a*) $\bar{x} = 2.4$; *b*) $\bar{x} = 2$; *c*) $m = 3$

8.5 *a*) $\bar{x} = 3.2$ segundos; *b*) $\bar{x} = 3.1$ segundos

8.7 *a*) 53.75; *b*) 75 y 100

8.9 *a*) El rango es 10; *b*) $s = 3.307$

8.11 *a*) 2.971; *b*) 2.971

8.13 $s = 0.585$

8.15 *a*) 45.9; *b*) 5.1

8.17 0.3159

8.19 *a*) Se reduce de 0.7 a 0.4;
b) Se incrementa de 0.2 a 0.8

8.21 Sí.

8.23 *a*) $\mu = 5.3$; $\sigma^2 = 0.81$;
b) $\mu_{\bar{X}} = 5.3$; $\sigma^2_{\bar{X}} = 0.0225$;
c) 0.9082

8.25 *a*) 0.6398; *b*) 7.35

8.29 0.5596

8.33 *a*) 0.1977; *b*) No

8.35 *a*) 1/2; *b*) 0.3085

8.37 $P(\bar{X} \leq 775|\mu = 760) = 0.9332$

8.39 *a*) 27.488; *b*) 18.475; *c*) 36.415

8.41 *a*) 0.297; *b*) 32.852; *c*) 46.928

8.43 *a*) 0.05; *b*) 0.94

8.47 *a*) 0.975; *b*) 0.10; *c*) 0.875; *d*) 0.99

8.49 *a*) 2.500; *b*) 1.319; *c*) 1.714

8.51 No; $\mu > 20$

8.53 *a*) 2.71; *b*) 3.51; *c*) 2.92;
d) 0.47; *e*) 0.34

8.55 La razón F es 1.44. Las varianzas no son significativamente diferentes.

Capítulo 9

9.5 $0.3097 < \mu < 0.3103$

9.7 *a*) $22{,}496 < \mu < 24{,}504$; *b*) error ≤ 1004

9.9 35

9.11 56

9.13 $0.978 < \mu < 1.033$

9.15 $47.722 < \mu < 49.278$

9.17 323.946 hasta 326.154

9.19 11,426 hasta 35,574

9.23 La varianza de S'^2 es menor.

9.25 (6.05,16.55)

9.27 (1.6358, 5.9376)

9.29 Límite de predicción superior: 9.42;
Límite de tolerancia superior: 11.87

9.33 Sí, el valor de 6.9 está fuera del intervalo de predicción.

9.35 $2.9 < \mu_1 - \mu_2 < 7.1$

9.37 $2.80 < \mu_1 - \mu_2 < 3.40$

9.39 $1.5 < \mu_1 - \mu_2 < 12.5$

9.41 $0.70 < \mu_1 - \mu_2 < 3.30$

9.43 $-6{,}536 < \mu_1 - \mu_2 < 2{,}936$

9.45 $(-0.74, 6.29)$

9.47 $(-6.92, 36.70)$

9.49 $0.54652 < \mu_B - \mu_A < 1.69348$

9.51 *a*) $0.498 < p < 0.642$; *b*) error ≤ 0.072

9.53 $0.194 < p < 0.262$

9.55 *a*) $0.739 < p < 0.961$; *b*) no

9.57 *a*) $0.644 < p < 0.690$; *b*) error ≤ 0.023

9.59 2,576

9.61 160

9.63 16,577

9.65 $-0.0136 < p_F - p_M < 0.0636$

9.67 $0.0011 < p_1 - p_2 < 0.0869$

9.69 $(-0.0849, 0.0013)$; no es significativamente diferente.

9.71 $0.293 < \sigma^2 < 6.736$; la afirmación es válida

9.73 $1.863 < \sigma < 3.578$

9.75 $9.265 < \sigma < 34.16$

9.77 $0.545 < \sigma_1/\sigma_2 < 2.690$

9.79 $0.016 < \sigma_1^2/\sigma_2^2 < 0.454$; no

9.81 $\frac{1}{n}\sum_{i=1}^{n} x_i$

9.83 *a)* $L(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma^n}\prod_{i=1}^{n}\left(\frac{1}{x_i}\right)e^{-\sum_{i=1}^{n}(\ln x_i-\mu)^2/2\sigma^2}$

b) $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln x_i$;

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\ln x_i - \left(\frac{1}{n}\sum_{j=1}^{n}\ln x_j\right)\right]^2$

9.85 $x\ln p + (1-x)\ln(1-p)$. Sea la derivada con respecto a $p = 0$; $\hat{p} = x = 1.0$

Capítulo 10

10.1 *a)* Concluya que menos del 30% del público son alérgicos a ciertos productos de queso cuando, de hecho, 30% o más son alérgicos.

b) Concluya que menos del 30% del público son alérgicos a ciertos prdoductos de queso cuando, de hecho, menos del 30% lo son.

10.3 *a)* La empresa no es culpable;

b) la empresa es culpable.

10.5 *a)* 0.1286;

b) $\beta = 0.0901$; $\beta = 0.0708$.

c) La probabilidad de un error tipo I es un tanto grande.

10.7 *a)* 0.0559;
b) $\beta = 0.0017$; $\beta = 0.00968$; $\beta = 0.5557$

10.9 *a)* $\alpha = 0.0032$; *b)* $\beta = 0.0062$

10.11 *a)* $\alpha = 0.1357$; *b)* $\beta = 0.2578$

10.13 *a)* $\alpha = 0.0094$; *b)* $\beta = 0.0122$

10.15 *a)* $\alpha = 0.0718$; *b)* $\beta = 0.1151$

10.17 *a)* $\alpha = 0.0384$; *b)* $\beta = 0.5$; $\beta = 0.2776$

10.19 $z = -1.64$; valor $P = 0.10$

10.21 $z = -2.76$; sí, $\mu < 40$ meses;
valor $P = 0.0029$

10.23 $z = 8.97$; sí $\mu > 20{,}000$ kilómetros;
valor $P < 0.001$

10.25 $t = 0.77$; no rechace H_0.

10.27 $z = 12.72$; valor $P < 0.0005$

10.29 $t = -1.98$; rechace H_0; valor $P = 0.0312$

10.31 $z = -2.60$; concluya que $\mu_A - \mu_B \leq 12$ kilogramos.

