ESTIMACIÓN

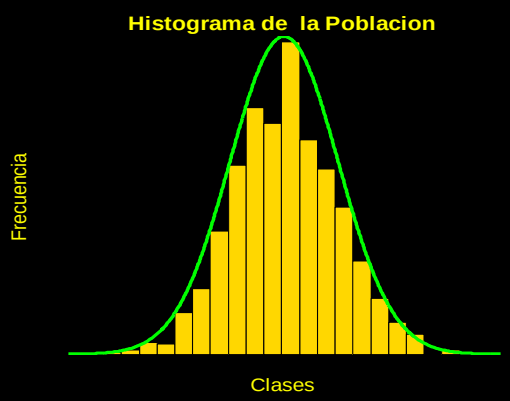
Ing. Adriana M. Apaza

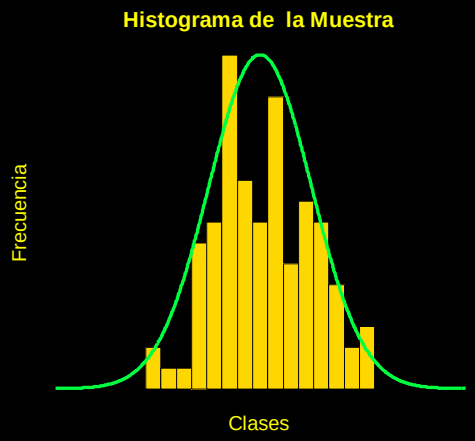
Parámetros poblacionales y Estadísticos Muestrales

Datos
(Población de Interés)

Muestreo

Muestras





Parámetros:

Media (μ)

Varianza(σ^2)

Desv. Est. (σ)

Etc.

Inferencias

Estadísticos:

Promedio ($\bar{x}$)

Varianza muestral(S^2)

Desv. Est. muestral(S)

Etc.

Inferencia estadística

La inferencia estadística puede dividirse en dos áreas principales

- Estimación de Parámetros
- Prueba de Hipótesis

ESTIMACIÓN DE PARÁMETROS

- Consiste en la obtención de valores aproximados para las características desconocidas (parámetros) de la distribución de la población.

- Tipos de estimación:
  - Puntual: un valor.
  - Por intervalos: un intervalo con garantías de contener al parámetro.

Sea θ un parámetro poblacional cuyo valor se desea conocer a partir de una muestra.

Sea $\hat{\Theta}$ un **estadístico** (función de la muestra) que utilizamos para estimar el valor de θ , y lo llamaremos **estimador**.

Una estimación puntual de algún parámetro poblacional θ es un valor único $\hat{\theta}$ del estadístico $\hat{\Theta}$.

Por ejemplo, cuando obtenemos una media aritmética a partir de una muestra, tal valor puede ser empleado como un estimador para el valor de la media poblacional.

PROPIEDADES DE LOS ESTIMADORES

Estimador **insesgado**. Diremos que $\widehat{\theta}$ es un estimador insesgado de θ si:

$$E\left(\widehat{\Theta}\right) = \theta$$

Vimos que la **media muestral** es un **estimador insesgado** de la **media poblacional**.

La **varianza muestral** S^2 es un estimador insesgado de la **varianza poblacional.**

La desviación estándar muestral S es un estimador sesgado de σ_X, con la tendencia a hacer el sesgo insignificante en muestras grandes.

- Sean $\hat{\Theta}_1$ y $\hat{\Theta}_2$ dos estimadores insesgados del parámetro θ.

Si $\sigma_{\hat{\theta}_1} < \sigma_{\hat{\theta}_2}$ decimos que $\hat{\Theta}_1$ es más **eficiente** que $\hat{\Theta}_2$.

Entre todos los estimadores insesgados de θ, el que tenga menor varianza es el **estimador insesgado de mínima varianza** o estimador más eficiente de θ.

En la explicación previa, un estimador $\hat{\Theta}$ produce un valor $\hat{\theta}$ que pretende aproximar a un parámetro θ.

