DISTRIBUCIONES MUESTRALES

La necesidad de las distribuciones muestrales

Recordemos que la **estadística** trata de la ***toma de decisiones*** basada en ***datos observados*** en presencia de ***incertidumbre.***

En el análisis de datos se obtienen ***estadísticos*** a fin de estimar los valores correspondientes en la población o ***parámetros*** .

Un ***estadístico*** puede definirse como una ***función de las variables aleatorias que se pueden observar en una muestra*** y de las constantes conocidas y se utilizan para hacer ***inferencias***, estimaciones o tomar decisiones , con respecto a los parámetros poblacionales desconocidos.

Precisamente se conoce como **inferencia estadística** al proceso de ***generalizar*** estos ***resultados muestrales a la población.***

En la práctica se usa la muestra para obtener conclusiones de la población o universo , para ello ***se selecciona al azar una muestra única de tamaño predeterminado***. De este modo, si se quiere emplear algún estadístico para estimar el correspondiente parámetro poblacional, se observar el estadístico para ***toda muestra posible que pudiera haber ocurrido*** y a la ***distribución de los resultados obtenidos para el estadístico*** se le conocerá como una **distribución muestral** .

Un ***estadístico*** resulta una ***variable aleatoria*** donde sus valores dependen de la muestra observada y de su tamaño y tendrá , en consecuencia , una distribución de probabilidad asociada para cada tamaño de la muestra

> La **distribución de probabilidad de un estadístico** se conoce como **distribución muestral**

Distribución Muestral de la Media

Dado que la media aritmética es un estadístico que presenta propiedades matemáticas importantes que justifican su elección entre las demás medidas de tendencia central para estimar la media poblacional , analizaremos el caso de la distribución muestral de la media.

El estadístico **media muestral** , se define :

$$Y = \overline{X} = (X_1 + X_2 + \ldots + X_n)/n = \left(\sum_{i=1}^{n} X_i\right) / n$$

donde X_1 , X_2 , ... X_n son una muestra aleatoria de tamaño n o de n observaciones y son **variables aleatorias en un muestreo repetido.**

$\overline{X}$ es una **variable aleatoria** pues es función de las variables aleatorias observadas en una muestra

Este estadístico se emplea para estimar la **media poblacional o esperanza de una variable aleatoria** X que simbolizamos μ_X , donde $\mu_X = E(X)$

La distribución muestral de un estadístico depende del *tamaño de la población* , del *tamaño de la muestra* y del *método de selección* de estas últimas .

Obtendremos la distribución muestral de la media para el siguiente *ejemplo :*

Para una población de cuatro máquinas ,únicas disponibles en el mercado para la fabricación de tornillos, a las que se sometió a la misma prueba ,se registró el número de piezas defectuosas producidas en la prueba obteniéndose los siguientes resultados:

MAQUINA	Cantidad de piezas defectuosas
A	3
B	2
C	1
D	4

La variable aleatoria **X** es *" cantidad de piezas defectuosas producidas por las máquinas disponibles "*

La variable aleatoria **X** tiene una **distribución uniforme** ya que las únicas cuatro máquinas disponibles producen distinta cantidad de piezas defectuosas y cada una tiene la misma probabilidad de ser seleccionada. La probabilidad de seleccionar una cualquiera de las máquinas es de 1/4 y ésta es, entonces, la probabilidad de seleccionar una máquina que produzca 1 , 2 , 3 ó 4 piezas defectuosas.

$$p_X(x) = \begin{cases} 1/4 & \text{si } x = 1, 2, 3, 4. \\ 0 & \text{en otro caso} \end{cases}$$

Dado que se dispone de información sobre toda la población , calculamos la media , μ_X , y la desviación estándar σ_X :

Esperanza o Media poblacional : $E(X) = \mu_X = 2{,}5$ piezas defectuosas

Varianza : $VAR(X) = \sigma_X^2 = E[(X_i - \mu_X)^2] = 1{,}25$

Desviación Estándar : $\sigma_X = \sqrt{VAR(X)} = 1{,}12$ piezas defectuosas

Si las muestras de dos máquinas se seleccionan con reposición de esta población , hay 16 muestras posibles que se podrían seleccionar y obtendríamos la siguiente tabla de medias muestrales:

TABLA : Total de muestras posibles de n = 2 Máquinas para una población de N = 4 Máquinas cuando se muestrea **con reposición**

Muestra	Máquinas	Resultados	Media Muestral
1	A , A	3 - 3	3,0
2	A , B	3 - 2	2,5
3	A , C	3 - 1	2,0
4	A , D	3 - 4	3,5
5	B - A	2 - 3	2,5
6	B , B	2 - 2	2,0
7	B , C	2 - 1	1,5
8	B , D	2 - 4	3,0
9	C , A	1 - 3	2,0
10	C , B	1 - 2	1,5
11	C , C	1 - 1	1,0
12	C , D	1 - 4	2,5
13	D , A	4 - 3	3,5
14	D , B	4 - 2	3,0
15	D , C	4 - 1	2,5
16	D , D	4 - 4	4,0
TOTAL			40

Resulta una población de tamaño N = 16 para la Variable Aleatoria $\overline{X}$: *" Media muestral para muestras de tamaño n = 2 "* tiene los siguientes parámetros:

Esperanza o Media : $E(\overline{X}) = \mu_{\overline{X}} = (40 / 16) = 2{,}5$

$VAR(\overline{X}) = \sigma_{\overline{X}}^2 = E[(\overline{X}_i - \mu_{\overline{X}})^2] = (10/16) = 0{,}625$

Desviación Estándar : $\sigma_{\overline{X}} = \sqrt{VAR(\overline{X})} = 0{,}79$

Para determinar los valores de la varianza empleamos como elemento auxiliar la tabla de frecuencias siguiente

Tabla de Frecuencias

Media muestral	Frecuencia (fi)	$fi . (\overline{X}_i - \mu_X)^2$
1,0	1	2,25
1,5	2	2,00
2,0	3	0,75
2,5	4	0,00
3,0	3	0,75
3,5	2	2,00
4,0	1	2,25
TOTAL	16	10,00

Podemos representar el **histograma de frecuencias** a partir de esta tabla de frecuencias y tendremos la forma en que se distribuye la variable .

Se observa que :

- El promedio de todas las medias muestrales posibles para muestras de tamaño 2 es igual a la media de la población :

$$\mu_X = \mu_{\overline{X}}$$

- *Las medias muestrales son menos variables que los datos de la población en sí.*

Una media muestral en particular promedia todos los valores de la muestra . La población puede constar de resultados individuales con una amplia escala de valores desde extremadamente pequeños hasta extremadamente grandes . Sin embargo , si un valor extremo cae dentro de la muestra , aunque tenga efecto sobre la media , el efecto se reducirá puesto que se está promediando junto con los demás valores de la muestra. Es más , según aumenta el tamaño de la muestra , el efecto de un solo valor extremo se hace aún más pequeño puesto que se está promediando con más observaciones.