10.33 $t = 1.50$; no hay evidencia suficiente para concluir que el incremento en la concentración de sustrato causaría un incremento en la velocidad media en más de 0.5 micromoles por 30 minutos.

10.35 $t = 0.70$; no hay suficiente evidencia de que el suero sea efectivo.

10.37 $t = 2.55$; rechace H_0: $\mu_1 - \mu_2 > 4$ kilómetros.

10.39 $t' = 0.22$; no rechace H_0.

10.41 $t' = 2.76$; rechace H_0.

10.43 $t = 2.48$; valor $P < 0.02$; rechace H_0.

10.45 $t = -2.53$; rechace H_0; la afirmación es válida.

10.47 $n = 6$

10.49 $78.28 \approx 79$ debido al redondeo.

10.51 5

10.53 *a)* H_0: $M_{\text{caliente}} - M_{\text{frío}} = 0$,
H_1: $M_{\text{caliente}} - M_{\text{frío}} \neq 0$,

b) t apareada, $t = 0.99$; valor $P > 0.30$; no rechace H_0.

10.55 Valor $P = 0.4044$ = (con una prueba de una cola); no se refuta la afirmación.

10.57 Valor $P = 0.0207$; sí, la moneda no es legal.

10.59 $z = -5.06$ y valor $P \approx 0$; concluya que menos de 1/5 de los hogares se calienta con petróleo.

10.61 $z = 1.44$; no rechace H_0.

10.63 $z = 2.36$ con valor $P = 0.0182$; sí, la diferencia es significativa.

10.65 $z = 1.10$ con valor $P = 0.1357$; no tenemos evidencia suficiente para concluir que el cáncer de mama sea más frecuente en las comunidades urbanas.

10.67 $\chi^2 = 18.13$ con valor $P = 0.0676$ (de la salida del programa); no rechace H_0: $\sigma^2 = 0.03$.

10.69 $\chi^2 = 63.75$ con valor $P = 0.8998$ (de la salida del programa); no rechace H_0.

10.71 $\chi^2 = 42.37$ con valor $P = 0.0117$ (de la salida del programa); la máquina está fuera de control.

10.73 $f = 1.33$ con valor $P = 0.3095$ (de la salida del programa); no rechace H_0: $\sigma_1 = \sigma_2$.

10.75 $f = 0.75$ con valor $P = 0.3186$ (de la salida del programa); no rechace H_0: $\sigma_1 = \sigma_2$.

10.77 $f = 19.6$ con valor $P = 0.0008$ (de la salida del programa); rechace H_0: $\sigma_1 = \sigma_2$.

10.79 $\chi^2 = 4.47$; no hay evidencia suficiente para afirmar que la moneda no sea legal.

10.81 $\chi^2 = 10.14$; rechace H_0, la razón no es 5:2:2:1.

10.83 $\chi^2 = 2.33$; no rechace H_0: distribución binomial.

10.85 $\chi^2 = 2.57$; no rechace H_0: distribución geométrica.

10.89 $\chi^2 = 5.19$; no rechace H_0: distribución normal.

10.91 $\chi^2 = 5.47$; no rechace H_0.

10.93 $\chi^2 = 124.59$; sí la ocurrencia de estos tipos de delitos es dependiente del distrito de la ciudad.

10.95 $\chi^2 = 31.17$ con valor $P < 0.0001$; las actitudes no son homogéneas.

10.97 $\chi^2 = 5.92$ con valor $P = 0.4332$; no rechace H_0.

10.99 $\chi^2 = 1.84$; no rechace H_0.

Capítulo 11

11.1 *a*) $a = 64.529$, $b = 0.561$;
b) $\hat{y} = 81.4$

11.3 *a*) $\hat{y} = 6.4136 + 1.8091x$;
b) $\hat{y} = 9.560$ a temperatura 1.75

11.5 *a*) $\hat{y} = 5.8254 + 0.5676x$;
c) $\hat{y} = 34.205$ a 50 °C

11.7 *b*) $\hat{y} = 343.706 + 3.221x$;
c) $\hat{y} = \$456$ donde los costos de publicidad son \$35

11.9 *a*) $\hat{y} = 153.175 - 6.324x$;
b) $\hat{y} = 123$ a $x = 4.8$ unidades

11.11 *b*) $\hat{y} = -1847.633 + 3.653x$;

11.13 *b*) $\hat{y} = 31.709 + 0.353x$;

11.17 *a*) $s^2 = 176.4$;
b) $t = 2.04$; no rechace H_0: $\beta = 0$.

11.19 *a*) $s^2 = 0.40$;
b) $4.324 < \alpha < 8.503$;
c) $0.446 < \beta < 3.172$

11.21 *a*) $s^2 = 6.626$;
b) $2.684 < \alpha < 8.968$;
c) $0.498 < \beta < 0.637$

11.23 $t = -2.24$; rechace H_0: $\beta < 6$

11.25 *a*) $24.438 < \mu_Y|_{24.5} < 27.106$;
b) $21.88 < y_0 < 29.66$

11.27 $8.856 < \mu_Y|_{1.6} < 9.761$

11.29 *a*) 17.1812;
b) no (el intervalo de confianza de 95% sobre la media mpg es [27.95, 29.60]);
c) las millas por galón probablemente excederán 18

11.33 *b*) $\hat{y} = 3.4156x$

11.35 *a*) $b = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$;
b) $\hat{y} = 2.003x$

11.37 $E(B) = \beta + \gamma \dfrac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1) x_{2i}}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}$.

11.39 *a*) $a = 10.812$, $b = -0.3437$;
b) $f = 0.43$; la regresión es lineal.

11.41 $f = 1.12$; la regresión es lineal.

11.43 $f = 1.71$ y valor $P = 0.2517$; la regresión es lineal.

11.45 *a*) $\hat{P} = -11.3251 - 0.0449T$;
b) sí;
c) $R^2 = 0.9355$;
d) sí

11.47 *b*) $\hat{N} = -175.9025 + 0.0902Y$; $R^2 = 0.3322$.

11.49 $r = 0.240$

11.53 *a*) $r = 0.392$;
b) $t = 2.04$; no rechace H_0: $\rho = 0$; sin embargo, el valor $P = 0.053$ es marginal.

Capítulo 12

12.1 *a*) $\hat{y} = 27.547 + 0.922x_1 + 0.284x_2$;
b) $\hat{y} = 84$ a $x_1 = 64$ y $x_2 = 4$.