A este enfoque se le llama estimación puntual

A pesar de la indudable utilidad de la estimación puntual, en la práctica, cuando se realiza la estimación de un parámetro se necesita obtener una medida de la fiabilidad de dicha estimación, de este modo, surge la necesidad de encontrar un método que permita calcular una región que contenga al valor del parámetro con una cierta garantía.

Ejemplo

Una empresa tabacalera desea estudiar el nivel medio de nicotina de sus cigarros. A la compañía le interesa que el nivel medio de nicotina se encuentre entre unos márgenes debido a que un nivel medio alto supone que el cigarro es muy perjudicial para la salud y un nivel medio bajo implica que el cigarro carece de sabor. De este modo, si el nivel de nicotina de un cigarro viene dado por una variable aleatoria, X, tal que E[X] =θ , donde θ es un parámetro desconocido, se desea, a partir de una muestra X1;X2;…;Xn, obtener LI y LS tal que

$$P[LI < \theta < LS] = 1 - \alpha$$

En el enfoque de **estimación de intervalos**, para un parámetro θ no se estima un valor, sino un intervalo de la forma $\hat{\theta}_l < \theta < \hat{\theta}_s$, donde los valores extremos dependen del valor numérico del estadístico $\hat{\theta}$ para una muestra en particular y de la distribución muestral de ese estadístico.

Partiendo de la distribución muestral para $\hat{\Theta}$, es posible determinar valores de $\hat{\theta}_i$, $\hat{\theta}_s$ tal que $P(\hat{\Theta}_i < \theta < \hat{\Theta}_s)$ sea igual a cualquier valor que se desee especificar. Por ejemplo, si

$$P(\hat{\Theta}_i < \theta < \hat{\Theta}_s) = 1 - \alpha \qquad (0 < \alpha < 1)$$

Significa que si se toman todas las muestras posibles de tamaño n de una población la probabilidad que el parámetro θ esté en el intervalo $(\hat{\Theta}_i, \hat{\Theta}_s)$ es de 1-α

Si estimo θ, a partir de una muestra única se obtiene una estimación puntual $\hat{\theta}$ del estadístico $\hat{\Theta}$, la que permitirá calcular un intervalo $\theta_i < \theta < \theta_s$ para θ con $1\text{-}\alpha$ de probabilidad de que ese intervalo sea uno que contenga al parámetro poblacional θ.

Al intervalo $\theta_i < \theta < \theta_s$ que se calcula a partir de la muestra seleccionada se denomina intervalo de confianza del $(1-\alpha)$ 100 %.

Intervalo de Confianza

Se considera una m.a. X1;X2;…;Xn procedente de una población definida por una variable aleatoria X, cuya distribución depende de un parámetro desconocido . El objetivo es desarrollar un método para calcular intervalos de confianza a partir de una función de la muestra que contenga al parámetro y cuya distribución no dependa de él.

Por la importancia de la Distribución Normal . Se construirán para este caso intervalos de confianza para su valor medio y su varianza considerando diferentes casos que dependen del conocimiento o no de los otros parámetros

Estimación de la media

Situación: Se tiene una población con media desconocida μ, pero se supone conocida la varianza σ^2.

Se toma una muestra aleatoria (X_1,X_2,...,Xn). Con esta muestra se calcula el estadístico $\bar{X}$ el cual es un estimador puntual insesgado para la media μ desconocida. Se puede obtener un intervalo de confianza del (1-α) 100% para μ si consideramos los siguientes hechos acerca de la distribución de:

1. Si la población es Normal, la distribución de $\bar{X}$ es Normal

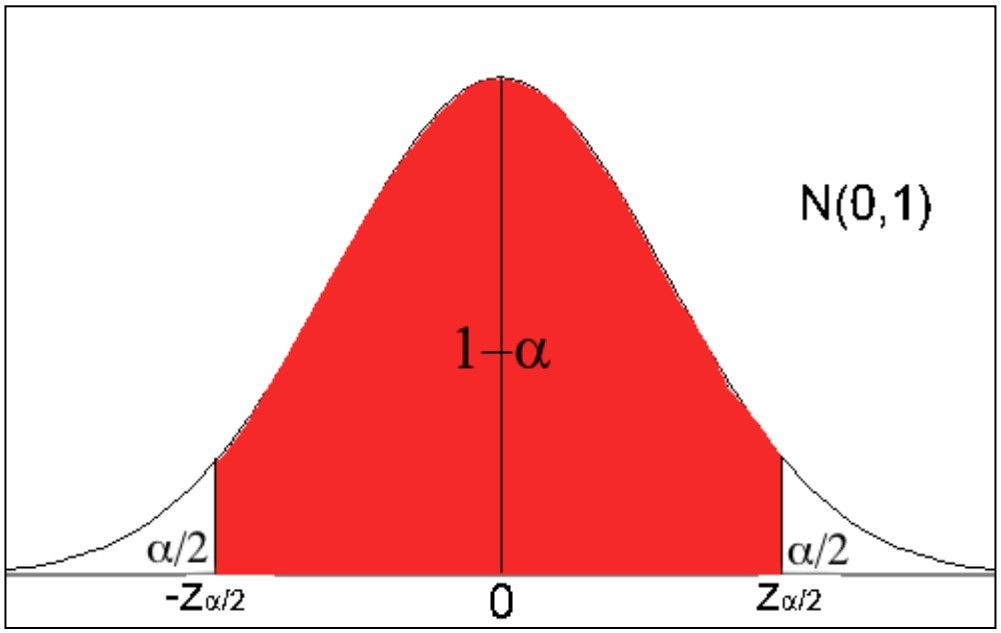
2. Si la población no es Normal, el Teorema del límite central nos garantiza una distribución de $\bar{X}$ aproximadamente normal cuando n $\to\infty$

3. La media de $\bar{X}$ es μ ($\bar{X}$ es insesgado)

4. La varianza de $\bar{X}$ es σ^2/n

entonces la variable aleatoria $$Z = \frac{\overline{x} - \mu_x}{\sigma_X / \sqrt{n}}$$

tiene una distribución normal estándar Z~N(0,1)

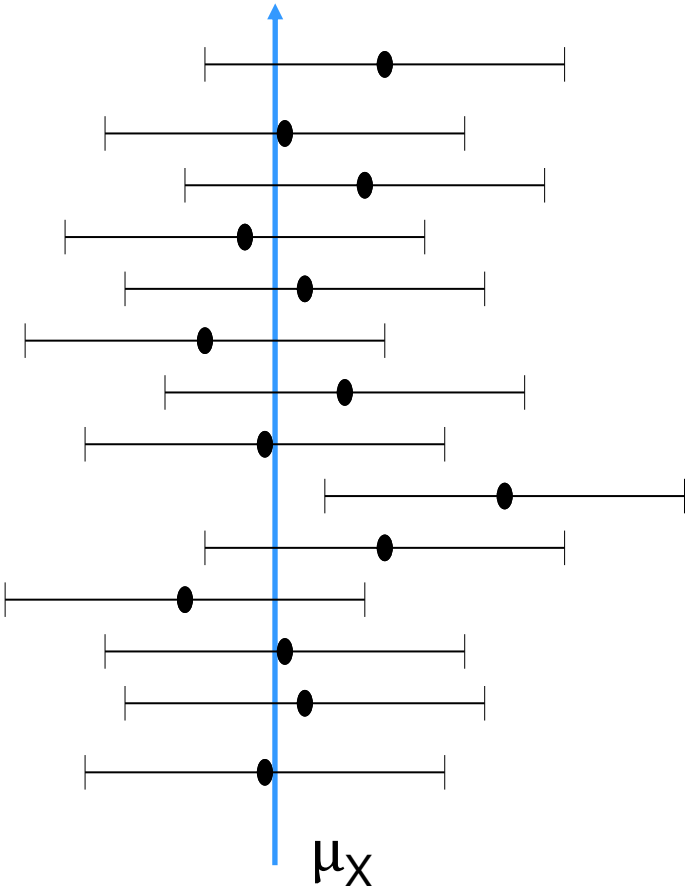


de la figura: $P\{-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\} = 1-\alpha$.

