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.

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

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

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

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

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

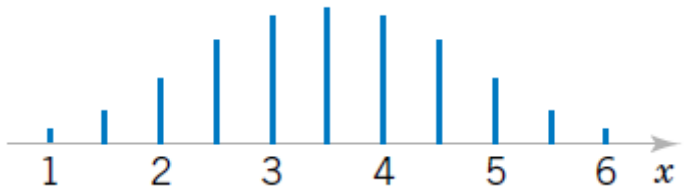
$$P\left(\frac{\bar{x} - E\left(\bar{x}\right)}{\sigma\left(\bar{x}\right)} \leq x\right) \rightarrow_{n\to\infty} N(0,1)$$

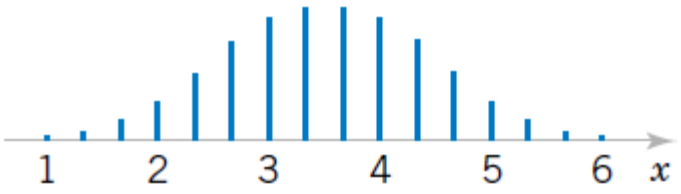
Se concluye, por lo tanto, que la variable aleatoria *media muestral* se distribuye normalmente con parámetros $E\left(\bar{x}\right)$ y $\sigma\left(\bar{x}\right)$

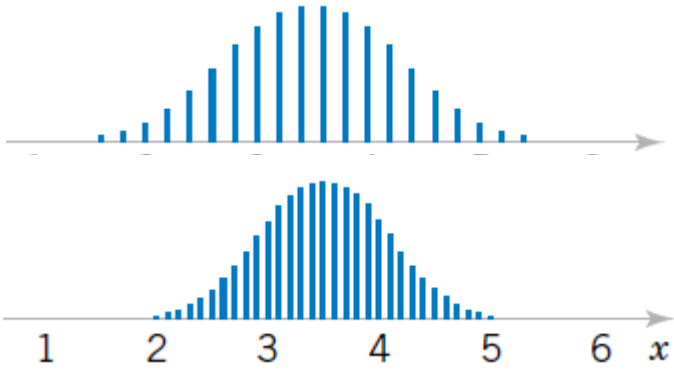
La conclusión es muy importante, ya que la mayor parte de los procedimientos de inferencia se basan en $\bar{x}$ .Si las variables que conforman la muestra se distribuyen normalmente, entonces la $\bar{x}$ también será distribuida normalmente, y así se puede

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

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

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

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:

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

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

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:

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

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:

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

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) %.

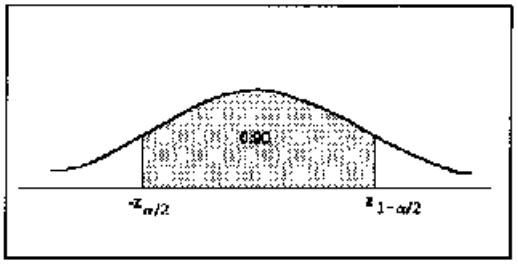
Ejemplo

-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

UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas

CÁTEDRA ESTADÍSTICA

Profesor Titular: Ing. Susana Vanlesberg

PRUEBAS DE HIPÓTESIS

Introducción

En la parte anterior hemos visto cómo construir una estimación a partir de un intervalo de confianza para un parámetro a partir de datos de la muestra. Sin embargo, muchos problemas de ingeniería requieren que decidamos si aceptar o rechazar un supuesto acerca de algún parámetro. El supuesto se denomina hipótesis, y el procedimiento de toma de decisiones acerca de la hipótesis se denomina pruebas de hipótesis. Este es uno de los aspectos más útiles de la inferencia estadística, ya que muchos problemas de toma de decisiones, pruebas o experimentos en el mundo de la ingeniería se pueden formular como problemas de prueba de hipótesis. Además, como veremos más adelante, hay una conexión muy estrecha entre pruebas de hipótesis e intervalos de confianza.

Prueba de hipótesis estadística y estimación de intervalos de confianza de los parámetros son los métodos fundamentales utilizados en la etapa de análisis de los datos de un experimento en el que el ingeniero está interesado, por ejemplo, en la comparación de la media de una población con un valor especificado.

Una hipótesis estadística es una declaración o supuesto acerca de los parámetros de una o más poblaciones.

En muchas circunstancias, las decisiones se deben tomar con base solo en la información de la muestra. Un gerente de control de calidad debe determinar si su

proceso funciona correctamente. Un director de una repartición provincial debe determinar si una nueva estrategia de manejo de recursos es adecuada para su provincia. Un ingeniero proyectista saber cuál es el valor de la precipitación media mensual en un lugar en el que se diseñará un desagüe.