12.3 $\hat{y} = 0.5800 + 2.7122x_1 + 2.0497x_2$.

12.5 *a*) $\hat{y} = 56.4633 + 0.1525x - 0.00008x^2$;
b) $\hat{y} = 86.7\%$ cuando la temperatura es 225 °C.

12.7 $\hat{y} = 141.6118 - 0.2819x + 0.0003x^2$.

12.9 *a*) $\hat{y} = -102.7132 + 0.6054x_1 + 8.9236x_2 + 1.4374x_3 + 0.0136x_4$;
b) $\hat{y} = 287.6$

12.11 $\hat{y} = 3.3205 + 0.4210x_1 - 0.2958x_2 + 0.0164x_3 + 0.1247x_4$.

12.13 $\hat{y} = -6.5122 + 1.9994x_1 - 3.6751x_2 + 2.5245x_3 + 5.1581x_4 + 14.4012x_5$.

12.15 *a*) $\hat{y} = 350.9943 - 1.2720x_1 - 01539x_2$;
b) $\hat{y} = 140.9$

12.17 0.1651

12.19 242.72

12.21 *a*) $\hat{\sigma}^2_{B_2} = 28.0955$; *b*) $\hat{\sigma}_{B_1B_2} = -0.0096$

12.23 $29.93 < \mu_{Y|19.5} < 31.97$

12.25 $t = 2.86$; rechace H_0 y favorezca $\beta_2 > 0$.

12.27 $t = 3.55$ con valor $P = 0.01$; rechace H_0 y favorezca $\beta_1 > 2$.

12.29 *a*) t $= -1.09$ con valor $P = 0.3562$;
b) $t = -1.72$ con valor $P = 0.1841$;
c) sí; no hay suficiente evidencia que demuestre que los valores de x_1 y x_2 son significativos.

12.31 $R^2 = 0.9997$

12.33 $f = 5.106$ con valor $P = 0.0303$; la regresión no es significativa en el nivel 0.01.

12.35 $f = 34.90$ con valor $P = 0.0002$; rechace H_0 y concluya que $\beta_1 > 0$.

12.37 $f = 10.18$ con valor $P < 0.01$; x_1 y x_2 son significativos en la presencia de x_3 y x_4.

12.39 Es mejor el modelo de dos variables.

12.41 Primer modelo: $R^2_{\text{aju}} = 92.7\%$, $CV = 9.0385$;
Segundo modelo: $R^2_{\text{aju}} = 98.1\%$, $CV = 4.6287$;
La prueba f parcial muestra un valor $P = 0.0002$; es mejor el modelo 2.

12.43 Utilizando x_2 solo no es muy diferente de usar x_1 y x_2 juntos, ya que las R^2_{aju} son 0.7696 contra 0.7591.

12.45 $\widehat{mpg} = 5.9593 - 0.00003773$ odómetro $+ 0.3374$ octanaje $- 12.6266z_1 - 12.9846z_2$;
b) sedán
c) no son significativamente diferentes.

12.47 *b*) $\hat{y} = 4.690$ segundos;
c) $4.450 < \mu_{Y|\{180,260\}} < 4.930$

12.49 $\hat{y} = 2.1833 + 0.9576x_2 + 3.3253x_3$

12.51 *a*) $\hat{y} = -587.211 + 428.433x$;
b) $\hat{y} = 1180 - 191.691x + 35.20945x^2$;
c) modelo cuadrático

12.53 $\hat{\sigma}^2_{B_1} = 20,588$; $\hat{\sigma}^2_{B_{11}} = 62.6502$;
$\hat{\sigma}_{B_1,B_{11}} = -1103.5$

12.55 *a*) Es mejor el modelo de intersección.

12.57 *a*) $\hat{y} = 3.1368 + 0.6444x_1 - 0.0104x_2 + 0.5046x_3 - 0.1197x_4 - 2.4618x_5 + 1.5044x_6$;
b) $\hat{y} = 4.6563 + 0.5133x_3 - 0.1242x_4$;
d) $\hat{y} = 4.6563 + 0.5133x_3 - 0.1242x_4$;
e) dos observaciones tienen valores grandes de R de student y deben verificarse.

12.59 *a*) $\hat{y} = 125.8655 + 7.7586x_1 + 0.0943x_2 - 0.0092x_1x_2$;
b) el modelo con x_2 solo es el mejor.

Capítulo 13

13.3 $f = 0.31$; no hay evidencia suficiente para sostener que existen diferencias entre las 6 máquinas.

13.5 $f = 14.52$; sí, la diferencia es significativa.

13.7 $f = 2.25$; no hay evidencia suficiente para sostener que las diferentes concentraciones de $MgNH_4PO_4$ afectan significativamente la altura que alcanzan los crisantemos.

13.9 $f = 8.38$; las actividades específicas promedio difieren de manera significativa.

13.11 *a*) $f = 14.27$; rechace H_0;
b) $f = 23.23$; rechace H_0;
c) $f = 2.48$; no rechace H_0.

13.13 *a*) $f = 13.50$; las medias de los tratamientos son diferentes;
b) $f(1\ vs.\ 2) = 29.35$; significativo;
$f(3\ vs.\ 4) = 3.59$; no significativo

13.15

$\bar{x}_3$	$\bar{x}_1$	$\bar{x}_4$	$\bar{x}_2$
56.52	59.66	61.12	61.96

13.17 *a*) $f = 9.01$; sí, significativo;
b)

Disminución	De Hess modificado	Remoción del sustrato, de Kicknet	Surber	Kicknet

13.19 Comparación del control para 1 y 2; significativo; Comparación del control para 3 y 4; no significativo.

13.21 La absorción media para el agregado 4 es significativamente menor que para el otro agregado.

13.23 $f = 70.27$ con valor $P < 0.0001$; rechace H_0.

$\bar{x}_0$	$\bar{x}_{25}$	$\bar{x}_{100}$	$\bar{x}_{75}$	$\bar{x}_{50}$
55.167	60.167	64.167	70.500	72.833

La temperatura es importante; tanto 75 como 50° (C) producen baterías con vida activa significativamente más larga.

13.27 *a*) f(fertilizante) = 6.11; significativo;
b) $f = 17.37$; significativo;
$f = 0.96$; no significativo

13.29 $f = 5.99$; el porcentaje de aditivos externos no es el mismo para las tres marcas de mermelada; marca A.