Con lo cual el intervalo de confianza del $(1-\alpha)100\%$ para la media es

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_X \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Muestras diferentes darán valores diferentes de $\bar{x}$ y, por lo tanto producirán diferentes estimaciones por intervalos de confianza del parámetro μ_X



Para algunas muestras la estimación del intervalo de confianza será correcto y para otras no. En la práctica seleccionamos sólo una muestra y no conocemos la μ de la población y tampoco podemos determinar si nuestra estimación es correcta. Podemos determinar la porción de muestras que producen resultados que nos llevan a construir intervalos de confianza que nos conducen a conclusiones correctas respecto a la μ poblacional.

Error en la estimación de μ_x

La precisión del intervalo de confianza es $z_{\alpha/2}\,\sigma/\sqrt{n}$ esto significa que al usar $\bar{X}$ para estimar μ, el error de estimación, dado por $E=|\bar{X}-\mu| \leq z_{\alpha/2}\,\sigma/\sqrt{n}$, con una confianza de $(1-\alpha)100\%$.

Con frecuencia se desea determinar el tamaño de la muestra para asegurar que el error en la estimación de μ_X será menor que una cantidad especificada e.

$$z\alpha_{/2}\frac{\sigma_X}{\sqrt{n}} \leq e$$

Si se utiliza $\bar{x}$ como una estimación de μ_X, se puede tener una confianza del $(1-\alpha)100\%$ de que el error no excederá una cantidad e cuando el tamaño de la muestra es:

$$n \geq \left(z\alpha_{/2}\frac{\sigma_X}{e}\right)^2$$

Ejemplo: Los siguientes son datos de conductividad térmica de cierto tipo de hierro (en BTU/hr-ft-°F):

41.60 41.48 42.34 41.95 41.86
42.18 41.72 42.26 41.81 42.04

Hallar un intervalo de confianza del 95 % y uno del 99% para la media.

Se supone que la población tiene una distribución Normal con σ=0.3

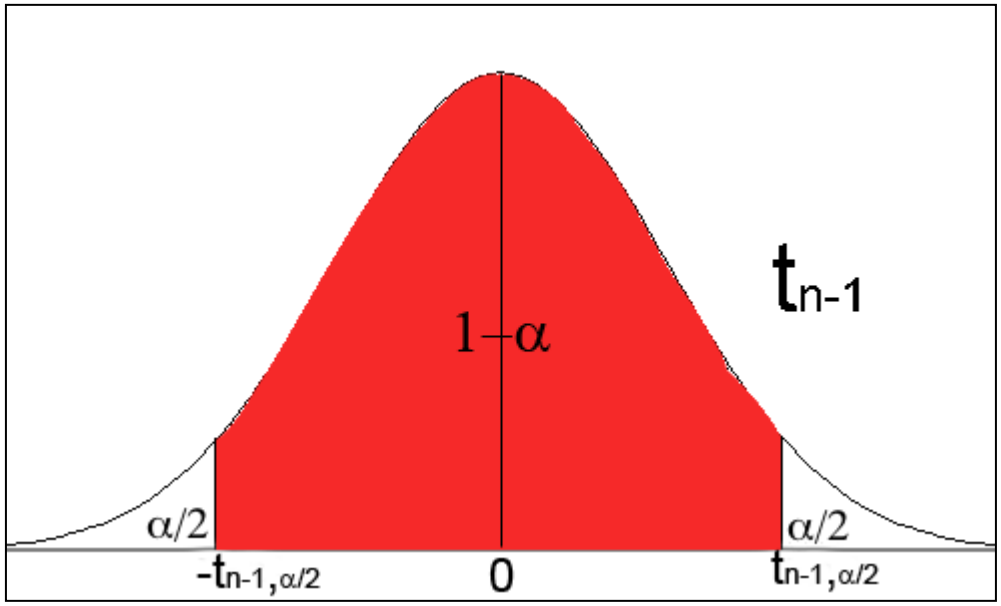
Estimación de μ_X cuando se desconoce σ^2_X

Si no se conoce la varianza σ^2 de la población, una posibilidad es utilizar la varianza muestral S^2 en las ecuaciones obtenidas para estimar intervalos en el caso de varianza conocida.

