Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

ESTADÍSTICA

Ingeniería Informática

TEORÍA

Mg.Ing. Susana Vanlesberg
Profesor Titular

UNIDAD 6

INFERENCIA ESTADÍSTICA DISTRIBUCIÓN EN EL MUESTREO -ESTIMACIÓN

DISTRIBUCIÓN EN EL MUESTREO

Hemos desarrollado en la unidad anterior lo referido al análisis de datos, ahora bien cuando se enfrenta un trabajo, generalmente el objetivo es conocer más acerca de la población de referencia, entonces las características muestrales serán el comienzo para este proceso.

Si la obtención de características muestrales se repite un determinado número de veces, es decir se sacan muestras de igual extensión, de la misma población que tiene una distribución dada, y en todas ellas se obtiene la misma función (la misma característica), los valores variarán de muestra a muestra. Esto permite considerar a las características muestrales como variables aleatorias. Como variables aleatorias tienen una distribución de probabilidad que les es propia. Generalmente se las conoce como ***distribución del estadístico por muestreo***. Estas distribuciones tienen propiedades bien definidas.

El proceso de analizar los datos tratando de traducir lo que ellos dicen en términos de probabilidad, con el fin de obtener conclusiones respecto a la población es lo que se denomina ***Inferencia Estadística***

ESTADÍSTICOS TRATADOS COMO VARIABLES ALEATORIAS

1- La media **y** la varianza muestral son dos de los estadísticos más importantes que serán estudiadas.

La media muestral será estudiada respondiendo a las siguientes preguntas:

- ¿Cuál es su valor medio?
- ¿Cuál es su varianza?
- ¿Cuál es su distribución?

Para empezar a responder se considera que los valores muestrales x_i son **independientes e idénticamente distribuidos**, con esperanza común μ y varianza σ^2 ya que provienen de la misma población.

$$E(\bar{x}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\, n\, E(x) = \mu$$

Este resultado indica que el valor medio de la variable aleatoria media muestral es igual al valor medio de la población.

La varianza de la variable aleatoria media muestral se obtiene por hallar la varianza del promedio de n valores independientes o idénticamente distribuidos:

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n^2}\cdot \mathrm{Var}\sum_{i=1}^{n} x_i = \frac{1}{n^2}\cdot \sum_{i=1}^{n} \mathrm{Var}(x_i) = \frac{1}{n^2}\cdot n \cdot \sigma^2$$

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}$$

El error estándar de la media muestral, que mide la variabilidad casual en medias de muestras es:

$$\sigma(\bar{x}) = \sqrt{\mathrm{Var}(x)} = \frac{\sigma}{\sqrt{n}}$$

La expresión anterior nuestra que el desvío de la media muestral es menor que el desvío de la población. Además cuando n tiende a infinito el desvío de la media muestral tiende a cero, esto significa que cuanto mayor es la extensión de la muestra, menor será el error o fluctuación de las medias de una muestra a otra.

Si las muestras son extraídas de una población finita y el muestreo se realiza sin reposición, se debe introducir un factor de corrección por población finita en el error de la media:

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

N extensión de la población y n: extensión de la muestra.

Llegados hasta aquí es conveniente presentar un teorema básico para el desarrollo de la Teoría de Inferencia, ***Teorema del Limite Central.***

Teorema del Límite Central.

Este teorema es muy útil, ya que es importante saber más acerca de la distribución de una suma de variables aleatorias. Su enunciado es el siguiente:

Sí se considera la suma de n variables* aleatorias x *independientes e idénticamente distribuidas, cada una con media y varianza finita, cuando el número de variables involucradas es mayor, la distribución de la suma se aproxima a una distribución Normal.

El valor de este teorema es que no requiere condiciones para las distribuciones de las variables aleatorias que se suman, sólo es necesario que cada una tenga un efecto insignificante sobre la distribución de la suma. Además brinda un método práctico para calcular valores de probabilidad aproximados asociados con sumas de variables aleatorias independientes distribuidas arbitrariamente

Este teorema es muy usado, ya que muchas variables aleatorias pueden considerarse como la suma de efectos independientes.

$$S = x_1 + x_2 + .. + x_n$$

$$S = \sum_{i=1}^{n} x_i = n\bar{x} \qquad E(x_i) = \mu \qquad \sigma^2(x_i) = \sigma^2$$

$$E[S] = E\left[\sum_{i=1}^{n} x_i\right] = E\left[n \cdot \bar{x}\right] = nE\left[\bar{x}\right]$$

$$\mathrm{Var}[S] = \mathrm{Var}\left[\sum_{i=1}^{n} x_i\right] = \mathrm{Var}\left[n\bar{x}\right] = n^2.\mathrm{Var}\left[\bar{x}\right]$$

$$\sigma(S) = n.\sigma\left(\bar{x}\right)$$

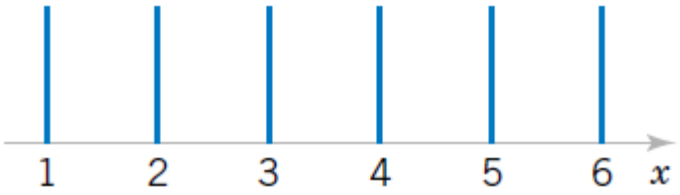
$$P\left(\frac{\bar{x} - E\left(\bar{x}\right)}{\sigma\left(\bar{x}\right)} \leq x\right) \rightarrow_{n\to\infty} N(0,1)$$