El tomador de decisiones querría cerciorarse, hasta donde sea posible, de que ha llegado a la conclusión correcta.

Conceptos principales

En muchos casos, los resultados de observaciones se utilizan para verificar suposiciones, ***Hipótesis***, respecto a algunas propiedades de la distribución de una población.

Generalmente la distribución de la variable x se conoce y, basándose en la muestra de observaciones, es necesario comprobar la hipótesis sobre los valores de los parámetros de esta distribución. ***Estas hipótesis se denominan paramétricas.***

La hipótesis sujeta a verificación se denomina ***hipótesis nula*** H_0. Paralelamente con esta, se analiza la hipótesis denominada ***hipótesis alternativa*** H_1, que suele ser la opuesta a la hipótesis nula.

Por ejemplo, si se comprueba la hipótesis de que el parámetro θ es igual a cierto valor θ_0 (H_0: $\theta = \theta_0$) puede analizarse como hipótesis alternativa cualquiera de las siguientes:

a) H_1: $\theta > \theta_0$ **b)** H_1: $\theta < \theta_0$ **c)** H_1: $\theta \neq \theta_0$

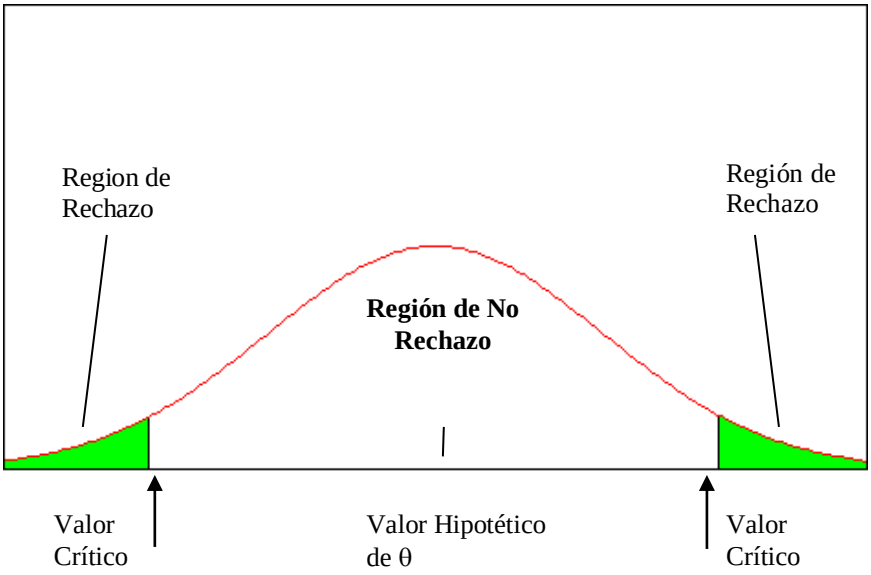
dependiendo de la formulación del problema.

Región de Rechazo y Región de Aceptación

Para cada tipo de procedimiento de prueba de hipótesis, se puede calcular una ***prueba estadística*** apropiada. Esta prueba estadística mide el acercamiento del valor de la muestra a la hipótesis nula.

La distribución apropiada de la prueba estadística se divide en dos regiones: una ***región de rechazo*** y una ***región de aceptación***.

Al tomar la decisión con respecto a la hipótesis nula, en la distribución estadística se debe determinar el ***valor crítico*** que divide la región de aceptación de la región de rechazo; este valor crítico depende del tamaño de la región de rechazo. El gráfico siguiente muestra las dos zonas.



ERRORES DE TIPOS I Y II

Este procedimiento de decisión puede conducir a una de dos conclusiones erróneas. Por ejemplo, dado el supuesto como verdad del valor del parámetro y formulado en la hipótesis nula, sin embargo, para los valores muestrales seleccionados al azar, podríamos observar un valor de la estadística de prueba que cae en la región crítica. Entonces podríamos rechazar la hipótesis nula hipótesis H_0 en favor de la alternativa cuando, de hecho, H_0 es realmente cierta. Este tipo de mal conclusión se llama un error de tipo I.

Rechazar la hipótesis nula H_0 cuando es verdadera se define como un error de tipo I.

Supongamos ahora que el verdadero valor del parámetro es diferente del valor plateado en H_0, sin embargo, el valor de la característica de muestra cae en la región de aceptación. En este caso queremos dejar de rechazar H_0 cuando es falsa.

Aceptar la hipótesis nula cuando en realidad es falsa se denomina error de tipo II.