Si la población es Normal, la variable aleatoria T/

$$T = \frac{\bar{X} - \mu_X}{\frac{S}{\sqrt{n}}}$$

tiene una distribución t de Student con n-1 grados de libertad



De la figura tenemos:

$$P\left(-t\alpha_{/2} < T < t\alpha_{/2}\right) = 1 - \alpha$$

Para el caso de una muestra aleatoria de tamaño n, de una población normal, se calculan $\bar{x}$ y s y se obtiene el siguiente intervalo de confianza del (1-α)100% para μ_X cuando se desconoce $\boldsymbol{\sigma^2}_{\boldsymbol{X}}$

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu_X < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Ejemplo:

El contenido de siete contenedores similares de ácido sulfúrico son

9,8 10,2 10,4 9,8 10,0 10,2 9,6 litros

Encuentre un intervalo de confianza del 95% para la media de todos los contenedores si se supone una distribución aproximadamente normal

Estimación de la diferencia entre dos medias con varianzas conocidas

Población 1
$X_1 \sim N(\mu_{X_1}, \sigma^2_{X_1})$
n_1, $\bar{x}_1$

Población 2
$X_2 \sim N(\mu_{X_2}, \sigma^2_{X_2})$
n_2, $\bar{x}_2$

Se puede esperar que la DM$(\bar{X}_1 - \bar{X}_2)$ sea aproximadamente en forma Normal con

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{X_1} - \mu_{X_2} \qquad \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma^2_{X_1}}{n_1} + \frac{\sigma^2_{X_2}}{n_2}$$

$$\boldsymbol{P\left(-z\alpha_{/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\frac{\sigma^2_{X_1}}{n_1} + \frac{\sigma^2_{X_2}}{n_2}}} < z\alpha_{/2} \right) = 1 - \alpha}$$

$$P\left((\overline{x}_1 - \overline{x}_2) - z\alpha_{/2}\sqrt{\frac{\sigma^2{}_{X_1}}{n_1} + \frac{\sigma^2{}_{X_2}}{n_2}} < (\mu_{X_1} - \mu_{X_2}) \right.$$

$$\left. < (\overline{x}_1 - \overline{x}_2) + z\alpha_{/2}\sqrt{\frac{\sigma^2{}_{X_1}}{n_1} + \frac{\sigma^2{}_{X_2}}{n_2}} \right) = 1 - \alpha$$

Ejemplo (varianzas conocidas)

Se quiere estudiar la diferencia de las vidas medias de dos tipos de lámparas. Para ello, se toma una muestra de 150 lámparas de tipo H y otra, independiente de la anterior, de 200 lámparas de tipo N, obteniéndose que las de tipo H tienen una vida media de 1400 horas y una desviación típica de 120, y que las de tipo N tienen una vida media de 1200 horas y desviación típica 80.

Ejemplo: Se piensa que la concentración del ingrediente activo de un detergente líquido para ropa, es afectada por el tipo de catalizador utilizado en el proceso de fabricación. Se realizan 10 observaciones con cada catalizador, y se obtienen los datos siguientes:

Catalizador 1: 57.9, 66.2, 65.4, 65.4, 65.2, 62.6, 67.6, 63.7, 67.2, 71.0
Catalizador 2: 66.4, 71.7, 70.3, 69.3, 64.8, 69.6, 68.6, 69.4, 65.3, 68.8

a) Encuentre un intervalo de confianza del 95% para la diferencia entre las medias de las concentraciones activas para los dos catalizadores. Asumir que ambas muestras fueron extraídas de poblaciones normales con varianzas iguales. b) ¿Existe alguna evidencia que indique que las concentraciones activas medias dependen del catalizador utilizado?