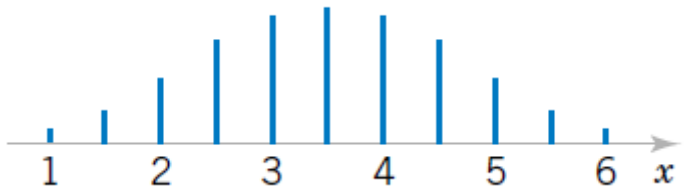
Se concluye, por lo tanto, que la variable aleatoria *media muestral* se distribuye normalmente con parámetros $E\left(\bar{x}\right)$ y $\sigma\left(\bar{x}\right)$

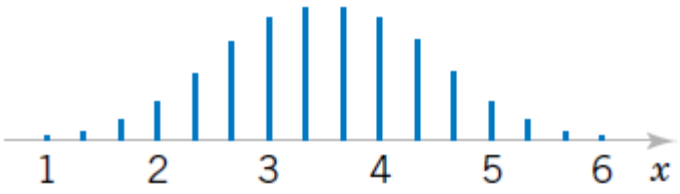
La conclusión es muy importante, ya que la mayor parte de los procedimientos de inferencia se basan en $\bar{x}$ .Si las variables que conforman la muestra se distribuyen normalmente, entonces la $\bar{x}$ también será distribuida normalmente, y así se puede

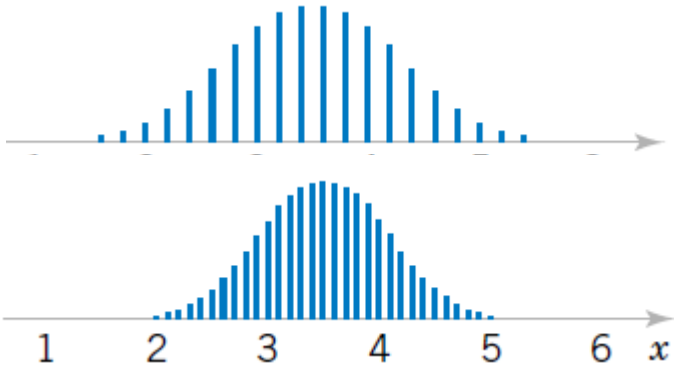
aplicar la teoría sobre variables distribuidas normalmente. En cambio, si las variables que conforman la muestra no son normales, entonces, para aplicar este teorema, es necesario que la extensión de la muestra n sea grande, y así $\bar{x}$ puede considerarse como distribuida normalmente.

Aunque el Teorema del Límite Central va a funcionar bien para muestras pequeñas (n= 4, 5) en la mayoría de los casos, sobre todo cuando la población es continua, unimodal y simétrica, se requiere muestras más grandes en otras situaciones, dependiendo de la forma de la población. En muchos casos de interés práctico, si n≥ 30, la aproximación normal será satisfactoria independientemente de la forma de la distribución de la población.









Se dijo anteriormente que la varianza muestral es, junto a la media, uno de los estimadores más importantes. Tratada como una variable aleatoria, es necesario obtener sus momentos:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \mu) - (\bar{x} - \mu)\right]^2$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{2}{n}\left(\sum(x_i - \mu)(\bar{x} - \mu)\right) + \frac{1}{n}\sum(\bar{x} - \mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - 2\left(\frac{\sum x_i}{n} - \frac{n.\mu}{n}\right)(\bar{x} - \mu) + \frac{n}{n}(\bar{x} - \mu)^2$$

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

luego $E[S^2] = \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \mu)^2\right] - E\left[(\bar{x} - \mu)^2\right]$

$$E[S^2] = \sigma_x^2 - \sigma_x^2 = \sigma_x^2 - \frac{\sigma_x^2}{n} = \sigma_x^2 \cdot \frac{n-1}{n}$$

$$E[S^2] \neq \sigma^2$$

La diferencia (n-1)/n *se denomina sesgo* y tiene realmente importancia cuando n es pequeño, ya que en caso contrario, el sesgo tiende a 1.

Un estimador insesgado de σ^2 es la varianza muestral corregida:

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Con lo cual:

$$E[S'^2] = E\left(\frac{n.S^2}{n-1}\right) = E\left(\frac{n}{n-1} \cdot \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}\right)$$

$$E[S'^2] = \sigma^2$$

Para seguir el mismo razonamiento que con la media haría falta encontrar la varianza de la varianza muestral pero es muy extensa esta demostración.