Este fenómeno se expresa en forma estadística en el valor de la desviación estándar de la media muestral que resulta menor al de la población. Es ***la medida de la variabilidad de la media de una muestra a otra*** y se conoce como ***error estándar de la media*** y se simboliza $\sigma_{\overline{X}}$

Muestreo de poblaciones normales

Veremos a continuación como responder la pregunta : Qué distribución seguirá la variable aleatoria $\overline{X}$ donde $\overline{X}$ es *" la media de muestras de tamaño n "* ? .Se demuestra que :

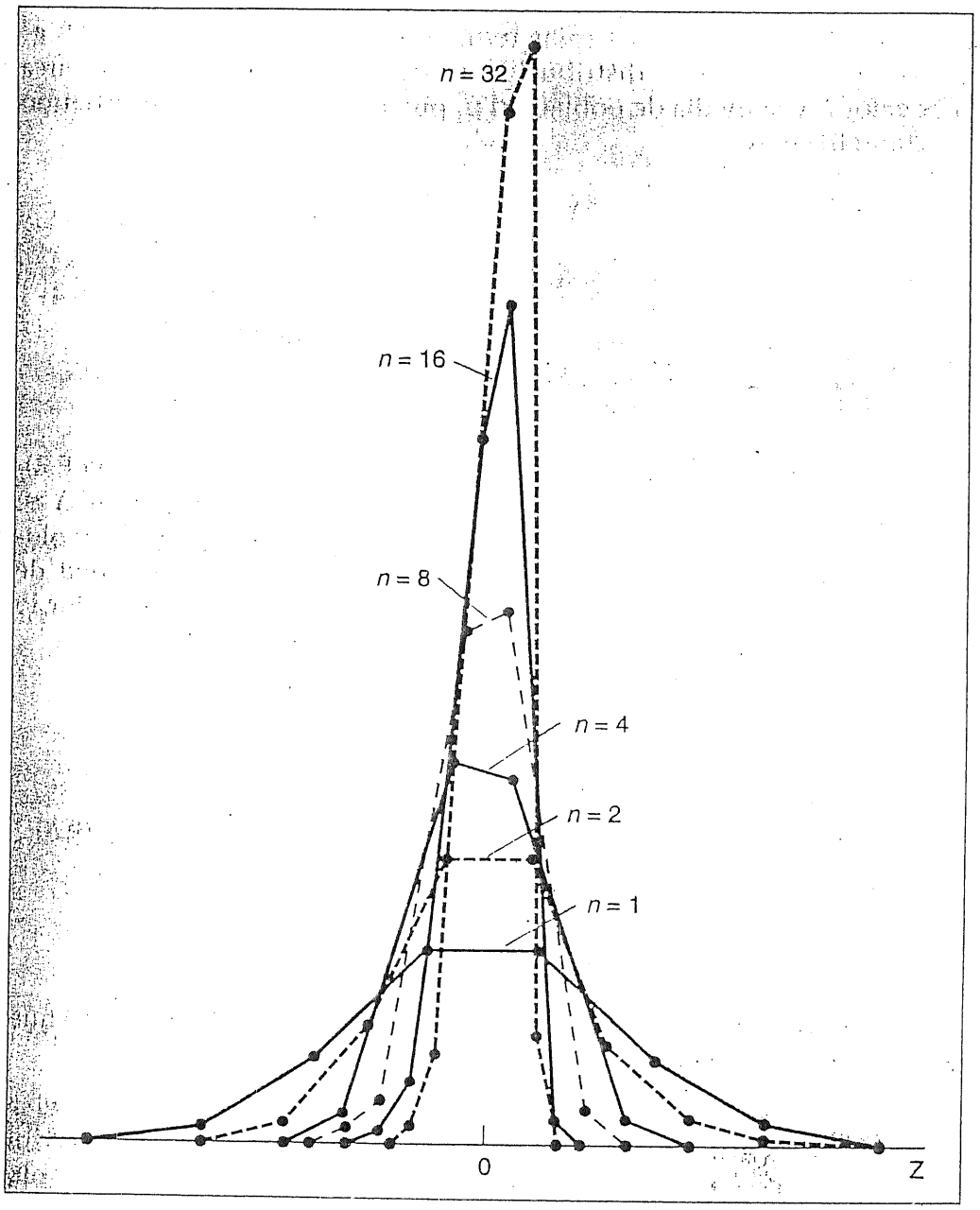
En un *muestreo con reposición de una variable aleatoria que tiene distribución normal de parámetros* μ_X (media poblacional) y σ_X (desviación estándar poblacional) , es decir ;

$$X \sim N(\mu_X, \sigma_X^2)$$

la **distribución muestral de la media** también tendrá **distribución normal** para cualquier tamaño n de la muestra con **media** $\mu_{\overline{X}} = \mu_X$ y **desviación estándar** $\sigma_{\overline{X}} = \sigma_X / \sqrt{n}$; o sea

$$\overline{X} \sim N(\mu_X, \sigma_X^2 / n)$$

Veamos el siguiente resultado, representado en la figura, donde se evidencia que las distribuciones muestrales de medias de una población normal son normales y que además conforme aumenta el tamaño de la muestra , la distribución muestral de la media continúa con una distribución normal con media $\mu_{\bar{x}} = \mu_x$, mientras que disminuye el error estándar de la media por lo que una proporción mayor de medias muestrales están más cercanas a la media.



Distribuciones muestrales de la media de 500 muestras de tamaño
N = 1 , 2 , 3, 8 ,16 y 32 seleccionadas de una población normal

Se seleccionaron aleatoriamente 500 muestras de tamaños 1, 2, 4 , 8 , 16 y 32 de una población que tiene distribución normal .

Trazamos los polígonos de frecuencias para los resultados obtenidos en donde observamos que aunque la distribución muestral de la media es aproximadamente normal para cada tamaño de la

muestra ; las medias muestrales están distribuidas de un modo más ajustado alrededor de la media de la población según aumenta el tamaño de la muestra.

La aproximación a la normal se debe a que se han seleccionado sólo 500 muestras de un número mucho mayor de muestras posibles, resulta entonces que las distribuciones muestrales representadasen el gráfico son aproximaciones de las reales.

Esto significa que **si definimos la variable aleatoria** Z / $Z = (\overline{X} - \mu_X) / (\sigma_X / \sqrt{n})$; Z **tendrá una distribución normal estándar** $Z \sim N(0, 1)$

Esto permitirá obtener, la proporción de todas las medias posibles que pueden encontrarse , por ejemplo , en un intervalo de la media de la población para muestras de tamaño n .Se emplearán para ello las tablas de la función de distribución acumulada para la distribución normal estándar y calcularemos este valor como el área bajo la curva en ese intervalo.

Si observamos el ejemplo resuelto para una población de cuatro máquinas, obtuvimos :

$\mu_{\bar{x}} = \mu_X$ y $\sigma_{\bar{x}} = \sigma_X / \sqrt{n} = 1{,}12 / \sqrt{2} \sim 0{,}79$ y la población no es normal

¿ Podrá generalizarse este resultado a poblaciones no normales ?

Muestreo de poblaciones no normales

Hemos examinado la distribución muestral de la media en el caso de que la variable aleatoria tuviera una distribución normal . Sin embargo, se debe tener en cuenta que en muchos casos se conocerá que la población no está distribuida en forma normal o quizá se piense que resulte poco realista asumir una distribución normal . Por lo tanto es necesario examinar la distribución muestral de la media para poblaciones que no tengan distribución normal.

Veamos los siguientes ejemplos de distribuciones muestrales de medias correspondientes a diferentes poblaciones:

Cada una de las distribuciones muestrales se han obtenido utilizando una computadora para seleccionar 500 muestras diferentes de su respectiva población. Estas muestras se seleccionaron para diferentes tamaños , $n = 2$, 4 , 8 , 16 , 32 ; de tres diferentes distribuciones continuas ; normal, uniforme y exponencial .

- La **FIGURA 1** es un ejemplo de la distribución normal de la media seleccionada de una población normal Si la población tiene una distribución normal , la distribución muestral de la media estará distribuida en forma normal independientemente del tamaño de la muestra . El examen de las distribuciones muestrales de la figura da una evidencia empírica de esta afirmación pues para cada tamaño de muestra estudiado, la distribución muestral de la media tiene una distribución aproximadamente normal.

- La **FIGURA 2** presenta una distribución muestral de la media en base a una población que sigue una distribución continua uniforme (rectangular) . Para muestras de tamaño $n = 1$, cada valor de la población es igualmente probable ; sin embargo , cuando se seleccionan muestras de tamaño $n = 2$ ya hay un efecto de " punto máximo " o de " límite central " en operación. Por lo tanto , en este caso, se pueden observar más valores cercanos a la media de la población que a lo lejos en los extremos , y ; según aumenta el tamaño de la muestra , la distribución muestral de la media se aproxima rapidamente a una distribución normal. Una vez que se cuenta con muestras de por lo menos $n = 8$, la media muestral tendrá una distribución aproximadamente normal.