13.31 f(estación) = 26.14; significativo

13.33 f(dieta) = 11.86; significativo

13.35 $f = 0.58$; no significativo

13.39 $f = 5.03$; las calificaciones no están influidas por los diferentes profesores.

13.41 $p > 0.0001$;
$f = 122.37$; la cantidad de tinta sí influye en el color de la tela.

13.43 *a*) $f = 14.9$; los operadores difieren significativamente;
b) $\hat{\sigma}_\alpha^2 = 28.91$; $s^2 = 8.32$.

13.45 *a*) $f = 3.33$; sin diferencia significativa; pero el valor $P = 0.0564$ es marginal;
b) $\hat{\sigma}_\alpha^2 = 1.08$; $s^2 = 2.25$.

13.49 9.

13.51 *a*) $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $\alpha_i \sim n(x; 0, \sigma_\alpha)$;
b) $\hat{\alpha}_\alpha^2 = 0$ (el componente de varianza estimado es -0.00027); $\hat{\sigma}^2 = 0.0206$.

13.53 *a*) $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $\alpha_i \sim n(x; 0, \sigma_\alpha)$;
b) sí; $f = 5.63$ con valor $P = 0.0121$;
c) hay un componente de varianza del telar significativo.

Capítulo 14

14.1 *a*) $f = 8.13$; significativo;
b) $f = 5.18$; significativo;
c) $f = 1.63$; no significativo

14.3 *a*) $f = 14.81$; significativo;
b) $f = 9.04$; significativo;
c) $f = 0.61$; no significativo;

14.5 *a*) $f = 34.40$; significativo;
b) $f = 29.95$; significativo;
c) $f = 20.30$; significativo;

14.7 Prueba del efecto de la cantidad de catalizador: $f = 46.63$ con valor $P = 0.0001$;
Prueba del efecto de la temperatura: $f = 10.85$ con valor $P = 0.0002$;
Prueba del efecto de la interacción: $f = 2.06$ con valor $P = 0.074$.

14.9 *a*)

Fuente de variación	gl	Suma de cuadrados	Cuadrados medios	f	P
Velocidad de corte	1	12.000	12.000	1.32	0.2836
Geometría de la herramienta	1	675.000	675.000	74.31	< 0.0001
Interacción	1	192.000	192.000	21.14	0.0018
Error	8	72.667	9.083		
Total	11	951.667			

b) El efecto de la interacción oculta el efecto de la velocidad de corte;

c) $f_{\text{geometría de la herramienta=1}} = 16.51$ y valor $P = 0.0036$;

d) $f_{\text{geometría de la herramienta=2}} = 5.94$ y valor $P = 0.0407$.

14.11 *a*)

Fuente de variación	gl	Suma de cuadrados	Cuadrados medios	f	P
Método	1	0.00010414	0.00010414	6.57	0.0226
Laboratorio	6	0.00805843	0.00134307	84.70	< 0.0001
Interacción	6	0.00019786	0.00003298	2.08	0.1215
Error	14	0.000222	0.00001586		
Total	27	0.00858243			

b) La interacción no es significativa;

c) Ambos efectos principales son significativos;

e) $f_{\text{laboratorio=1}} = 0.01576$ y valor $P = 0.9019$; no hay diferencia significativa de los métodos en el laboratorio 1;
$f_{\text{geometría de la herramienta=2}} = 9.081$ y valor $P = 0.0093$.

14.13 *b*)

Fuente de variación	gl	Suma de cuadrados	Cuadrados medios	f	P
Método	1	0.06020833	0.06020833	157.07	< 0.0001
Laboratorio	1	0.06020833	0.06020833	157.07	< 0.0001
Interacción	1	0.00000833	0.00000833	0.02	0.8864
Error	8	0.00306667	0.00038333		
Total	11	0.12349167			

c) Tanto el tiempo como el tratamiento influyen significativamente en la absorción del magnesio, aunque no existe interacción significativa entre ambos.

d) $y = \mu + \beta_T \text{ Tiempo} + \beta_Z Z + \beta_{TZ} \text{ Tiempo} * Z + \epsilon$, donde $Z = 1$ cuando el tratamiento $= 1$ y $Z = 0$ cuando el tratamiento $= 2$;

e) $f = 0.02$ con valor $P = 0.8864$; la interacción en el modelo no es significativa

14.15 *a*) AB: $f = 3.83$; significativo;
AC : $f = 3.79$; significativo;
BC : $f = 1.31$; no es significativo;
ABC : $f = 1.63$; no es significativo;

b) A : $f = 0.54$; no es significativo;
B : $f = 6.85$; significativo;
C : $f = 2.15$; no es significativo;

c) La presencia de la interacción AC oculta el efecto principal C.

14.17 *a*) Esfuerzo cortante $f = 45.96$ con valor $P < 0.0001$;
recubrimiento $f = 0.05$ con valor $P = 0.8299$;
humedad $f = 2.13$ con valor $P = 0.1257$;
recubrimiento × humedad $f = 3.41$ con valor $P = 0.0385$;
recubrimiento × e. cortante $f = 0.08$ con valor $P = 0.9277$;
humedad × e. cortante $f = 3.15$ con valor $P = 0.0192$;
recubrimiento × humedad × e. cortante $f = 1.93$ con valor $P = 0.1138$.

b) La mejor combinación parece ser sin recubrimiento, humedad media y nivel de e. cortante de 20.

14.19

Efecto	f	P
Temperatura	14.122	< 0.0001
Superficie	6.70	0.0020
HRC	1.67	0.1954
T×S	5.50	0.0006
T×HRC	2.69	0.0369
S×HRC	5.41	0.0007
T×S×HRC	3.02	0.0051

14.21 *a*) sí; marca × tipo; marca × temperatura;

b) sí;

c) marca *Y*, detergente en polvo, alta temperatura.