Las probabilidades de cometer estos errores pueden considerarse como los riesgos de tomar decisiones incorrectas. La probabilidad máxima de cometer un error del tipo I se llama nivel de significación, y generalmente se representa con α:

$$\boldsymbol{\alpha} = \text{Máx P (I)} = \text{Máx P } (H_1/H_0) = \text{Máx P(rechazar } H_0/H_0 \text{ es verdadera)}$$

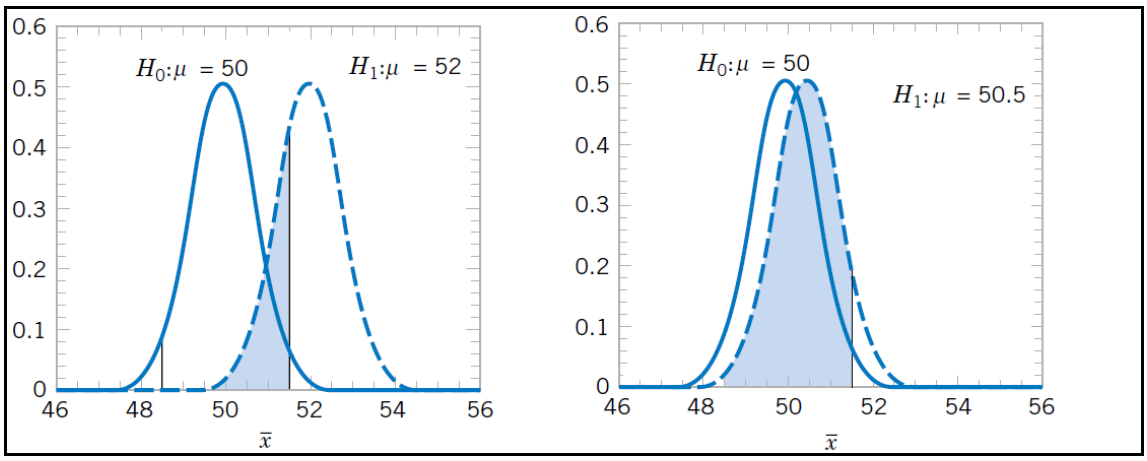
La probabilidad máxima de cometer un error del tipo II se denomina β:

$$\boldsymbol{\beta} = \text{Máx P (II)} = \text{Máx P } (H_0/H_1) = \text{Máx P(aceptar } H_0/H_0 \text{ es falsa)}$$

$$P(\alpha) = \int_{-\infty}^{k_c} f \left(k \ / \ H_0 \right) dk$$

$$P(\beta) = \int_{-\infty}^{k_c} f \left(k \ / \ H_1 \right) dk$$

Una buena prueba estadística es aquella en donde tanto α como β son pequeñas, porque permitirá tomar una decisión correcta, con menor riesgo de equivocarse.



Probabilidad de error tipo II para dos H_1 propuestas

1.- El tamaño de la región crítica, y en consecuencia la probabilidad de un error de tipo I, siempre se puede reducir mediante la selección apropiada de los valores críticos.
2.- Errores tipo I y tipo II están relacionados. Una disminución en la probabilidad de un tipo de error siempre resulta en un aumento en la probabilidad de la otra, siempre que el tamaño de la muestra no cambie.
3.- Un aumento del tamaño de la muestra generalmente reduce tanto α como β siempre que los valores críticos se mantengan constantes.
4.- Cuando la hipótesis nula es falsa β aumenta a medida que el verdadero valor del parámetro se acerca al valor propuesto en la hipótesis nula.

En general, el analista controla la probabilidad de error tipo I. Dado que el analista puede controlar directamente la probabilidad de rechazar erróneamente H_0, siempre pensamos en el rechazo de la hipótesis nula H_0 como una fuerte conclusión.

Por otra parte, la probabilidad de error tipo II β no es una constante, sino que depende del verdadero valor del parámetro. También depende del tamaño de la muestra que hemos seleccionado.

Debido a que la probabilidad de error de tipo II, β, es una función tanto del tamaño de la muestra y la medida en que la hipótesis nula H_0 es falsa, se acostumbra a pensar en la decisión de aceptar H_0 como débil conclusión, a menos que sepamos que β es aceptablemente pequeña. Por lo tanto, en lugar de decir que "Aceptamos H_0", preferimos la terminología " No se rechaza H_0." No rechazar H_0 implica que no se ha

encontrado evidencia suficiente para rechazar H_0, es decir, para hacer una declaración fuerte. El no poder rechazar H_0 no significa necesariamente que hay una alta probabilidad de que H_0 sea cierta. Se puede simplemente significar que se necesitan más datos para llegar a una conclusión firme. Esto puede tener importantes implicaciones para la formulación de hipótesis.