Decimos únicamente que para poblaciones Normales la varianza del estimador S^2 es:

$$Var(S^2) = \frac{2\sigma^4}{\upsilon}$$

Para obtener la distribución por muestreo que le corresponderá a S^2 es necesario recordar la variable χ^2 que surge como la suma de cuadrados de variables aleatorias normales estandarizada:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

Si μ se desconoce se lo estima a través de la media muestral con lo cual la expresión anterior se transforma en:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^2 = \frac{n S^2}{\sigma^2} \sim \chi^2_{n-1}$$

O bien utilizando la varianza muestral corregida:

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

En los problemas que ocurren frecuentemente en ingeniería se necesita hacer estimación, relacionada generalmente a:

- ✓ **La media de la población**
- ✓ **La varianza de la población**
- ✓ **La proporción p de elementos en una población que pertenecen a una clase de interés**
- ✓ **La diferencia entre las medias de dos poblaciones**
- ✓ **La diferencia entre las proporciones de dos poblaciones**

Los estimadores razonables de esos parámetros son:

Para μ, la media muestral $\bar{x}$

Para σ^2, la varianza muestral S^2
Para π, la proporción muestral p
Para la diferencia de medias poblacionales μ_1- μ_2, la diferencia de medias muestrales $\bar{x}_1 - \bar{x}_2$
Para la diferencia de proporciones poblacionales π_1-π_2, la diferencia de proporciones muestrales p_1-p_2

Por lo tanto se debe analizar de todos estos estimadores su distribución muestral:

DISTINTOS CASOS DE DISTRIBUCIÓN POR MUESTREO

1.-Distribución por muestreo de medias

Población Normal con desvío σ conocido

Estandarizando la variable aleatoria media muestral se obtiene una variable Normal estándar:

$$X \sim N \quad (\mu, \sigma) \quad ; \quad \bar{x} \sim N \quad \left(\mu, \frac{\sigma}{\sqrt{n}}\right);$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N \quad (0,1)$$

En el caso de que la población tenga una distribución aproximadamente Normal, los resultados son similares sólo que le variable z será también distribuida en forma aproximadamente Normal.

Población Normal, σ desconocido, muestra chica (n < 30)

Si las variables que constituyen la muestra son independientes e idénticamente distribuidas con media y varianza finita, pero como sucede generalmente, la varianza se desconoce será reemplazada por la varianza muestral, entonces la variable resultante se distribuye como t de Student. Esto es debido a que como ya se demostró, una variable ***t*** se genera como el cociente de una variable Normal y la raíz cuadrada de una variable chi-cuadrado dividida por sus grados de libertad; entonces:

$$t = \frac{z}{\sqrt{\frac{\chi^2}{v}}} \quad siendo \quad z \sim N \quad (0,1)$$

$$= \frac{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\left(\frac{nS^2}{\sigma^2}\right)}{n-1}}} = \frac{\left(\bar{x} - \mu\right)\sqrt{n}}{\frac{\sqrt{n}.S}{\sigma\sqrt{n-1}}}$$

$$= \frac{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\left(\frac{nS^2}{\sigma^2}\right)}{n-1}}} = \frac{\left(\bar{x} - \mu\right)\sqrt{n}}{\frac{\sqrt{n}.S}{\sigma\sqrt{n-1}}}$$

$$t_{n-1} = \frac{\bar{x} - 1}{\frac{S}{\sqrt{n-1}}} \quad o \quad bien \quad t_{n-1} = \frac{\bar{x} - \mu}{\frac{S'}{\sqrt{n}}}$$

Población Normal, σ desconocido, muestra grande (n ≥ 30)

Cuando la muestra es grande se puede considerar la varianza poblacional desconocida reemplazada por la varianza muestral y la distribución de la variable resultante sigue siendo Normal:

$$z = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

2.-Distribución muestral de la varianza

Provenientes los valores muestrales de una población con distribución Normal recordar, ya se ha mostrado que la varianza muestral tiene una distribución Chi-cuadrado.

$$\chi^2_{n-1} = \frac{n.S^2}{\sigma^2}$$

3.-Distribución muestral de proporciones

En la población, una proporción se define como:

$$\pi = \frac{K}{N}$$

Siendo K el número de elementos que tienen una característica deseada y N el total de elementos de la población. En la muestra, se define como:

$$p = \frac{x}{n}$$

siendo p la proporción muestral, x la cantidad de elementos que poseen la categoría deseada y n la extensión de la muestra. Suele considerarse a p como la proporción de éxitos, y por esto se la asocia a la distribución Binomial (recuérdese que E = n.p y Var = n.p.q).