- La **FIGURA 3** es un ejemplo de la distribución muestral de la media obtenida de una población , con un gran sesgo hacia la derecha, conocida como distribución exponencial. Al aumentar el tamaño de la muestra la distribución muestral presenta menos sesgo. Para $n = 16$, la distribución de la

media tiene un ligero sesgo , mientras que para muestras de tamaño 32 la distribución muestral de la media parece tener distribución normal.

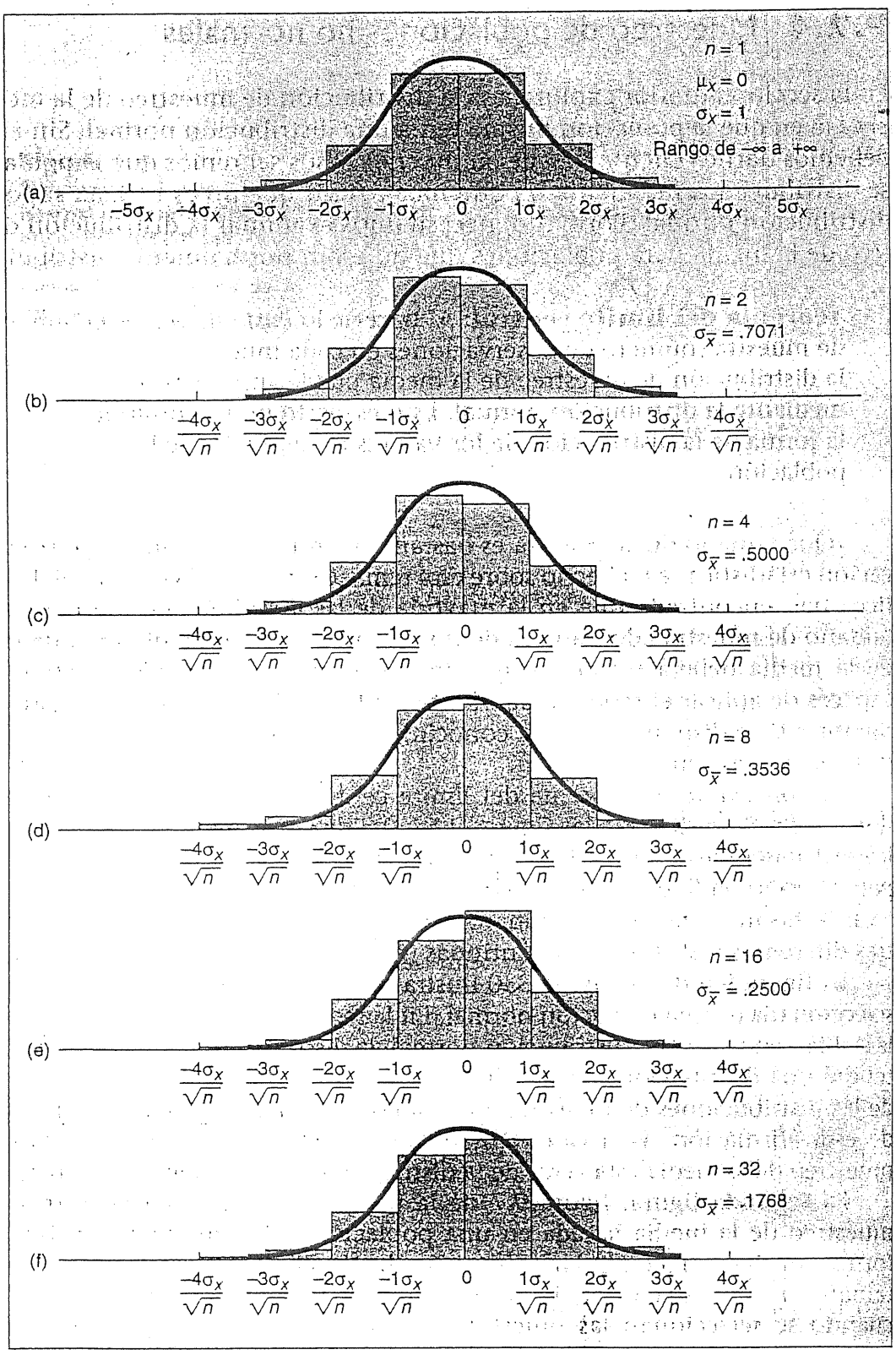


FIGURA 1 :
Distribución normal y distribución muestral de la media de 500 muestras de tamaño n = 2 , 4 , 8 , 16 , 32.

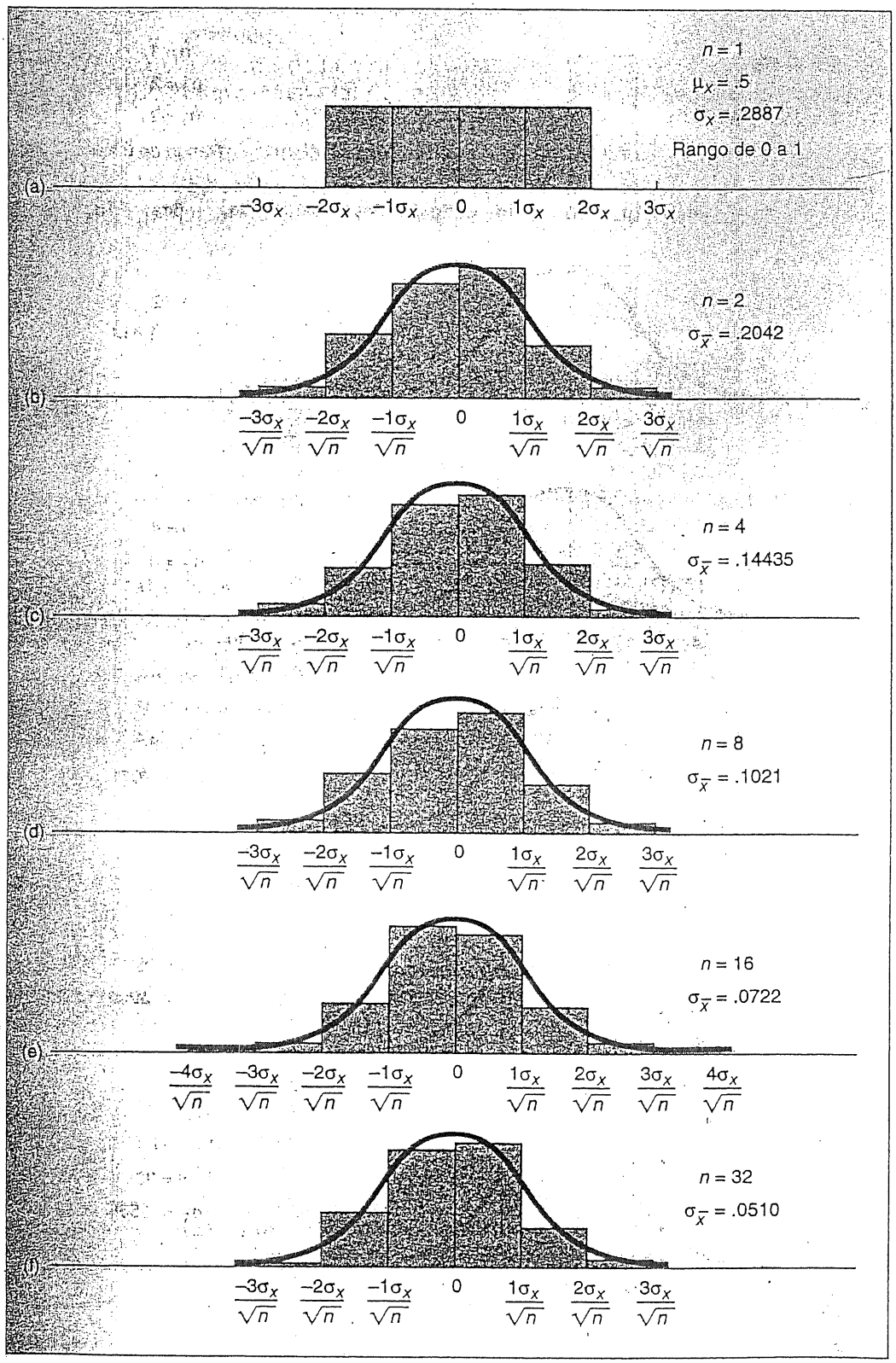


FIGURA 2 :
Distribución continua uniforme (rectangular) y distribución muestral de la media de 500 muestras de tamaño n = 2 , 4 , 8 , 16 , 32.