Ejemplo:
Una muestra de 6 soldaduras de un tipo tenía promedio de prueba final de resistencia de 83.2 ksi y desviación estándar de 5.2. Y una muestra de 10 soldaduras de otro tipo tenía resistencia promedio de 71.3 ksi y desviación estándar de 3.1. supongamos que ambos conjuntos de soldaduras son muestras aleatorias de poblaciones normales. Se desea encontrar un intervalo de confianza de 95% para la diferencia entre las medias de las resistencias de los dos tipos de soldaduras.

Observaciones pareadas

Consideramos los procedimientos de estimación para la diferencia de medias cuando las muestras no son independientes y las varianzas de las dos poblaciones no son necesariamente iguales

Definimos la variable aleatoria diferencia $D = X_1 - X_2$
Las diferencias se asumen con distribución normal con media $\mu_D = \mu_1 - \mu_2$ y varianza $\sigma^2{}_D$ desconocida. Se estima por $S^2{}_d$ varianza muestral de las diferencias D. Para el valor promedio de las diferencias $\overline{D}$, la distribución muestral es t de Student con γ=n-1 grados de libertad

$$T = \frac{\overline{D} - \mu_D}{\frac{S_d}{\sqrt{n}}}$$

Emplearemos las DM(T) para determinar el intervalo de confianza de μ_D

$$P\left(-t\alpha_{/2} < T < t\alpha_{/2}\right) = 1 - \alpha$$

Resulta

$$P\left(-t\alpha_{/2} < \frac{\overline{D} - \mu_D}{\frac{S_d}{\sqrt{n}}} < t\alpha_{/2}\right) = P\left(\overline{D} - t\alpha_{/2}\frac{S}{\sqrt{n}} < \mu_D < \overline{D} + t\alpha_{/2}\frac{S}{\sqrt{n}}\right)$$
$$= 1 - \alpha$$

Por ejemplo, supongamos que se mide el tiempo en segundos que un individuo tarda en hacer una maniobra de estacionamiento con dos automóviles diferentes en cuanto al tamaño de la llanta y la relación de vueltas del volante

	Automóvil 1	*Automóvil 2*	*diferencia*
sujeto	*(observación* x_{1j}*)*	*(observación* x_{2j}*)*	D_j
1	37.0	17.8	19.2
2	25.8	20.2	5.6
3	16.2	16.8	-0.6
4	24.2	41.4	-17.2
5	22.0	21.4	0.6
6	33.4	38.4	-5.0
7	23.8	16.8	7.0
8	58.2	32.2	26.0
9	33.6	27.8	5.8
10	24.4	23.2	1.2
11	23.4	29.6	-6.2
12	21.2	20.6	0.6
13	36.2	32.2	4.0
14	29.8	53.8	-24.0

Estimación de Proporción

Un estimador puntual de la proporción p en un experimento binomial está dado por el estadístico $\hat{P} = \frac{X}{n}$ donde X es el número de éxitos en n intentos .

La proporción muestral $\hat{p} = \frac{x}{n}$ se utiliza como una estimación puntual del parámetro p. Siempre que la proporción desconocida p no se acerque demasiado a 0 o a 1, se puede establecer un intervalo de confianza para p considerando la DM($\widehat{P}$)

La DM($\widehat{P}$) se puede considerar aproximadamente Normal con media p y varianza p(1-p)/n

$$\hat{P} \approx N\left(p, \frac{pq}{n}\right) \qquad z = \frac{\bar{P} - p}{\sqrt{\frac{\hat{P}\hat{Q}}{n}}}$$

$$P\left(-z\alpha_{/2} < Z < z\alpha_{/2}\right) = 1 - \alpha$$

$$P\left(\hat{P} - z\alpha_{/2}\sqrt{\frac{\hat{P}\hat{Q}}{n}} < p < \hat{P} - z\alpha_{/2}\sqrt{\frac{\hat{P}\hat{Q}}{n}}\right) = 1 - \alpha$$

Para n grande se emplean estimaciones puntuales, y se obtiene el intervalo de confianza de (1-α)100%