Luego, las características de esta variable aleatoria son:

$$E[p] = E\left[\frac{x}{n}\right] = n\frac{\pi}{n} = \pi$$

$$Var[p] = Var\left[\frac{x}{n}\right] = \frac{n\,\pi\,(1-\pi)}{n^2} = \frac{\pi\,(1-\pi)}{n}$$

El error estándar de p mide las variaciones casuales de proporciones de muestra de una muestra a otra:

$$\sigma_p = \sqrt{\frac{\pi\,(1-\pi)}{n}}$$

Este error debe ajustarse por un factor de corrección por población finita, si el muestreo se hace sin reposición:

$$\sigma_p = \sqrt{\frac{\pi\ (1-\pi\ )}{n}}\sqrt{\frac{(N-n\ )}{n\ -1}}$$

Luego, la distribución muestral de es la siguiente:

$$p \sim N\left(\pi;\sqrt{\frac{\pi\ (1-\pi\ )}{n}}\right)$$

$$z = \frac{p-\pi}{\sqrt{\frac{\pi\ (1-\pi\ )}{n}}} \sim N(0,1)$$

4.-Distribución muestral de la diferencia de dos medias muestrales

Varianzas poblacionales conocidas

Cuando sea de interés comparar las medias de dos variables aleatorias, esto se hará sobre la base de dos muestras extraídas de las poblaciones cuyas medias se quiere comparar.

$$x \sim N(\mu_x,\sigma_x) \qquad y \sim N(\mu_y,\sigma_y)$$

$$\bar{x} \sim N\left(\mu_x,\frac{\sigma_x}{\sqrt{n_x}}\right) \qquad \bar{y} \sim N\left(\mu_y,\frac{\sigma_y}{\sqrt{n_y}}\right)$$

siendo $\bar{x}$ independiente de $\bar{y}$

Usando los resultados de combinaciones lineales de variables distribuidas normalmente puede decirse que la variable aleatoria *diferencia de medias* muestrales se distribuye normalmente. Aún sin saber si las poblaciones son normales, si las extensiones de muestras son suficientemente grandes, como cada media muestral se distribuye normalmente, es de esperar que la diferencia de medias muestrales sea también normalmente distribuida. Los parámetros de esta distribución normal son:

$$E\left(\bar{x}-\bar{y}\right) = E\left(\bar{x}\right) - E\left(\bar{y}\right) = \mu_x - \mu_y$$

$$\mathrm{Var}\left(\bar{x}-\bar{y}\right)=\mathrm{E}\left[\bar{x}-\bar{y}-\mathrm{E}\left(\bar{x}-\bar{y}\right)\right]^2=$$

$$=\mathrm{E}\left[\left(\bar{x}-\bar{y}\right)-\mathrm{E}\left(\bar{x}\right)+\mathrm{E}\left(\bar{y}\right)\right]^2=\mathrm{E}\left[\left(\bar{x}-\mathrm{E}\left(\bar{x}\right)\right)-\left(\bar{y}-\mathrm{E}\left(\bar{y}\right)\right)\right]+\mathrm{E}\left[\left(\bar{y}-\mathrm{E}\left(\bar{y}\right)\right)\right]^2=$$

$$=\mathrm{E}\left[\bar{x}-\mathrm{E}\left(\bar{x}\right)\right]^2-2\mathrm{E}\left[\left(\bar{x}-\mathrm{E}\left(\bar{x}\right)\right)\cdot\left(\bar{y}-\mathrm{E}\left(\bar{y}\right)\right)\right]+\mathrm{E}\left[\left(\bar{y}-\mathrm{E}\left(\bar{y}\right)\right)\right]^2$$

Como la covarianza de variables aleatorias independientes es igual a cero, luego:

$$\mathrm{Var}\left(\bar{x}-\bar{y}\right)=\mathrm{Var}\left(\bar{x}\right)+\mathrm{Var}\left(\bar{y}\right)=\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}$$

Entonces, si las varianzas de ambas poblaciones se conocen, se obtiene:

$$\frac{\bar{x}-\bar{y}-\left(\mu_x-\mu_y\right)}{\sqrt{\left(\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}\right)}}=z\sim N\left(0,1\right)$$

Varianzas poblacionales desconocidas - Muestras grandes

La mayoría de las veces las varianzas poblacionales se desconocen y deben ser estimadas. En este caso, es decir muestras grandes, la distribución de la diferencia de medias muestrales sigue siendo Normal pero con la variable z con la siguiente forma:

$$z=\frac{\bar{x}-\bar{y}-\left(\mu_x-\mu_y\right)}{\sqrt{\left(\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}\right)}}\sim N\left(0,1\right)$$

Varianzas poblacionales desconocidas - Muestras chicas

En este caso debe hacerse la siguiente consideración respecto a las varianzas poblacionales desconocidas:

Varianzas poblacionales desconocidas pero supuestas iguales e iguales a un valor constante.