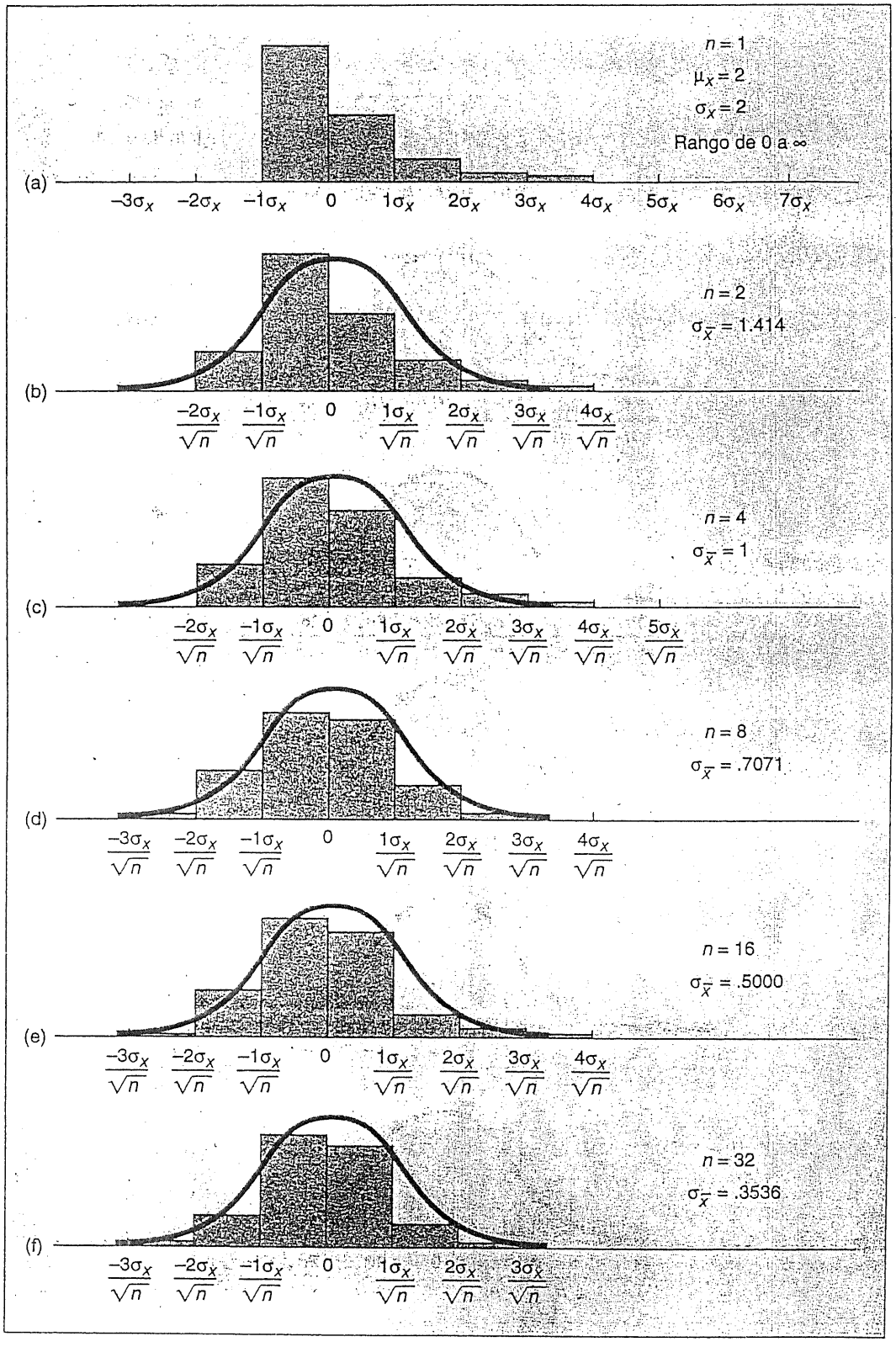


FIGURA 3 :
Distribución exponencial y distribución muestral de la media de 500 muestras de tamaño n = 2 , 4 , 8 , 16 , 32.

Los resultados observados en estos *ejemplos* , se reflejan en el siguiente teorema :

Teorema del límite central

Si $\overline{X}$ es la **media de muestras aleatorias de tamaño** n que se toma de una población con media μ_X y varianza σ_X^2 , entonces la forma límite de la distribución de Z /

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}}$$

se aproxima a la distribución normal estándar cuando n se hace infinito.

Es decir $\overline{X} \sim N (\mu_X , \sigma_X^2 / n)$ **para n suficientemente grande**

En otras palabras , este teorema establece que , ***si el tamaño de la muestra es suficientemente grande se puede aproximar mediante una distribución normal la distribución muestral de la media***. Esto es cierto independientemente de la forma de distribución de los valores individuales de la población.

Como regla general, los estadísticos han encontrado que para la mayor parte de las distribuciones poblacionales , siempre que el tamaño de la muestra sea por lo menos 30 ($n \geq 30$) , la distribución muestral de la media es aproximadamente normal.

De lo expuesto se pueden resumir las siguientes ***conclusiones*** :

- Para la mayor parte de las distribuciones, independientemente de su forma, la distribución muestral de la media tendrá distribución aproximadamente normal si se seleccionan muestras de por lo menos 30 observaciones .
- Si la distribución de la población es bastante simétrica ; la distribución muestral de la media será aproximadamente normal si se seleccionan muestras de por lo menos 15 observaciones.
- Si la población tiene una distribución normal , las distribución muestral de al media será normal independientemente del tamaño de la muestra.

El teorema del límite central permite al investigador hacer inferencias sobre la media de la población sin tener que conocer la forma específica de la distribución de la población para muestras de por lo menos 30 observaciones .

Ejemplo :

Supóngase que el equipo de empaque en un proceso de llenado de paquetes de cereal de 368 gr se ajusta en forma tal que la cantidad de cereal en la caja tiene distribución normal con una media de 368 gr y se sabe , por experiencia , que la desviación estándar de la población es de 15 gr.

Si de varios millares de cajas que se llenan en un día se seleccionan al azar 25 cajas y se calcula el peso promedio para esta muestra , qué resultado se puede esperar ?

La población tiene una distribución normal, la media de la muestra actúa como una representación en miniatura de la población y tiene una buena posibilidad de encontrarse cerca de los 368 gr. Entonces surge la pregunta :

- ¿ Cuál es la probabilidad de que una muestra de 25 cajas tenga una media entre 365 y 368 gr. ?