Un concepto importante es la potencia de una prueba estadística.

La potencia de una prueba estadística es la probabilidad de rechazar la hipótesis nula H_0 cuando la hipótesis alternativa es verdadera.

La potencia se calcula como **1-β**, y puede ser interpretada como la probabilidad de rechazar correctamente una hipótesis nula falsa.

Para una probabilidad dada del error del primer género α, la del error de segundo género β puede disminuirse aumentando el volumen de la muestra.

El error α está bajo control del investigador y se elige o establece antes de realizar la prueba de hipótesis, es el nivel de significación para la prueba de hipótesis, entonces como se puede controlar α , también 1-α está controlada.

Nivel de significación de la prueba

La probabilidad α es el nivel de significación de la prueba, es el riesgo o la probabilidad que el investigador asume de manera voluntaria de equivocarse al rechazar la hipótesis nula, cuando en realidad es verdadera.

Es también la confiabilidad de decidir si se rechaza o no la hipótesis nula. Los niveles de significación más usados son: 0,01, 0,05 y 0,10.

Cuando se rechaza la hipótesis nula, se dice que hay significancia estadística, pero cuando no se rechaza la hipótesis nula significa que "no existe suficiente información como para rechazarla", es errado afirmar que se acepta la hipótesis nula. No se puede aceptar algo que no sabemos si es verdadero o falso.

Que una prueba sea estadísticamente significativa, es decir, rechazar la hipótesis nula no asegura que la hipótesis alternativa sea cierta ante la evidencia de datos muestrales, sino que los datos muestrales discrepan con el supuesto bajo la hipótesis nula.

Recordar que la muestra es aleatoria, también lo son los estadísticos que se usan para someter a prueba hipótesis estadísticas.

Por tanto se recomienda no ser mecanicistas y estar dependiendo del valor α porque lo estadísticamente significativo no siempre es relevante para la investigación.

Ahora se usan los softwares estadísticos como SPSS, MINITAB, SAS, entre otros y es preocupante ver como se los usa de manera indiscriminada, sin sustento, se cree que es solo poner los datos y ver el resultado si es o no estadísticamente significativo. No hay que contentarse con que sea estadísticamente significativo sino que sea relevante para la investigación.

Estadístico de prueba

Para rechazar o no la hipótesis nula se toma una muestra aleatoria de la población bajo estudio y los resultados contenida en ella se usa en expresiones llamadas estadísticos o estadísticas de prueba e indican el grado de discrepancia entre la hipótesis nula y los datos muestrales que están resumidos en las estadísticas.

Cuando la discrepancia es "grande", es decir la evidencia de la muestra (datos muestrales) difiere del valor supuesto para el parámetro bajo la hipótesis nula; se rechaza la hipótesis nula en caso contrario no se rechaza.

La regla de acuerdo a la cual se toma la decisión de aceptar o rechazar la hipótesis H_0 se llama ***criterio*** *k.* Ya que la decisión se adopta basándose en la muestra de observaciones de la variable aleatoria, es necesario elegir un estadístico adecuado que se denomina, en este caso, ***estadístico del criterio o estadístico de prueba***. Se trata, en general, de que el estadístico de prueba sea uno cuya distribución en el muestreo se conozca, en el supuesto de que H_0 sea cierta. Generalmente resulta el estimador del parámetro que se quiere probar en la hipótesis nula.

Región de rechazo

Al conjunto de valores de la estadística de prueba para los que la hipótesis nula se rechaza se llama "región de rechazo o región crítica".

El establecimiento de la región de rechazo depende de la distribución de probabilidad de la estadística de prueba, el punto de corte (punto o valor que divide a la región crítica de la no crítica) se llama también "valor crítico o punto crítico", cuyo su valor depende de la distribución de probabilidad de la estadística de prueba.

El conjunto de todos los valores del estadístico del criterio para los cuales se toma la decisión de rechazar H_0 se denomina **dominio crítico**. El conjunto contrario se llama **dominio de aceptación.**

El valor del estadístico de prueba que separa a ambos conjuntos se llama **valor crítico del estadístico de prueba θc.** Este valor depende del tamaño de la muestra, de α, de la forma de H_1 y de la distribución del estadístico.

Nivel crítico de una prueba de hipótesis (p-value)

Una forma de comunicar los resultados de una prueba de hipótesis consiste en afirmar que la hipótesis nula es aceptada o no a un valor determinado α o nivel de significación.

El valor P es el nivel mínimo de significación que conduciría a un rechazo de la hipótesis nula H_0 con los datos dados.