14.23 *a*)

Efecto	f	P
Tiempo	543.53	< 0.0001
Temperatura	209.79	< 0.0001
Solvente	4.97	0.0457
Tiempo × temperatura	2.66	0.1103
Tiempo × solvente	2.04	0.1723
Tiempo × solvente	0.03	0.8558
Tiempo × temperatura × solvente	6.22	0.0140

Aunque las tres interacciones bilaterales se muestran insignificantes, podrían estar ocultas por la interacción trilateral significativa;

14.25 *a*) $f = 1.49$; no hay interacción significativa;

b) f(operadores) = 12.45; significativo;
f(filtros) = 8.39; significativo;

c) $\hat{\sigma}_\alpha^2 = 0.1701$ (filtros);
$\hat{\sigma}_\beta^2 = 0.3514$ (operadores);
$s^2 = 0.1867$

14.27 *a*) $\hat{\sigma}_\beta^2$, $\hat{\sigma}_\gamma^2$, $\hat{\sigma}_{\alpha\gamma}^2$ son significativos;
b) $\hat{\sigma}_\gamma^2$ y $\hat{\sigma}_{\alpha\gamma}^2$ son insignificantes

14.29 0.59

14.31 *a*) modelo combinado;
b) Material: $f = 47.42$ con valor $P < 0.0001$;
Marca: $f = 1.73$ con valor $P = 0.2875$;
Material × marca: 16.06 con valor $P = 0.0004$;
c) no

14.33 *a*) $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, (modelo combinado);
A = flujo de energía, B = tipo de cereal;
$\beta_j \sim n(x; 0, \sigma_\beta^2)$, independiente;
$(AB)_{ij} \sim n(x; 0, \sigma_{\alpha\beta}^2)$, independiente;
$\epsilon_{ijk} \sim n(x; 0, \sigma^2)$, independiente;
b) no;
c) no

Capítulo 15

15.1 $SSA = 2.6667$, $SSB = 170.6667$, $SSC = 104.1667$, $SS(AB) = 1.5000$, $SS(AC) = 42.6667$, $SS(BC) = 0.0000$, $SS(ABC) = 1.5000$.

15.3 Los factores A, B y C tienen efectos negativos sobre el compuesto de fósforo, y el factor D tiene un efecto positivo. Sin embargo, la interpretación del efecto de los factores individuales debería implicar el uso de las gráficas de interacción.

15.5 Efectos significativos
A: $f = 9.98$; C: $f = 6.54$; BC: $f = 19.3$.
Efectos insignificantes
B: $f = 0.20$; D: $f = 0.02$; AB: $f = 1.83$;
AC: $f = 0.20$; AD: $f = 0.57$; BD: $f = 1.83$;
CD: $f = 0.02$.

15.9 *a*) $b_A = 5.5$, $b_B = -3.25$ y $b_{AB} = 2.5$;
b) Los valores de los coeficientes son de la mitad de los efectos;
c) $t_A = 5.99$ con valor $P = 0.0039$;
$t_B = -3.54$ con valor $P = 0.0241$;
$t_{AB} = 2.72$ con valor $P = 0.0529$;
$t_2 = F$.

15.11 *a*) $A = -0.8750$, $B = 5.8750$, $C = 9.6250$, $AB = -3.3750$, $AC = -9.6250$, $BC = 0.1250$ y ABC = -1.1250;
B, C, AB y AC parecen importantes con base en sus magnitudes.

b)

Efectos	**Valor *P***
A	0.7528
B	0.0600
C	0.0071
AB	0.2440
AC	0.0071
BC	0.9640
ABC	0.6861

c) sí;
d) Con un nivel alto de A, esencialmente C no tiene efecto. Con un nivel bajo de A, C tiene un efecto positivo.

15.13 A, B, C, AC, BC y ABC, cada uno con un grado de libertad pueden probarse usando un error cuadrático medio con 12 grados de libertad. Cada una de las tres réplicas contiene 2 bloques con AB confundidos.

15.15

Bloque 1	Bloque 2	Bloque 3	Bloque 4
(1)	*c*	*d*	*a*
ab	*abc*	*ac*	*b*
acd	*ad*	*bc*	*cd*
bcd	*bd*	*abd*	*abcd*

CD también se confunde con los bloques.

15.17

Réplica 1		Réplica 1		Réplica 1	
B1	B2	B1	B2	B1	B2
abc	*ab*	*abc*	*ab*	(1)	*a*
a	*ac*	*a*	*ac*	*c*	*b*
b	*bc*	*b*	*bc*	*ab*	*ac*
c	(1)	*c*	(1)	*abc*	*bc*
ABC		ABC		AB	
confundido		confundido		confundido	

15.19 *a*)

Máquina			
1	**2**	**3**	**4**
(1)	*c*	*a*	*ac*
ab	*d*	*b*	*ad*
cd	*e*	*acd*	*ae*
ce	*abc*	*ace*	*bc*
de	*abd*	*ade*	*bd*
abcd	*abe*	*bcd*	*be*
abce	*cde*	*bce*	*acde*
abde	*abcde*	*bde*	*bcde*

b) AB, CDE, $ABCDE$ (un diseño posible).

15.21 *a*) x_2, x_3, x_1x_2 y x_1x_3;

b) Curvatura: valor $P = 0.0073$;

c) Un punto de diseño adicional diferente de los originales.

15.23 $(0,-1)$, $(0,1)$, $(-1,0)$, $(1,0)$ podría utilizarse.

15.25 *a*) con BCD como el contraste definitorio, el bloque principal contiene (1), a, bc, abc, bd, abd, cd, acd;

b)

Bloque 1	Bloque 2
(1)	a
bc	abc
abd	bd
acd	cd

confundido por ABC;

c) el contraste definitorio BCD produce los siguientes resultados: $A = ABCD$, $B = CD$, $C = BD$, $D = BC$, $AB = ACD$, $AC = ABD$, y $AD = ABC$. Puesto que AD y ABC están confundidos con los bloques, hay sólo dos grados de libertad del error en las interacciones no confundidas.