Recordar que la distribución chi-cuadrado está asociada a la varianza muestral, luego, usando las propiedades reproductivas de la distribución Chi-cuadrado se obtiene:

$$\frac{(n_x-1)S'^2_x}{\sigma^2}+\frac{(n_y-1)S'^2_y}{\sigma^2} \sim \chi^2_{n_x+n_y-2}$$

con lo cual la distribución deja de ser Normal para transformarse en t de Student. Recordar como surge una variable t de Student: como el cociente entre une variable Normal (0,1) y la raíz cuadrada de una variable chi-cuadrado dividida por sus grados de libertad

$$= \frac{\dfrac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sigma\sqrt{\dfrac{1}{n_x}+\dfrac{1}{n_y}}}}{\dfrac{(n_x-1)S'^2_x+(n_y-1)S'^2_y}{\sigma^2(n_x+n_y-2)}} =$$

$$= \frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\dfrac{(n_x-1)S'^2_x+(n_y-1)S'^2_y}{n_x+n_y-2}}\sqrt{\left(\dfrac{1}{n_x}+\dfrac{1}{n_y}\right)}} = t_{n_x+n_y-2}$$

siendo $S_w=\sqrt{\dfrac{(n_x-1)S'^2_x+(n_y-1)S'^2_y}{n_x+n_y-2}}$ el estimador ponderado de σ

Varianzas poblacionales desconocidas y distintas

Los desvíos poblacionales desconocidos son reemplazados por los desvíos muestrales pero se obtiene una variable t de Student cuyos grados de libertad deben ser calculados:

$$t_v=\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\left(\dfrac{S'^2_x}{n_x}+\dfrac{S'^2_y}{n_y}\right)}}$$

5.-Distribución muestral de la diferencia de proporciones

Si de dos poblaciones independientes, cada una con distribución Binomial de parámetro π, se extrae una muestra, luego el estimador de la diferencia de proporciones poblacionales $\pi_1 - \pi_2$, será $p_1 - p_2$, de la cual se quiere determinar su distribución por muestreo:

$$P_1 \sim N\left(\pi_1 ; \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}\right)$$

$$P_2 \sim N\left(\pi_2 ; \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

$$E(p_1 - p_2) = \pi_1 - \pi_2$$

$$Var(p_1 - p_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

Luego si los tamaños de muestra son suficientemente grandes, la distribución de Δp por muestreo es aproximadamente Normal (por el Teorema del límite Central).

$$z = \frac{(p_1 - p_2) - \pi_1 - \pi_2}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$$

6.-Distribución muestral del cociente de varianzas

Suele ser de interés comparar la variabilidad de dos poblaciones, esto se puede realizar a través de la razón de varianzas muestrales. Si esta razón es cercana a la unidad, las variabilidades se puede decir que son casi equivalentes; Si por el contrario se aleja de uno, se dice que son no equivalentes; pero para que esta decisión pueda ser correcta, se deberá analizar la distribución de la razón de varianzas muestrales. Para esto se extrae una muestra aleatoria de tamaño n_x, de la primer población, constituida por variables independientes y distribuidas Normalmente, cada una con media μ_x y varianza σ^2_x; lo mismo se hace con la población dos, se extrae una muestra de extensión n_y de variables aleatorias independientes, cada una con media μ_y y varianza σ^2_y siendo X e Y independientes. Luego la distribución de cada varianza se vincula a la distribución χ^2 de la siguiente forma:

$$\frac{(n_x - 1)S'^2_x}{\sigma^2_x} = \frac{\sum_{i=1}^{n_x}(x_i - \bar{x})^2}{\sigma^2_x} \sim \chi^2 \text{ con } n_x - 1 \text{ grados de libertad.}$$

$$\frac{(n_y - 1)S'^2_y}{\sigma_y^2} = \frac{\sum_{j=1}^{n_y}(y_j - \bar{y})^2}{\sigma_y^2} \sim \chi^2 \text{ con } n_y - 1 \text{ grados de libertad.}$$

Al ser X y Y variables aleatorias independientes, entonces estas dos variables χ^2 también son independientes. De esta manera, el cociente de estas variables χ^2 origina une variable F de Snedecor, con n_x-1 y n_y-1 grados de libertad.

ESTIMACIÓN

Frecuentemente, los parámetros de las distribuciones son valores que se desconocen. Se busca, entonces, a partir de valores observados, estimar el o los valores desconocidos. Este procedimiento se denomina ***estimación de parámetros***.

Un estimador es una función de valores observados (muestra) que no depende de ningún parámetro desconocido. ***Un estimador es un estadístico*, *y* una *estimación es cualquiera* de sus *posibles valores*.**

Para estimar un parámetro pueden utilizarse distintos estadísticos (características de muestra). Es evidente que en calidad de estimación conviene tomar estadísticos cuyos valores, para distintas muestras de la población sean, por término medio, próximos al valor real del parámetro. También es deseable que con el aumento del tamaño de la muestra crezca la fiabilidad de la estimación.