La variable aleatoria es $\overline{X}$: *"peso promedio de muestras de tamaño 25 "*. El problema es obtener

$$P\ (\ 365\ gr\ <\ \overline{X}\ <\ 368\ gr\ )$$

Dado que $X \sim N\ (\ \mu_X = 368\ gr\ ,\ \sigma_X = 15\ gr\ )$ donde **X**: *"peso de la caja con cereal "*

resulta $\overline{X} \sim N\ (\ \mu_{\overline{X}} = 368\ gr\ ,\ \sigma_{\overline{X}} = (15\ gr / \sqrt{25}) = 3gr\ )$, entonces :

$P\ (\ 365\ gr < \overline{X} < 368\ gr) = P\ (\ -1 < Z < 0\ ) = 0,5 - 0,1587 = 0,3413$

Donde trabajamos con la variable aleatoria $Z\ /\ Z \sim N\ (\ \mu_Z = 0\ ,\ \sigma_Z^2 = 1)$ estandarizando la variable aleatoria X , entonces :

$Z = (\overline{X} - \mu_{\overline{X}}) / \sigma_{\overline{X}} = (\overline{X} - 368) / 3$ y si $x = 365\ gr \Rightarrow z = (365 - 368) / 3 = -1$

si $x = 368\ gr \Rightarrow z = (368 - 368) / 3 = 0$

y luego calculamos la probabilidad empleando la tabla de distribución acumulada para Z .

Obtuvimos que $P\ (\ 365\ gr < \overline{X} < 368\ gr\ ) = 0,3413$.

- Este resultado significa que el 34,13 % de todas las muestras posibles de tamaño 25 tendrían una media muestral entre 365 y 368 gr que no es lo mismo que decir que cierto porcentaje de cajas individuales tendrán entre 365 y 368 gr.

Si queremos calcular esto último hacemos :

$P\ (\ 365\ gr < X < 368\ gr\ ) = P\ (\ -0,20 < Z < 0\ ) = 0,5 - 0,4207 = 0,0793$

pues $Z = (\ X - \mu_X\ ) / \sigma_X = (\ X - 368\ ) / 15$ y si $x = 365\ gr \Rightarrow z = (\ 365 - 368)/15 = -0,20$

si $x = 368\ gr \Rightarrow z = 0$

lo que permitió obtener las probabilidades empleando la tabla de distribución acumulada para la normal estándar.

El resultado obtenido significa que se espera que el 7,93 % de las cajas individuales contengan entre 365 y 368 gr.

Al comparar estos resultados se observa que se encuentran muchas más medias muestrales que medias individuales entre 365 y 368 gr ; es decir la posibilidad de que la media de una muestra de 25 cajas esté cerca de la media de la población es mayor que la posibilidad de que lo esté un valor individual único.

- Veamos cómo se afectarían los resultados si se utilizara un tamaño de muestra diferente, por ejemplo, 100 cajas en vez de 25 .

En este caso $\overline{X} \sim N\ (\ \mu_{\overline{X}} = 368\ gr\ ,\ \sigma_{\overline{X}} = (15\ gr / \sqrt{100}\ ) = 1,5\ gr\ )$ y

$P\ (\ 365\ gr < \overline{X} < 368\ gr\ ) = P(\ -2 < Z < 0\ ) = 0,5 - 0,0228 = 0,4772$

Se puede esperar que el 47,72% de las muestras de tamaño 100 tengan medias entre 365 y 368 gr.

- Por último , en lugar de determinar la proporción de medias muestrales que se esperan que se encuentren en cierto intervalo , podemos encontrar el intervalo en el cual caería una proporción fija de medias muestrales.

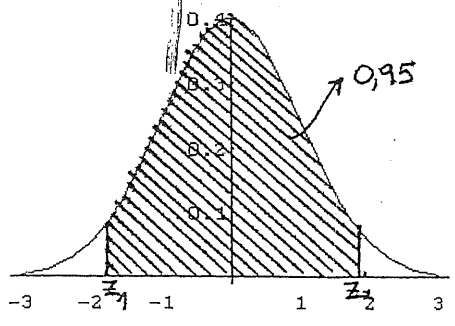
Sea conocer el intervalo en torno a la media de la población que incluye el 95 % de las medias muestrales en base a una muestra de 25 cajas.

Esto significa que $P\ (\ x_1 < \overline{X} < x_2\ ) = 0,95$ donde $\overline{X} \sim N\ (\ \mu_{\overline{X}} = 368\ gr\ ,\ \sigma_{\overline{X}} = 3\ gr\ )$

Resulta $Z = (\ \overline{X} - 368\ ) / 3$ una variable normal estándar para la que

$$z_1 = (x_1 - 368)/3 \quad y \quad z_2 = (x_2 - 368)/3$$

$$P(z_1 < Z < z_2) = 0.95 = P(Z < z_2) - P(Z < z_1) = 0{,}975 - 0{,}025$$



El área entre z_1 y z_2 es de 0,95

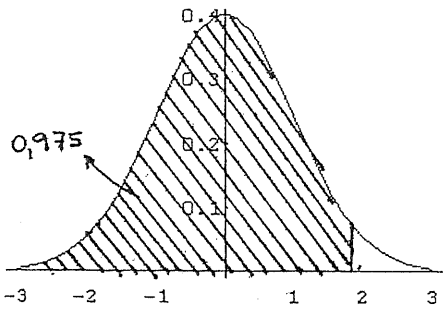
En la tabla de distribución acumulada de la normal estándar, el valor z_2 que deja un área de 0,975 a su izquierda es

$z_2 = 1{,}96$, $\phi(1{,}96) = P(Z < 1{,}96) = 0{,}975$

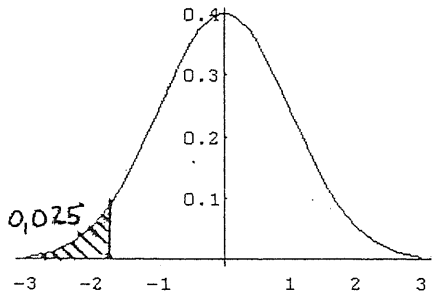
El valor z_1 que deja un área de 0,025 a su izquierda es

$z_1 = -1{,}96$, $\phi(-1{,}96) = P(Z < -1{,}96) = 0{,}025$

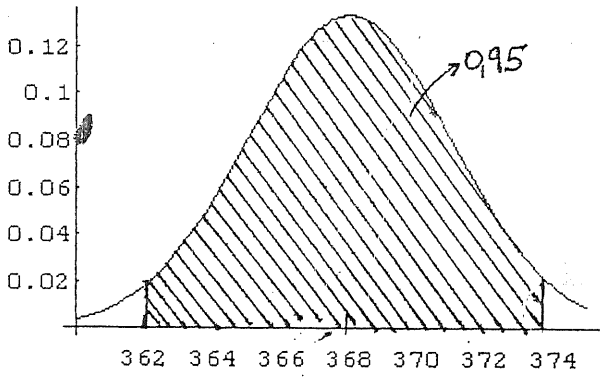
$P(-1{,}96 < Z < 1{,}96) = 0{,}975 - 0.025 = 0{,}95$



$\phi(1{,}96) = P(Z < 1{,}96) = 0{,}975$



$\phi(-1{,}96) = P(Z < -1{,}96) = 0{,}025$



Los valores buscados de x_1 y x_2 se obtienen de las ecuaciones correspondientes a z_1 y z_2 y resultan

$x_1 = 368 - 1{,}96.3 = 362{,}12$

$x_2 = 368 + 1{,}96.3 = 373{,}88$

Entonces el 95 % de las medias muestrales de muestras de tamaño 25 se encuentran en el intervalo $362{,}12 \text{ gr} < \overline{X} < 373{,}88 \text{ gr}$.

Esto me puede servir , por ejemplo , para decidir si el equipo funciona correctamente. Si la media de una muestra de tamaño 25 se encuentra en este intervalo decimos que el funcionamiento es correcto , de lo contrario debo efectuar el ajuste correspondiente.