Fuente de variación	Grado de libertad
A	1
B	1
C	1
D	1
Bloques	1
Error	2
Total	7

15.27 *a*) Con el contraste definitorio $ABCE$ y $ABDF$, el bloque principal contiene (1), ab, acd, bcd, ce, $abce$, ade, bde, acf, bcf, df, $abdf$, aef, bef, $cdef$, $abcdef$;

b) $A \equiv BCE \equiv BDF \equiv ACDEF$,
$AD \equiv BCDE \equiv BF \equiv ACEF$,
$B \equiv ACE \equiv ADF \equiv BCDEF$,
$AE \equiv BC \equiv BDEF \equiv ACDF$,
$C \equiv ABE \equiv ABCDF \equiv DEF$,
$AF \equiv BCEF \equiv BD \equiv ACDE$,
$D \equiv ABCDE \equiv ABF \equiv CEF$,
$CE \equiv AB \equiv ABCDEF \equiv DF$,
$E \equiv ABC \equiv ABDEF \equiv CDF$,
$DE \equiv ABCD \equiv ABEF \equiv CF$,
$F \equiv ABCEF \equiv ABD \equiv CDE$,
$BCD \equiv ADE \equiv ACF \equiv BEF$,
$AB \equiv CE \equiv DF \equiv ABCDEF$,
$BCF \equiv AEF \equiv ACD \equiv BDE$,
$AC \equiv BE \equiv BCDF \equiv ADEF$;

Fuente de variación	Grado de libertad
A	1
B	1
C	1
D	1
E	1
F	1
AB	1
AC	1
AD	1
BC	1
BD	1
CD	1
Error	3
Total	15

15.29

Fuente	gl	SS	MS	f	P
A	1	6.1250	6.1250	5.81	0.0949
B	1	0.6050	0.6050	0.57	0.5036
C	1	4.8050	4.8050	4.56	0.1223
D	1	0.2450	0.2450	0.23	0.6626
Error	3	3.1600	1.0533		
Total	7	14.9400			

15.31

Fuente	gl	SS	MS	f	P
A	1	388129.00	388129.00	3585.49	0.0001
B	1	277202.25	277202.25	2560.76	0.0001
C	1	4692.25	4692.25	43.35	0.0006
D	1	9702.25	9702.25	89.63	0.0001
E	1	1806.25	1806.25	16.69	0.0065
AD	1	1406.25	1406.25	12.99	0.0113
AE	1	462.25	462.25	4.27	0.0843
BD	1	1156.00	1156.00	10.68	0.0171
BE	1	961.00	961.00	8.88	0.0247
Error	6	649.50	108.25		
Total	15	686167.00			

Todos los efectos principales son significativos en el nivel 0.05; AD, BD y BE son también significativos en el nivel 0.05.

15.33 El bloque principal contiene af, be, cd, abd, ace, bcf, def, $abcdef$.

15.35 $A \equiv BD \equiv CE \equiv CDF \equiv BEF \equiv ABCF \equiv ADEF \equiv ABCDE$;
$B \equiv AD \equiv CF \equiv CDE \equiv AEF \equiv ABCE \equiv BDEF \equiv ABCDF$;
$C \equiv AE \equiv BF \equiv BDE \equiv ADF \equiv CDEF \equiv ABCD \equiv ABCEF$;
$D \equiv AB \equiv EF \equiv BCE \equiv ACF \equiv BCDF \equiv ACDE \equiv ABDEF$;
$E \equiv AC \equiv DF \equiv ABF \equiv BCD \equiv ABDE \equiv BCEF \equiv ACDEF$;

$F \equiv BC \equiv DE \equiv ACD \equiv ABE \equiv ACEF \equiv ABDF \equiv BCDEF$.

15.37 $\hat{y} = 12.7519 + 4.7194x_1 + 0.8656x_2 - 1.4156x_3$; las unidades están centradas y a escala; prueba de falta de ajuste, $F = 81.58$, con valor $P < 0.0001$.

15.39 *AFG*, *BEG*, *CDG*, *DEF*, *CEFG*, *BDFG*, *BCDE*, *ADEG*, *ACDF*, *ABEF* y *ABCDEFG*.

Capítulo 16

16.1 $x = 7$ con valor $P = 0.1719$; no rechace H_0.

16.3 $x = 3$ con valor $P = 0.0244$; rechace H_0.

16.5 $x = 4$ con valor $P = 0.3770$; no rechace H_0.

16.7 $x = 4$ con valor $P = 0.1335$; no rechace H_0.

16.9 $w = 43$; no rechace H_0.

16.11 $w_+ = 17.5$; no rechace H_0.

16.13 $z = -2.13$; rechace H_0 a favor de $\mu_1 - \mu_2 < 8$.

16.15 $u_1 = 1$; la afirmación es válida.

16.17 $u_2 = 5$; A opera más.

16.19 $u = 15$; no rechace H_0.

16.21 $h = 10.47$; los tiempos de operación son diferentes.

16.23 $v = 7$ con valor $P = 0.910$; muestra aleatoria.

16.25 $v = 6$ con valor $P = 0.044$; no rechace H_0.

16.27 $z = 1.11$; muestra aleatoria.

16.29 0.70

16.31 0.995

16.33 *a*) $rs = 0.39$; no rechace H_0.

16.35 *a*) $rs = 0.72$; *b*) rechace H_0 :, de manera que $\rho > 0$.

16.37 *a*) $rs = 0.71$; *b*) rechace H_0 :, de manera que $\rho > 0$.

Capítulo 18

18.1 $p^* = 0.173$

18.3 *a*) $f(p|x = 1) = 40p(1 - p)^3/0.2844$; $0.05 < p < 0.15$;
b) $p^* = 0.106$

18.5 $8.077 < \mu < 8.692$

18.7 *a*) 0.2509; *b*) $68.71 < \mu < 71.69$;
c) 0.0174

18.11 $p^* = \dfrac{6}{x + 2}$

Índice analítico

A

Agrupamiento en modelos multifactoriales, 594
Análisis de varianza (ANOVA), 264, 415, 511
  de un solo factor, 513
    comparaciones con un grado de libertad, 523
    contraste de, 524
    efecto del i-ésimo tratamiento, 514
    suma de cuadrados de los contrastes y, 525
    tratamiento del, 513
Aproximación
  binomial a hipergeométrica, 155
  normal a binomial, 187, 188
  poisson a binomial, 163

B

Bernoulli
  experimento de, 143, 144
  proceso de, 143
Bloques, 513

C

Coeficiente
  de correlación, 121, 432
    de la población, 434
    de rangos Spearman, 690, 691
    muestral, 434
    producto-momento de Pearson, 434
  de determinación, 407, 435
    ajustado, 467
    múltiple, 465
    variación, 474
Combinaciones, 46
Componentes de la varianza, 557
Confianza
  coeficiente de, 273
  grado de, 273
  intervalo de, 273, 274, 285
    límites de, 273, 275
Contrastes ortogonales, 525
Control
  de datos, 500
  en, estadístico, 698
    gráficas R, 704
    gráficas S, 711
    gráficas U, 720
  estadístico de la calidad, 697
    gráfica de, 697, 698
  fuera de, 698
  límites de, 699
Corrección de continuidad, 190
Correlación, 432
Covarianza, 115, 119
Cuadrados latinos, 549
Cuantil, 239
Curva característica de operación, 338