Si se ha obtenido un estimador puntual, es conveniente tener una medida de precisión atribuida al estimador. La precisión de un estimador se mide por el error estándar del estimador. Es decir, cuanto menor sea este error, tanto más preciso será el estimador. Es bueno, entonces, que cuando se de una estimación, también se brinde el error estándar de la estimación.

Hay varios métodos para realizar estimación de los cuales se va a desarrollar uno de ellos.

Si se quiere una expresión más formal de la estimación y su precisión, se puede obtener lo que se denomina ***estimación por intervalos***. Es la estimación de un parámetro por un intervalo al azar, que se denomina ***intervalo de confianza***, cuyos extremos son funciones de las variables aleatorias observadas.

Se llama ***intervalo de confianza para el parámetro θ*** al intervalo (θ_1,θ_2) que contiene el valor real del parámetro, con una probabilidad dada 1-α, siendo ésta la probabilidad confidencial. Las cotas del intervalo, como se dijo, son funciones de las observaciones y, por lo tanto, son variables aleatorias. Es por esto que se dice que el intervalo de confianza "cubre" al parámetro que se estima con una probabilidad l- α, o bien, en el 100 (1- α) % de los casos. La elección de la

probabilidad confidencial se determina por las condiciones concretas; por regla general se utilizan 0.90, 0.95 y 0,99.

Esta formulación puede expresarse de forma general como sigue:

$$P\left(\left|\theta - \hat{\theta}\right| \leq k\,\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

1-α coeficiente de confianza; k constante no negativa que depende de la distribución por muestreo del estimador $\hat{\theta}$

Esta desigualdad puede escribirse de la siguiente manera:

$$P\left(-k\,\sigma_{\hat{\theta}} \leq \theta - \hat{\theta} \leq k\,\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

$$P\left(\hat{\theta} - k\,\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + k\,\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

Con esto puede obtenerse una expresión general de un estimador por intervalo de confianza simétrico, para un parámetro:

$$P\left(\hat{\theta} - k\,\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + k\,\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

Siendo $\hat{\theta} - k\,\sigma_{\hat{\theta}}$ = L límite inferior

$\hat{\theta} + k\,\sigma_{\hat{\theta}}$ = U límite superior

Esta expresión permite obtener intervalos de confianza para cualquier parámetro, sea la distribución del estimador simétrica o no.

La constante k depende de la distribución muestral del estimador y del valor de 1-α.

En la estimación por intervalos se desea obtener intervalos de poca amplitud, ya que esto hará más precisa la estimación. El ancho real de un intervalo es dictado por el coeficiente de confianza y por el tamaño de la muestra, entre otras cosas. Dados la extensión de la muestra y el error estándar del estimador, cuanto más corto es el intervalo, tanto menor es el nivel de confianza.

Es posible obtener el tamaño de muestra adecuado a la precisión y a la confianza con la cual se quiere trabajar:

$$\text{Error de estimación} = |z_{\left(\frac{\alpha}{2}\right)}| \frac{\sigma}{\sqrt{n}}$$

$$n = \frac{z^2_{\frac{\alpha}{2}} \cdot \sigma^2}{\text{Error}^2}$$

Esto permite variar el nivel de confianza sin aumentar el error de estimación, sólo variando el tamaño de muestra; o bien reducir el error de estimación sin variar el nivel de confianza.

INTERVALOS PARA PARÁMETROS

Intervalos para la media poblacional

a -- Población Normal con desvío parámetro conocido

El intervalo para la media poblacional se basa en el estimador media muestral. En este caso su distribución muestral es la siguiente:

$$x \sim N(\mu;\sigma)$$

$$\bar{x} \sim N\left(\mu;\frac{\sigma}{\sqrt{n-1}}\right)$$

Luego, estandarizando la variable media muestral se obtiene:

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = z \sim N(0,1)$$

De la expresión general de intervalos se deduce el intervalo correspondiente:

$$\theta = \bar{x} \qquad \sigma_\theta = \frac{\sigma}{\sqrt{n}} \qquad k = |z_{\left(\frac{\alpha}{2}\right)}|$$

$$\left(\bar{x} \pm |z_{\left(\frac{\alpha}{2}\right)}| \frac{\sigma}{\sqrt{n}} \right)$$

b - Población Normal con desvío parámetro desconocido Muestra grande(n > 30)

EL desvío poblacional que se desconoce se estima por S, con lo cual la distribución del estimador media muestral se transforma en:

$$\bar{x} \sim N\left(\frac{S}{\sqrt{n}} \right)$$

con lo cual el intervalo de acuerde a la expresión general es:

$$P\left(\bar{x} - z_{\left(\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\left(\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n}} \right) = 1 - \alpha$$

Recordar que si el muestreo es sin reposición de una población finita debe hacerse la corrección del error estándar de la media muestral.