Muestreo de poblaciones finitas

Cuando el muestreo se realiza con reposición o la población es infinita valen los resultados que hemos presentado. Cuando el muestreo se realiza **sin reposición de una población de tamaño finito** N y en especial cuando **el tamaño de la muestra n no es pequeño** en comparación con el tamaño de la población N ($n / N > 0{,}05$, es decir se muestrea más del 5% de la población) se debe usar un **factor de corrección para población finita** al definir el error estándar de la media . Resulta

$\sigma_{\overline{x}} = (\sigma_X / \sqrt{n}) \sqrt{(N - n)/(N - 1)}$ $\quad k = \sqrt{(N - n)/(N - 1)}$ es el factor de corrección población finita

Distribución Muestral de la Diferencia de Medias

Supóngase que se tienen dos poblaciones , una con media μ_{x1} y varianza σ_{x1}^2 y la segunda con media μ_{x2} y varianza σ_{x2}^2 . Si el estadístico $\bar{X}_1$ representa la media de una muestra aleatoria de tamaño n_1 seleccionada de la primera población y el estadístico $\bar{X}_2$ representa la media de una muestra aleatoria de tamaño n_2 seleccionada de la segunda población, independientemente de la muestra de la primera población; entonces la distribución muestral de la diferencia de medias $\bar{X}_1 - \bar{X}_2$ para muestras repetidas de tamaños n_1 y n_2 seleccionadas independientemente de dos poblaciones responde al siguiente teorema:

Si se sacan al azar muestras independientes de tamaños n_1 y n_2 de dos poblaciones, discretas o continuas, con medias μ_{x1} y μ_{x2} y varianzas σ_{x1}^2 y σ_{x2}^2 respectivamente; entonces, la distribución muestral de la diferencia de medias $\bar{X}_1 - \bar{X}_2$, está distribuida aproximadamente en forma normal con media y varianzas :

$$\mu_{x1-x2} = \mu_{x1} - \mu_{x2} \qquad \sigma^2_{x1-x2} = (\sigma_{x1}^2 / n_1) + (\sigma_{x2}^2 / n_2)$$

De modo que $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_{x1} - \mu_{x2})}{\sqrt{(\sigma_{x1}^2 / n_1) + (\sigma_{x2}^2 / n_2)}}$$ es aproximadamente una variable normal estándar

Si $n_1 \geq 30$ y $n_2 \geq 30$ la aproximación normal de $\bar{X}_1 - \bar{X}_2$ es muy buena , sin importar las formas de las dos poblaciones.

Si ambas poblaciones son normales, entonces $\bar{X}_1 - \bar{X}_2$ tiene una distribución normal sin importar qué valores tengan n_1 y n_2

Ejemplo:
Los cinescopios de televisión del fabricante A tiene una duración promedio de 6, 5 años y una desviación estándar de 9 años , mientras que los del fabricante B tienen una vida promedio de 6,0 años con una desviación estándar de 0,8 años. ¿Cuál es la probabilidad de que una muestra aleatoria de 36 cinescopios del fabricante A tenga una duración promedio que sea al menos un año más que la duración promedio de una muestra de 49 cinescopios del fabricante B?

La información proporcionada por el problema es la siguiente:

Población 1 Duración de los cinescopios de A	Población 2 Duración de los cinescopios de B
$\mu_{x1} = 6,5$	$\mu_{x2} = 6,0$
$\sigma_{x1}^2 = 0,9^2$	$\sigma_{x2}^2 = 0,8^2$
$n_1 = 36$	$n_2 = 49$

$$\mu_{x1-x2} = 6,5 - 6 = 0,5 \qquad \sigma_{x1-x2} = \sqrt{(0,9^2/36) + (0,8^2/49)} = 0,189$$

La variable aleatoria $\bar{X}_1 - \bar{X}_2$ tiene una distribución normal con media $\mu_{x1-x2} = 0,5$ y desviación estándar $\sigma_{x1-x2} = 0,189$, entonces la variable $Z = [(\bar{X}_1 - \bar{X}_2) - 0,5] / 0,189$ es normal estándar y podemos calcular probabilidades empleando la función de distribución acumulada de la normal estándar:

$$P(\bar{X}_1 - \bar{X}_2 \geq \bar{x}_1 - \bar{x}_2) = P(\bar{X}_1 - \bar{X}_2 \geq 1) = 1 - P(\bar{X}_1 - \bar{X}_2 < 1) =$$

$$P(\bar{X}_1 - \bar{X}_2 \geq 1) = 1 - P(Z < (1 - 0,5)/0,189) = 1 - P(Z < 2,65) = 1 - 0,9960 = 0,0040$$

La probabilidad de que la media de 36 cinescopios del fabricante A sea al menos 1 año más grande que la media de 49 cinescopios del fabricante B es del 0,4 %

Distribución Muestral de $(n-1)\ S^2/\sigma^2$

Si se toma una muestra aleatoria de tamaño n de una población normal con media μ_X y varianza σ_X^2 y se calcula la varianza muestral s^2 se obtiene un valor del estadístico S^2. La variable aleatoria S^2 es la varianza de todas y cada una de las muestras aleatorias de tamaño n que se pueden tomar de la población normal. Para calcular probabilidades relacionadas con S^2 se trabaja con la distribución muestral del estadístico $(n-1)\ S^2/\sigma^2$ que tiene una función de densidad conocida como ji cuadrado. A esta variable aleatoria se la simboliza χ^2. Así, se enuncia el siguiente teorema :

Si S^2 es la varianza de muestras aleatorias de tamaño n tomadas de una población normal que tiene varianza σ^2, entonces el estadístico

$\chi^2 = (n-1)\ S^2/\sigma^2$ tiene una distribución ji cuadrado con $\nu = n-1$ grados de libertad

Decimos que una variable aleatoria continua X tiene una distribución ji cuadrada con ν grados de libertad si su función de densidad es

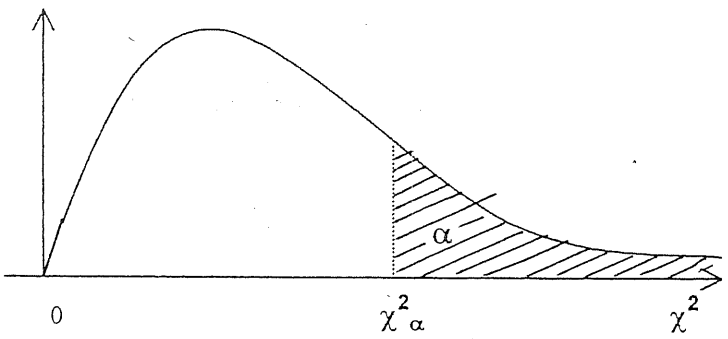
$$f_X(x) = \begin{cases} (x^{(\nu/2)-1}\ e^{-x/2}) / 2^{\nu/2}\ \Gamma(\nu/2) & x > 0 \\ 0 & \text{en otro caso} \end{cases}$$

donde ν es un entero positivo

Los valores de la variable aleatoria χ^2, se calculan a partir de cada muestra $\chi^2 = (n-1)\ s^2/\sigma^2$, $\nu = n-1$

La probabilidad de que una muestra aleatoria produzca un valor χ^2 más grande que algún valor especificado es igual al área bajo la curva a la derecha de ese valor.

Así, $P(\chi^2 > \chi^2_\alpha) = \alpha$, donde χ^2_α representa el valor de χ^2 arriba del cual se encuentra un área α. Esta situación se representa en la figura:



Buscaremos en una Tabla de doble entrada los valores de χ^2_α. En la tabla se entra con valores de ν y de α.