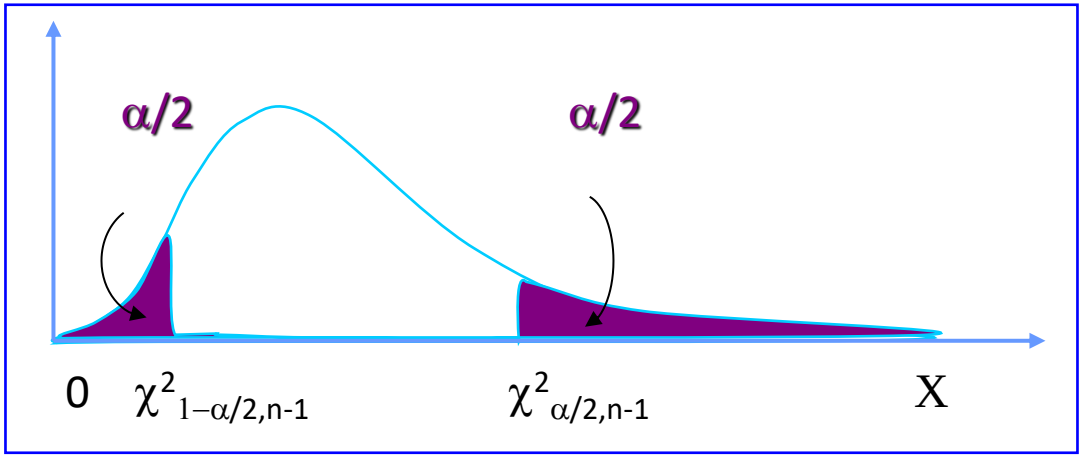
$$\hat{p} - z\alpha_{/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z\alpha_{/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Estimación de la varianza

Si la Población es Normal, la distribución muestral del estadístico siguiente

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2_X}$$

donde S^2 es la varianza muestral usada como estimador puntual de σ^2 , es de tipo Ji-cuadrada con n-1 grados de libertad



$$P(\chi^2_{1-\alpha/2,n-1} < \chi^2 < \chi^2_{\alpha/2,n-1}) = 1 - \alpha$$

$$P(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2_X} < \chi^2_{\alpha/2,n-1}) = 1-\alpha$$

$$P(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2_X < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}) = 1-\alpha$$

Si para una única muestra aleatoria de tamaño n se calcula la varianza muestral s^2, se obtiene el siguiente intervalo de confianza

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2_X < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