c - Población Normal con desvío parámetro desconocido. Muestra chica (n < 30)

$$x \sim N(\mu; \sigma)$$

$$\bar{x} \sim N\left(\mu; \frac{S}{\sqrt{n-1}} \right)$$

$$\text{con lo cual } \frac{\bar{x} - 1}{\frac{S}{\sqrt{n-1}}} = t_{n-1}$$

y el intervalo partiendo de la expresión general será:

$$p\left(\bar{x} - t_{\left(1-\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{\left(1-\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n-1}}\right) = 1-\alpha \quad \text{ó bien}$$

$$p\left(\bar{x} - t_{\left(1-\frac{\alpha}{2}\right)} \frac{S'}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\left(1-\frac{\alpha}{2}\right)} \frac{S'}{\sqrt{n}}\right) = 1-\alpha$$

$$\left(\bar{x} \pm |t_{\left(1-\frac{\alpha}{2}\right)}| \frac{S'}{\sqrt{n}}\right)$$

Intervalo para la varianza poblacional

Para obtener un intervalo para el desvío poblacional se toma como estimador la varianza o desvío muestral. Recordar la distribución muestral de la varianza muestral:

$$\frac{(n-1)\, S'^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{ó bien}$$

$$\frac{n\, S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Luego el intervalo buscado será;

$$P\left(\frac{n S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{n S^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha$$

Un intervalo para el desvío poblacional se deduce del anterior por obtener la raíz cuadrada de todos los términos de le desigualdad.

Intervalo para la proporción poblacional

La estimación de la proporción poblacional π se basa en la proporción muestral p. Recordando su distribución muestral y suponiendo una extensión de muestra suficientemente grande, entonces:

$$p \sim N\left(\pi \; ; \; \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

El intervalo buscado, de acuerde e le expresión general, será

$$\left(p \pm |Z_{\left(1-\frac{\alpha}{2}\right)}| \sqrt{\frac{\pi(1-\pi)}{n}} \right)$$

En la expresión del desvío de p, se deberá sustituir el parámetro desconocido π por su estimador puntual p para poder obtener un valor. Además, si el muestreo es sin reposición y la población finita, deberá corregirse este desvío de acuerde a la siguiente expresión:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

Intervalo para la diferencia de medias poblacionales

a – Poblaciones Normales. Desvíos parámetros conocidos

$$X \sim N(\mu_x; \sigma_x) \qquad Y \sim N(\mu_y; \sigma_y)$$

$$\bar{x} \sim N\left(\mu_x; \frac{\sigma_x}{\sqrt{n_x}}\right) \qquad \bar{y} \sim N\left(\mu_y; \frac{\sigma_y}{\sqrt{n_y}}\right)$$

El estimador, es en este caso, la diferencia de medias muestrales. Recordar su distribución muestral.

$$\bar{x} - \bar{y} \sim N\left(\mu_x - \mu_y \; ; \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$$

$$Z = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} = N(0,1)$$

$$\theta = \mu_x - \mu_y \qquad \hat{\theta} = \bar{x} - \bar{y} \qquad \sigma_\theta = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

Esto permite obtener con la expresión general de intervalo, el buscado para este caso:

$$\left(\left(\bar{x} - \bar{y}\right) \pm \; | Z_{\left(1-\frac{\alpha}{2}\right)} | \; \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right) = 1 - \alpha$$

b- Poblaciones normales, desvíos poblacionales desconocidos pero supuestos iguales

Las consideraciones son similares el caso anterior, solo que ahora σ_x y σ_y se desconocen, pero se consideran iguales (esto debe ser verificado previamente). Para este caso, la distribución por muestreo de la diferencia de medias muestrales es la siguiente:

$$\frac{\bar{x} - \bar{y} - \left(\mu_x - \mu_y\right)}{S_w \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = \; t_{n_x + n_y - 2}$$

Con $S_w = \sqrt{\frac{\left(n_x - 1\right) S_x'^2 + \left(n_y - 1\right) S_y'^2}{n_x + n_y - 2}}$ ó $S_w = \sqrt{\frac{n_x S_x^2 + n_y S_y^2}{n_x + n_y - 2}}$

Por lo tanto el intervalo, de acuerde a la expresión general, es el siguiente:

$$\left(\left(\bar{x} - \bar{y}\right) \; \pm \; | t_{\left(1-\frac{\alpha}{2}\right)} | \; S_w \; \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right)$$

c- Poblaciones normales, desvíos poblacionales desconocidos pero distintos

Si los desvíos no pueden considerarse iguales, esto lleva a una nueva expresión de la variable:

$$\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{S_x'^2}{n_x}+\frac{S_y'^2}{n_y}}} = t_\nu \quad , \quad \nu = \frac{\left(\frac{S_x'^2}{n_x}+\frac{S_y'^2}{n_y}\right)^2}{\frac{\left(\frac{S_x'^2}{n_x}\right)^2}{n_x-1}+\frac{\left(\frac{S_y'^2}{n_y}\right)^2}{n_y-1}} - 2$$