D

Datos históricos, 27
Desviación, 116
  estándar, 3, 15, 16, 28, 29, 116, 118, 131, 234
Diagrama de Venn, 36
Diseño
  completamente aleatorio, 513, 536
  experimental
    $\frac{1}{2}$ de un factorial, 648
    bloques, 535
    concepto de confusión, 639
    contraste, 613
    contraste definitorio, 640
    cuadrados latino, 549
    de bloques incompletos, 604
    de parámetros robustos, 661
    efectos principales, 574
    factores de control, 662
    factores de ruido, 662
    factorial fraccionario, 625, 647

interacción, 574
ortogonal, 631
por bloques completamente aleatorios, 537
relación definitoria, 648
resolución, 658
Distribución(es), 19, 23
a posteriori, 727
a priori, 726, 727
beta, 206
binomial, 143, 144, 188
de varianza, 147
negativa, 158-160
negativa y geométrica, 158, 160
condicional, 95
conjunta, 99
continua
de Weibull, 203, 204
exponencial, 194, 196
gamma, 195
logarítmica normal, 201
normal, 172
uniforme, 171
chi cuadrada, 200, 201
de Earlang, 206
de frecuencia, 23
de muestreo, 243, 244
de media, 244
de Poisson, 161, 162
de varianza, 163
de probabilidad, 80
conjunta, 91, 98
continuas, 84
de varianza, 116
de probabilidad discreta, 142
binomial, 143, 144
binomial negativa, 158, 159
geométrica, 158, 160
hipergeométrica, 152, 153
multinomial, 143, 149
exponencial, 194, 196
con tres factores, 590
de dos factores, 577
de falta (o fallo) de memoria y su efecto, 198
de varianza, 196
negativa, 196
relación con el proceso de Poisson, 196
F, 261-264
F no central, 560
gamma 194, 195
de varianza, 161
relación con el proceso de Poisson, 196
geométrica, 158, 160
de varianza, 161
hipergeométrica, 152, 154, 158, 160
de varianza, 154
multivariada, 156
logarítmica normal, 201
marginal, 93, 94, 98, 99
conjunta, 99
multinomial, 143, 149
normal, 172, 173
bivariada, 433
curva, 172, 173, 175
de varianza, 175
estándar, 177
normales, 19, 172, 173, 188
razón de varianza, 263
simétrica, 23
t, 257-259, 261
uniforme, 143, 171
discreta, 143, 144
Weibull, 202, 203
media de, 203
varianza de, 203

E
Efecto, 575
enmascarar, 575
Eliminación hacia atrás, 482
Erlang, distribución, 206
Error
al estimación de la media, 277
al estimación media, 276
estándar
de una estimación, 280
tipo I, 324
tipo II, 325
Espacio muestral, 31
continuo, 79
discreto, 79
partición, 53
Esperanza matemática, 107, 108, 111
Estadística
bayesiana, 732
descriptiva, 3
inferencial, 269
Estadístico, 232
C_p, 493, 494
de prueba, 324
Estimación, 270
de dos proporciones, 302
de la diferencia entre dos medias, 288

de la probabilidad máxima, 310, 311, 315
de la razón de dos varianzas, 308
de la varianza, 306
observaciones pareadas y, 294
por intervalo, 272, 273
bayesiano, 731
puntual, 270, 272
error estándar y, 280
proporción, 299
Estimado, 12
de unión de la varianza, 290
Estimador
de momentos, 317
de probabilidad máxima, 311, 313
insesgado, 270, 271
Estudio
observacional, 3, 26
retrospectivo, 27
Evento, 34
Experimentos
factorial, 573
en bloques, 594
en modelos multifactoriales, 594
interacción enmascarada, 575
modelo II, 600
modelo III, 602
modelo mixto, 602
factoriales 2^k
alias, 649
bloque principal, 641
bloques incompletos, 639
confusión, 639
contraste definitorio, 640
corridas centrales, 634
de efectos fijos, 555
diseño de Plackett-Burman, 660
diseño ortogonal, 631
factoriales fraccionarios, 647
parcialmente confundidos, 644
preparación de la regresión, 625
relación definitoria, 648
resolución, 658
sabio, 529

F

Factor, 26
Falta de ajuste, 419
Frecuencia relativa, 22, 23, 28, 29, 107
Función,
de la distribución acumulada, 81, 86
de masa de probabilidad, 80
o conjunta, 92
de pérdida del error cuadrático, 732
gamma, 194
probabilística, 80

G

Grado de libertad, 15, 16, 255, 256
Gráfica(s)
cusum, 721
de barras, 82
de caja, 3, 236
de control, 636
para atributos, 713
para variables, 700
de cuantiles, 236, 238, 239, 240, 241
de puntos, 3, 8, 29
de probabilidad normal, 236
R, 704
S, 711
U, 720
$\bar{X}$, 702

H

Hipótesis, 322
alternativa, 322
estadísticas, 321
nula, 322
pruebas de, 322, 323

I

Independencia, 60-62, 64
estadística, 97-99
Inferencia bayesiana, 726
Interacción, 25
Intervalo
bayesiano, 731
de confianza, 273, 274, 285
de la diferencia entre dos medias, 288, 289, 291, 293
de la diferencia entre las dos proporciones, 303
de la razón de dos varianzas, 308
de observaciones pareadas, 296
de proporción, 300
de una muestra, 274-279
de una muestra grande, 280
de varianza, 307
interpretación de, 292
para desviación estándar, 307
para razón de desviación estándar, 309
de predicción, 281, 282, 285
de una observación futura, 282