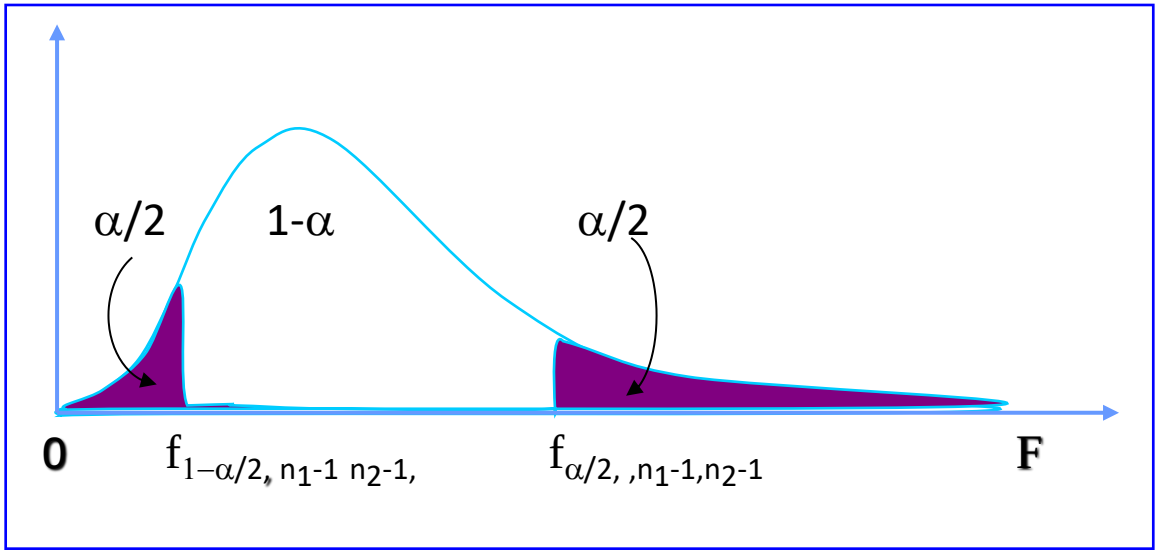
Ejemplo

Se sabe que el peso por comprimido de un cierto preparado farmacéutico se distribuye según una Normal. Con el objeto de estudiar la varianza de la distribución, se extrae una m.a. de 6 artículos. Sabiendo que la varianza muestral es igual a 40, se pretende estimar la varianza poblacional mediante un intervalo de confianza al 90 %.

Estimación de la razón de dos varianzas

Se tienen dos poblaciones normales e independientes con varianzas desconocidas σ_1^2, σ_2^2 respectivamente. Se tienen disponibles dos muestras aleatorias de tamaños n_1, n_2 una de cada población respectivamente. Sean S_1^2 , S_2^2 las varianzas muestrales respectivas.

Se puede determinar el intervalo de confianza de $\sigma_{X_1}^2 / \sigma_{X_2}^2$ empleando el estadístico F con una distribución muestral F de Fisher con γ_1, γ_2 grados de libertad



$$P(f_{1-a/2,\, n_1-1\ n_2-1} < F < f_{a/2,\ ,n_1-1,n_2-1}) = 1-\alpha$$

$$P\left(f_{1-\alpha/2,\, n_1-1\ n_2-1} < \frac{S_1^2}{S_2^2}\frac{\sigma_{X_2}^2}{\sigma_{X_1}^2} < f_{\alpha/2,\ ,n_1-1,n_2-1}\right) = 1-\alpha$$

$$P\left(\frac{1}{f_{\alpha/2,\ ,n_1-1,n_2-1}}\frac{S_1^2}{S_2^2} < \frac{\sigma_{X1}^2}{\sigma_{X2}^2} < \frac{1}{f_{1-\alpha/2,\, n_1-1\ n_2-1}}\frac{S_1^2}{S_2^2}\right) = 1-\alpha$$

$$P\left(\frac{1}{f_{\alpha/2,\,,n_1-1,n_2-1}}\frac{S_1^{\,2}}{S_2^{\,2}}<\frac{\sigma_{X1}^{\,2}}{\sigma_{X2}^{\,2}}<f_{\alpha/2,\,,n_2-1,n_1-1}\frac{S_1^{\,2}}{S_2^{\,2}}\right)=1-\alpha$$

Para dos únicas muestras de tamaño n_1 y n_2 tomadas en forma independiente de las dos poblaciones normales .

Se obtiene el siguiente intervalo de confianza de $(1-\alpha)100\%$ para $\sigma_{X_1}^{\,2}/\,\sigma_{X_2}^{\,2}$

$$\frac{1}{f_{\alpha/2,\,,n_1-1,n_2-1}}\frac{s_1^{\,2}}{s_2^{\,2}}<\frac{\sigma_{X1}^{\,2}}{\sigma_{X2}^{\,2}}<f_{\alpha/2,\,,n_2-1,n_1-1}\frac{s_1^{\,2}}{s_2^{\,2}}$$

Ejemplo:

Una compañía fabrica piezas para turbinas. Tiene dos procesos distintos para hacer el esmerilado de las piezas y ambos procesos producen terminados con la misma rugosidad promedio. El ingeniero del proceso desea seleccionar el proceso con la menor variabilidad en la rugosidad de la superficie. Para ello toma una muestra de 12 piezas del primer proceso, obteniendo una desviación estándar muestral $s_1 = 5.1$ micropulgadas, luego toma una muestra de 15 piezas del segundo proceso, obteniendo $s_2 = 4.7$. ¿Puede elegir el primer proceso con una confianza del 90% de tener menor variabilidad en la rugosidad?