Por ejemplo: para $\nu = 7$ y $\alpha = 0{,}05$ se obtiene $\chi^2_{0,05} = 14.067$

para $\nu = 7$ y $\alpha = 0{,}95$ se obtiene $\chi^2_{0,95} = 2{,}167$

El 90 % de la distribución ji cuadrado cae entre 2,167 y 14,067 , entonces

$$P(2{,}167 < \chi^2 < 14{,}067) = 0{,}90 \quad \text{si} \quad \nu = 7$$

El 95 % de la distribución estará entre $\chi^2_{0,975}$, el valor que deja un área de 0,975 a su derecha , y $\chi^2_{0,025}$, el valor que deja un área de 0,025 a su derecha . En términos de probabilidades resulta :

$$P (\chi^2_{0,975} < \chi^2 < \chi^2_{0,025}) = 0,95$$

Si a partir de una muestra, obtenemos un valor de ji cuadrado a la derecha de $\chi^2_{0,025}$, debido a la poca probabilidad que tenemos de obtener un valor en esa región se puede pensar que el valor supuesto para σ^2 es demasiado pequeño. En este caso $\chi^2 = (n - 1)\, s^2/\sigma^2$ es un valor demasiado grande.

Si, en cambio, obtenemos un valor de ji cuadrado a la izquierda de $\chi^2_{0,975}$, debido a la poca probabilidad que tenemos de obtener un valor en esa región se puede pensar que el valor supuesto para σ^2 es demasiado grande. En este caso $\chi^2 = (n - 1)\, s^2/\sigma^2$ es un valor demasiado pequeño.

Ejemplo:
Un fabricante de baterías para automóviles garantiza que sus batería durarán , en promedio, 3 años con una desviación estándar de 1 año . Si 5 de estas batería tienen duraciones de 1,9 - 2,4 - 3,0 - 3,5 y 4 , 2 años ¿ está el fabricante convencido aún que sus baterías tienen una desviación estándar de un año ?
La variable aleatoria duración de las baterías se distribuye normalmente.
La varianza muestral es $s^2 = 0,815$, $\nu = 4$ (el tamaño de la muestra es 5) y $\sigma^2 = 1^2 = 1$, entonces $\chi^2 = (n - 1)\, s^2/\sigma^2 = 4 \,.\, 0,815 / 1 = 3,26$

Dado que el 95 % de los valores de χ^2 con 4 grados de libertad cae entre $\chi^2_{0,975} = 0,484$ y $\chi^2_{0,005} = 11,143$, el valor calculado con $\sigma^2 = 1$ está en ese intervalo y por lo tanto es fabricante no tiene razón para sospechar que la desviación estándar sea distinta de 1 .

Si buscamos en la Tabla con $\nu = 4$, el valor de χ^2 más cercano a 3,26 es $\chi^2_{0,50} = 3,357$ que deja un área de 0,5 a su derecha , lo que llevaría a la misma conclusión .

Distribución del estadístico T

La mayoría de las veces no se conoce la varianza σ^2 de la población de la que se seleccionan las muestras aleatorias. Para muestras de tamaño $n \geq 30$ se obtiene una buena aproximación de σ^2 cuando se calcula S^2. Si en la variable aleatoria $Z = (\overline{X} - \mu_X) / (\sigma / \sqrt{n})$, se reemplaza σ por S se define una nueva variable aleatoria , el estadístico $T = (\overline{X} - \mu_X) / (S / \sqrt{n})$. La distribución de probabilidades de T se desvía en forma apreciable de la normal estándar si el tamaño de la muestra es pequeño, porque en este caso los valores de S^2 fluctúan considerablemente de una muestra a otra. La función de densidad del estadístico T recibe el nombre de distribución **t de Student con ν=n-1 grados de libertad** y está dada por :

$$h(t) = \left(\Gamma ((\nu+1)/2) \,/\, \Gamma (\nu/2) \sqrt{\pi \nu} \right) \left(1 + (t^2 / \nu) \right)^{-(\nu+1)/2} \qquad -\infty < t < \infty$$

Se considera que las muestras se extraen de una **población normal.** Sin embargo, puede demostrarse que puede emplearse para las poblaciones no normales que poseen distribuciones en forma de campana.

Al igual que la variable aleatoria Z , T tiene una distribución simétrica alrededor de una media igual a cero con forma de campana. Sin embargo, la distribución t varía más , debido a que los valores de T dependen de las variaciones de dos cantidades $\overline{X}$ y S^2 , mientras que los valores z dependen sólo de los cambios de $\overline{X}$ de una muestra a otra.

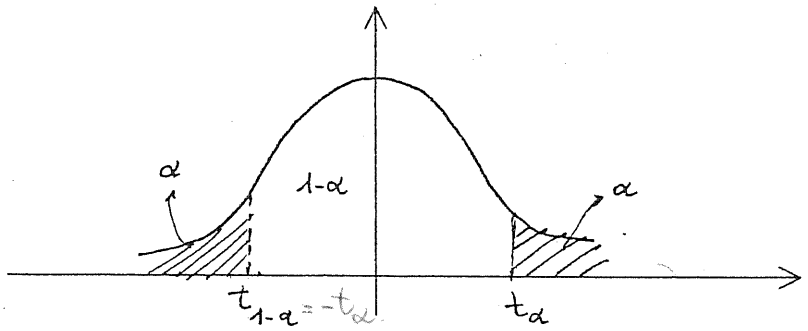
La varianza de T depende del tamaño de la muestra y siempre es mayor que 1.
A medida que el tamaño de la muestra aumenta las dos distribuciones se aproximan.

Las probabilidades de que una muestra aleatoria produzca un valor $t = (x - \mu_X) / (s / \sqrt{n})$ que caiga entre dos valores cualesquiera especificados es igual al área bajo la curva de la distribución t entre esos valores. Tenemos una distribución diferente para cada tamaño n de las muestras ($n \geq 30$).

Para determinar probabilidades se pueden emplear tablas que contienen los valores de t por encima del cual se encuentra un área específica α para cada grado de libertad $\nu=n-1$. Los valores de α que presenta la tabla son 0,1 - 0,05 - 0,025 - 0,01 - 0,005 . (al menos)

De este modo t_α representa el valor de t arriba del cual se encuentra un área igual a α.

Para $\nu = 10$, $t_{0,025} = 2,228$ es el valor de t arriba del cual se encuentra un área $\alpha = 0,025$.



Dado que la distribución t es simétrica respecto del eje de ordenadas, que coincide con la media de la distribución, se tiene que $t_{1-\alpha} = - t_\alpha$.

Esto significa que el valor de t que tiene un área de $1 - \alpha$ a la derecha , o α a la izquierda es opuesto al valor de t que deja un área igual a α hacia la cola derecha de la distribución. Así, resulta:

$$t_{0,95} = - t_{0,05} \qquad t_{0,99} = - t_{0,01}$$

Ejemplos

- Obtener el valor de t con $\nu = 14$ grados de libertad que tiene un área de 0,025 a su izquierda

Debo calcular $t_{0,975} = - t_{0,025} = - 2,145$

El valor de t que tiene un área de 0,025 a su izquierda es el que deja un área de 0,975 a su derecha y este valor es opuesto al valor de t que deja un área de 0,025 a su derecha debido a la simetría de la distribución t.

- La $P (- t_{0,025} < T < t_{0,05}) = P (t_{0,975} < T < t_{0,05}) = 0,95$

- Encontrar el valor de K de tal forma que la $P (K < T < -1,761) = 0.045$ para una muestra aleatoria de tamaño 15 seleccionada de una población normal.

De acuerdo a la tabla para $\nu = 14$, $t_{0,05} = 1,761$ es el valor de t que deja un área de 0,05 a su derecha , entonces $t_{0,95} = - 1,761$ deja un área de 0,95 a su derecha y de 0,05 a su izquierda y resulta $P (T < -1,761) = 0.05$

Como $P (K < T < -1,761) = P (T < -1,761) - P (T < K) = 0,05 - P (T < K) = 0,045$ se obtiene que $P (T < K) = 0,005$.

K es el valor de t que deja un área de 0,005 a su izquierda o de 0,995 a su derecha es decir

$t_{0,995} = - t_{0,005} = -2,977$

Se obtuvo que P (-2,977 < T < -1,761) = 0.045

Podemos decir que exactamente el 95 % de los valores de t de una distribución t con ν grados de libertad caen entre $-t_{0,025}$ y $t_{0,025}$. Un valor de t que caiga debajo de $-t_{0,025}$ o arriba de $t_{0,025}$ provocaría que se pensara que el valor de μ_X puede ser un error.