L
Límites de confianza unilaterales, 277

M
Media(s), 3, 11-13, 19, 28, 29, 107, 108, 110, 111, 229, 232
  cuadradas esperadas
    modelo ANOVA, 556
  cuadráticas, 416
  error cuadrado y, 287
  global, 514
  muestral, 232
  recortada, 12, 13
Mediana, 3, 11-13, 28, 29, 232
Métodos
  de distribución libre, 671
  no paramétricos, 671
    prueba de corridas y, 687
    prueba de Kruskall-Wallis y, 684
    prueba de la suma de rangos de Wilcoxon, 681
    prueba de rango con signo, 676
    prueba de signo, 672
    límites de tolerancia, 690
Metodología
  bayesiana, 269, 725
  de respuesta superficial, 452
Mínimos cuadrados y el modelo ajustado, 394, 395
Moda, 232,728
Modelo
  ANOVA, 561
  de efectos aleatorios, 555
  empírico, 391
  I, experimento del, 555
  II, experimento del, 555
  selección del, 479
    estadístico C_p, 493, 494
    eliminación hacia atrás, 482
    métodos secuenciales, 479
    presión, 490
    regresión progresiva, 483
    selección hacia delante, 482
Momentos
  alrededor del origen, 220
  y funciones generadoras, 219, 220
Muestra, 229, 230, 231
  aleatoria, 229, 231
  desviación estándar de la, 3, 15, 16, 28, 29, 234
  media de la, 3, 11-13, 19, 28, 29, 229, 232
  mediana de la, 3, 11-13, 28, 29, 232
  moda de la, 232
  rango, 15, 28, 29
  varianza, 15, 16, 28, 229, 232, 233
Muestreo
  aleatorio simple, 7
    comprobación de hipótesis, 352
    en estimación de la media, 277
    selección del tamaño de un, 301
  de aceptación, 153
Multicolinealidad, 479

N
Nivel
  de calidad
    aceptable, 721
    rechazable, 721
  de significancia, 324, 325

O
Observaciones pareadas, 294
  de estimación, 294

P
Parámetro de no centralidad, 560
Permutación, 43
  circular, 44
Perspectiva
  bayesiana, 726
  condicional, 726
Población, 2, 4, 229, 230
  varianza de la, 230
Predictor lineal, 501
Presentación de datos, 236
Probabilidad, 31, 48, 49
  condicional, 58-61, 64, 70-72
  subjetiva, 725
  total, 68
Proceso de Poisson, 161
  relación con distribución gamma, 196
Prueba
  de Bartlett, 519
  de Cochran, 521
  de comparaciones múltiples, 527
    de Duncan, 530
    de Dunnett's, 531
    de Turkey, 529
    tasa de error de experimento sabio y, 529
  de la bondad del ajuste, 240, 371, 372
  para la igualdad de diversas varianzas, 519
    de Bartlett, 519
    de Cochran, 521
  para la linealidad de la regresión, 417

pruebas de hipótesis, 270
de dos colas, 332
de dos muestras, 345
de dos varianzas, 367
de homogeneidad, 377
de independencia, 374
de la bondad del ajuste, 372
de una sola cola, 332
elección del tamaño de la muestra, 350, 353
estadística, 328
observaciones pareadas, 347
para varias proporciones, 378
propiedades importantes de una, 331
región crítica, 324
sobre una sola proporción, 361
tamaño de la prueba, 325
una sola muestra (varianza conocida), 338
una sola muestra (varianza desconocida), 342
valor crítico, 324
valores P, 334, 336
varianza muestral, 367
varianzas desconocidas pero diferentes, 347
varianzas desconocidas pero iguales, 346
sobre una sola proporción, 361

R
R^2, 407, 408
R^2 ajustado, 466, 467
Rango, 15, 28, 29
intercuartil, 236, 237
Reglas
aditivas, 52
de Bayes, 68, 70, 71
de eliminación, 68, 70
de multiplicación, 40
multiplicativa, 61
Regresión, 20
progresiva, 481
Regresión lineal
coeficiente de, 392
de determinación, 407
correlación, 432
de dos factores, 577
ecuación normal, 396
error
aleatorio y, 391
falta de ajuste, 419
medias de la respuesta, 394, 409
mínimos cuadrados, 394
modelo
empírico, 391
estadístico, 391
selección, 479, 490
múltiple, 390, 445
ANOVA, 457
coeficiente de determinación, 465
comparación de variables, 461
ecuación normal, 447
inferencia, 458
matriz de varianza-covarianza, 456
matriz TESTADA, 485
multicolinealidad, 479
polinimial, 448
R^2 ajustado, 466
regresión de suma de cuadrados, 463
residuo R de Student, 486
residuo studentizado, 486
suma de cuadrados no corregida, 462
valor extremo, 486
variables ortogonales, 470
predicción de, 409
intervalo, 410, 411
polinomial, 445, 448
prueba para la linealidad de la, 417
que pasa por el origen, 413
residual, 394
simple, 389, 390, 392
sobreajuste, 408
suma
cuadrática, 415
de cuadrados asociada a la media, 464
total de cuadrados, 415
transformación, 426
valores ajustados y, 417
variable
categóricas, 474, 475
dependiente, 389
independientes, 389
Regresión logística, 500
dosis eficaz, 502
razón de probabilidad, 503
Regresión no lineal, 499
control de datos, 500
logística, 500
respuesta binaria, 499
Regresores, 389
Residuo, 394, 429
Respuesta binaria, 499

S
Selección
del modelo, 479
eliminación hacia atrás, 482
estadístico C_p, 493, 494
métodos secuenciales, 479
presión, 490
regresión progresiva, 483
hacia delante, 482
Suma
cuadrática de la regresión, 415
de errores al cuadrado, 415
total de los cuadrados corregida, 407

T
Tabla de análisis de varianza, 416
Tabla de contingencia, 374
frecuencias marginales, 374
Tasa
de error del experimento sabio, 529
de falla, 204
Teorema
de Chebyshev, 131-133
del límite central, 245
Tolerancia
intervalo de, 284, 285
límites de, 283, 284
Transformación, 426
de variables discretas, 212
Tratamiento
efecto negativo del, 575
efecto positivo del, 575

U
Unidad experimental, 289, 295, 574

V
Valor(es)
esperado, 108-111
extremos, 236, 486
-P, 4, 334, 336
Validación cruzada, 490
Variabilidad, 8, 9, 14-16, 116, 131, 232, 261, 263, 264
dentro de las muestras, 264
entre muestras, 264
Variables
aleatorias, 77
continuas, 80
discretas, 79
varianza de, 116, 118
categóricas, 474
o indicadoras, 474
dependiente, 389
Varianza, 15, 16, 28, 115, 116, 118, 229, 232, 233
análisis de, (ANOVA), 264, 415, 511
componentes de la, 557
de la muestra, 233
distribución de razón de, 263

W
Weibull, distribución de, 202, 203