Con lo cual, el intervalo es el siguiente:

$$\left((\bar{x}-\bar{y}) \pm |t_{\left(1-\frac{\alpha}{2}\ ,\ \nu\right)}| \sqrt{\frac{S_x'^2}{n_x}+\frac{S_y'^2}{n_y}} \right)$$

Intervalo para la diferencia de proporciones poblacionales

El estimador de $\Delta\pi$ es $\Delta p = p_1$ y p_2, siendo p_1 y p_2 proporciones muestrales obtenidas de muestras al azar independientes de cada una de las poblaciones, con n_1 y n_2 suficientemente grandes:

$$p_1 \sim N\left(\pi_1 \ ; \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}\right)$$

$$p_2 \sim N\left(\pi_2 \ ; \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

$$\Delta p \sim N\left(\pi_1-\pi_2 \ ; \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

Con lo cual el intervalo será:

$$\Delta p \pm |Z_{\left(1-\frac{\alpha}{2}\right)}| \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}$$

Notar que en el desvío de Δp es necesario reemplazar los valores desconocidos de π_1 y π_2 por sus estimadores puntuales p1, y p_2.

Intervalo para la razón de varianzas poblacionales

Sean X y Y dos variables aleatorias independientes, con distribución Normal. Si interesa obtener un intervalo para la razón de las varianzas poblacionales, esto se obtiene a partir de la razón de varianzas muestrales de la siguiente manera:

$$X \sim N(\mu_x, \sigma_x) \qquad Y \sim N(\mu_y, \sigma_y)$$

conocidos $\bar{x},\ \bar{y},\ S_x^2\ y\ S_y^2,\ n_x\ y\ n_y$, luego

$$F_{n_x-1;n_y-1} = \frac{\frac{S_x^2}{\sigma_x^2}}{\frac{S_y^2}{\sigma_y^2}}$$

Entonces, el intervalo para la razón de varianzas es:

$$F\left(\frac{S_x^2}{S_y^2}\ F_{1-\frac{\alpha}{2}\ ;\ n_y-1\ ;\ n_x-1} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq \frac{S_x^2}{S_y^2}\ F_{\frac{\alpha}{2}\ ;\ n_y-1\ ;\ n_x-1}\right) = 1-\alpha$$

El intervalo para la razón de desvíos se obtiene directamente por hallar la raíz cuadrada a todos los términos de la desigualdad anterior.

$$P\left(\frac{S_x}{S_y}\ \sqrt{\frac{1}{F_{\frac{\alpha}{2}\ ;\ n_x-1\ ;\ n_y-1}}} \leq \frac{\sigma_x}{\sigma_y} \leq \frac{S_x}{S_y}\ \sqrt{F_{\frac{\alpha}{2}\ ;\ n_x-1\ ;\ n_y-1}}\right) = 1-\alpha$$

En la tabla final de esta unidad puede observarse, en forma resumida, todo lo expresado anteriormente.

PROCEDIMIENTO GENERAL PARA DETERMINAR LAS COTAS DE LOS INTERVALOS

1-Identificar el estimador apropiado para el parámetro que se desea estimar

2-Determinar su distribución por muestreo.

3-De acuerdo a la expresión general, plantear el intervalo.

4-Sustituir en la desigualdad los valores obtenidos de la muestra.

5-Una vez obtenido el intervalo, se concluye diciendo que el intervalo hallado cubre el valor del parámetro desconocido con una confianza de (1-α) %.

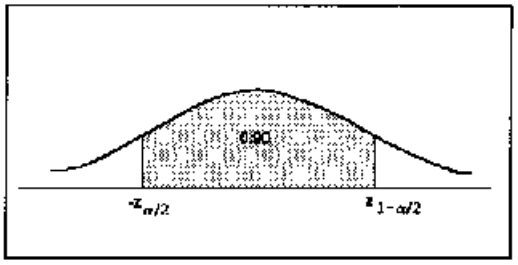
Ejemplo

-El tiempo de funcionamiento sin fallas de una máquina se sabe que es distribuido normalmente. Se realizaron 100 observaciones del tiempo sin fallas y se obtuvo un valor medio de 500 horas, conociendo que el desvío parámetro es de 10 horas. Se desea estimar, con 90% de confianza, el valor medio del tiempo de funcionamiento sin fallas.

$\bar{x} = 500 \ \ hs \ \ ; \ \ n = 100 \ \ ; \ \ \sigma = 10 \ \ hs$

$$\bar{x} \cong N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad ; \quad z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \quad ; \quad N(0,1)$$

$$P\left(\bar{x} - z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \ \leq \ \mu \ \leq \ \bar{x} + z_{\left(1-\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$



$$P\left(500 - 1.65 \cdot \frac{10}{\sqrt{100}} \ \leq \ \mu \ \leq \ 500 + 1.65 \frac{10}{\sqrt{100}}\right) = 0.90$$

$(498.35; 501.65)$ es posible decir que con 90% de confianza el valor media del tiempo de funcionamiento sin fallas se encuentra en este intervalo hallado