Ejemplo: $t > t_{0,025}$ $(\bar{x} - \mu_x) > 0$ y grande, entonces μ_x pequeño.
$t < t_{-0,025}$ $(\bar{x} - \mu_x) < 0$ y grande en v. abs., entonces μ_x grande

Un fabricante de focos afirma que su producto durará en promedio 500 hs de trabajo. Para conservar este promedio, esta persona verifica 25 focos cada mes. Si el valor de t calculado cae entre $-t_{0,05}$ y $t_{0,05}$ el fabricante mantiene esta afirmación ¿ Qué conclusión se obtiene de una muestra con x = 518 hs y desviación estándar s = 40 hs . Asuma que los tiempos de vida se distribuyen en forma normal.

Para $\nu = 24$, obtenemos : $t_{0,05} = 1,711$, $- t_{0,05} = t_{0,95} = - 1,711$, P (-1,711< T < 1,711) = 0,90

$t = (x - \mu_X) / (s / \sqrt{n}) = (518 - 500) / (40/5) = 2,25$

El valor de t calculado a partir de la muestra es superior a 1,711. La probabilidad de obtener un valor de $t \geq 2,25$ es menor que 0,05 y también es menor que 0,025 ($t_{0,025} = 2,064$ para $\nu = 24$) .

Si μ_X fuera mayor que **500**, el valor de t obtenido estaría más cercano a $t_{0,05}$. De aquí que el fabricante está en condiciones de concluir que sus focos son un producto mejor de lo que había pensado.

Distribución F

Si S_1^2 y S_2^2 son las varianzas muestrales de **variables aleatorias independientes** para muestras de tamaño n_1 y n_2 que se extraen de **poblaciones normales** con varianzas σ_1^2 y σ_2^2 entonces la variable aleatoria F

$$F = (S_1^2 / \sigma_1^2) / (S_2^2 / \sigma_2^2) = (S_1^2 / S_2^2) . (\sigma_2^2 / \sigma_1^2)$$

tiene una distribución F con $\nu_1 = n_1 - 1$ y $\nu_2 = n_2 - 1$ grados de libertad.

La función de densidad de f está dada por

$$f_F(f) = \begin{cases} [\Gamma((\nu_1 + \nu_2)/2) (\nu_1 / \nu_2)^{(\nu1/2)} / \Gamma(\nu_1/2) \Gamma(\nu_2/2)] [f^{(\nu1/2)-1} . (1 + (\nu_1 / \nu_2) f)^{-(\nu1+\nu2)/2}] & 0 < f < \infty \\ 0 \quad \text{en otro caso} \end{cases}$$

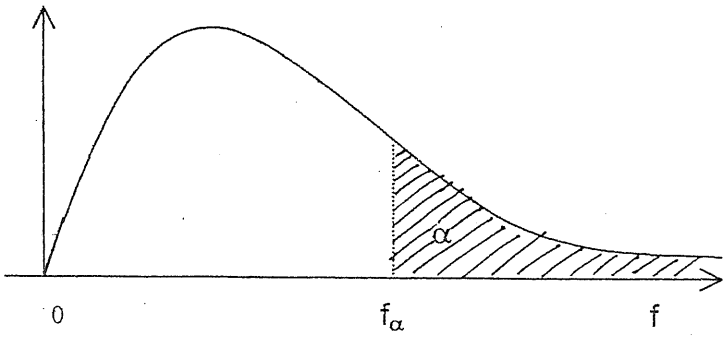
conocida como distribución F con ν_1 , ν_2 grados de libertad.

El número de grados de libertad ν_1 asociados con la variable aleatoria de la población que aparece en el numerador siempre se establece en primer término , después va el números de grados de libertad asociado a la variable aleatoria de la población que aparece en el denominador.

La curva de la distribución de probabilidades de F depende no sólo de ν_1 y ν_2 sino también del orden en el que se establecen . Una vez que se dan estos dos valores puede identificarse la curva.

Las tablas que emplearemos para calcular probabilidades relacionadas con la variable aleatoria F dan el valor de f por encima del cual se encuentra un área igual a α para ν_1 y ν_2 grados de libertad, dados en ese orden. Representamos ese valor como f_α

Esta situación se representa en la figura:



Disponemos de dos tablas de doble entrada para trabajar con F. Una para valores de $\alpha = 0,05$ y la otra para valores de $\alpha = 0,01$ para varias combinaciones de grados de libertad ν_1 y ν_2 .

De este modo se obtiene para $\nu_1 = 6$ y $\nu_2 = 10$ que $f_{0,05} = 3,22$

Propiedad :

Dada $f_\alpha(\nu_1, \nu_2)$ para f_α con ν_1 y ν_2 grados de libertad, se obtiene $f_\alpha(\nu_1, \nu_2) = 1 / f_\alpha(\nu_2, \nu_1)$

De este modo $f_{0,95}(6, 10) = 1 / f_{0,05}(10, 6) = 1 / 4,06 = 0,246$ si $\alpha = 0,05$, $\nu_1 = 6$, $\nu_2 = 10$

Ejemplos :

1) Si S_1^2 y S_2^2 son las varianzas muestrales de **variables aleatorias normales** para muestras independientes de tamaño $n_1 = 8$ y $n_2 = 12$ con varianzas iguales, $\sigma_1^2 = \sigma_2^2$, encuentre $P((S_1^2 / S_2^2) < 4, 89)$

En este caso, la variable aleatoria F resulta :

$$F = (S_1^2 / S_2^2) \cdot (\sigma_2^2 / \sigma_1^2) = (S_1^2 / S_2^2) \quad \text{pues } \sigma_1^2 = \sigma_2^2 \text{ y } (\sigma_2^2 / \sigma_1^2) = 1$$

con $\nu_1 = 7$ y $\nu_2 = 11$ grados de libertad.

En la tabla correspondiente a $\alpha = 0,01$, para el par ordenado $(\nu_1, \nu_2) = (7, 11)$, obtenemos :

$$f_{0,01}(7, 11) = 4,89$$

que significa que el valor 4,89 deja un área de 0,01 a su derecha y de 0,99 a su izquierda

Entonces, la $P((S_1^2 / S_2^2) < 4, 89) = 0,99$ pues corresponde al área a la izquierda de $f_{0,01}(7, 11)$

2) Si S_1^2 y S_2^2 son las varianzas muestrales de variables aleatorias normales con varianzas $\sigma_1^2 = 10$ y $\sigma_2^2 = 15$ para las que se tomaron muestras independientes de tamaños $n_1 = 25$ y $n_2 = 31$, encuentre $P((S_1^2 / S_2^2) > 1,26)$

La variable aleatoria F resulta :

$$F = (S_1^2 / S_2^2) \cdot (\sigma_2^2 / \sigma_1^2) = 1,5 (S_1^2 / S_2^2) \quad \text{pues} \quad (\sigma_2^2 / \sigma_1^2) = 1,5$$

con $\nu_1 = 24$ y $\nu_2 = 30$ grados de libertad.

Como $P((S_1^2 / S_2^2) > 1,26) = P((1,5 (S_1^2 / S_2^2)) > 1,89) = P(f_\alpha (24, 11) > 1,89) = 0,05$

y en la tabla correspondiente a $\alpha = 0,05$, para el par ordenado $(\nu_1, \nu_2) = (24, 11)$, obtenemos :

$$f_{0,05}(24, 11) = 1,89$$

que significa que el valor 1,89 deja un área de 0,05 a su derecha ; entonces,

$$P(f_\alpha (24, 11) > 1,89) = P(f_{0,05} (24, 11) > 1,89) = 0,05$$