Es costumbre decir que la estadística de prueba es significativa cuando la hipótesis nula H_0 se rechaza, por lo tanto, podemos pensar en el valor **p** como el mínimo valor en el que los datos son significativos.

El valor *p* es la probabilidad de obtener un estadístico de prueba igual o más exacto que el resultado obtenido a partir de los datos de la muestra, dado que la hipótesis nula es verdadera.

A menudo, al valor *p* se lo conoce como *nivel de significación observado,* que es el mínimo nivel al cual H_0 puede ser rechazada.

- **Si el valor *p* es menor que α, H_0 es rechazada.**
- **Si el valor *p* es mayor o igual que α, H_0 no es rechazada.**

Decisiones posibles en un test de hipótesis

	Aceptar H0	**Rechazar H0**
θ En la región de aceptación	No hay error	Error del tipo I
θ En la región de rechazo	Error del tipo II	No hay error

PASOS PARA LA VERIFICACIÓN DE UNA PRUEBA DE HIPÓTESIS

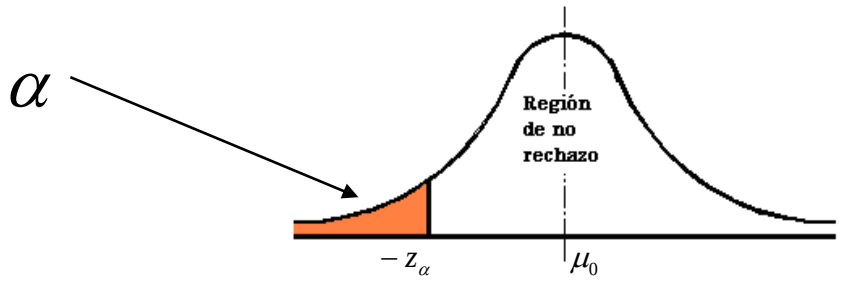
1) Expresar la hipótesis nula.
2) Expresar la hipótesis alternativa.
3) Especificar el nivel de significación, (α).
4) Determinar el tamaño de la muestra.
5) Establecer los valores críticos que dividen las regiones de rechazo y de no rechazo.
6) Seleccionar la prueba estadística según la distribución por muestreo del estadístico que se elige para llevar adelante la prueba referida al parámetro.
7) Coleccionar los datos y calcular el valor muestral de la prueba estadística.
8) Determinar si la prueba estadística ha caído en la región de rechazo o en la de aceptación.
9) Tomar la decisión.
10) Expresar la decisión estadística en términos del problema.

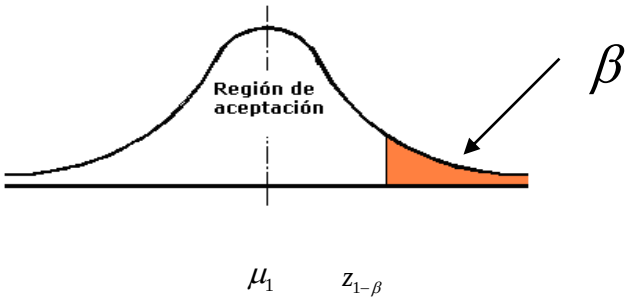
Observación: En las etapas 5 a 8 son utilizados estadísticos cuyos cuantiles se encuentran tabulados (Normal, Student, Chi-cuadrado, F de Snedecor).

La verificación de las hipótesis estadísticas puede realizarse basándose en los intervalos de confianza. En este caso, la hipótesis propuesta se aceptará si el valor θ_0 es cubierto por el intervalo respectivo.

Determinación del tamaño de muestra sin cambiar las probabilidades de los errores de Tipo I y II (válido para cualquiera de los casos que se presentarán)

Frente a α y β predeterminados, es posible encontrar un valor de **n** apropiado.
Supóngase una prueba unilateral por izquierda correspondiente al parámetro μ, con α prefijado y ubicado de acuerdo a H_1:





Luego, H_0 se rechazará si:

$$\bar{x} < \bar{x}_c = \mu_0 + Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

Para el β prefijado, β es la probabilidad de aceptar H_0 cuando lo es H_1; entonces se aceptará H_0 con una probabilidad de (1 - α) sólo si:

$$\bar{x} > \bar{x}_c = \mu_1 + Z_{1-\beta} \cdot \frac{\sigma}{\sqrt{n}}$$

De esta dos últimas ecuaciones puede obtenerse **n**, y luego de cualquiera de ellas el $\bar{x}_c$.

Si se quiere verificar que con **n** y $\bar{x}_c$ determinados anteriormente se mantienen los valores de α y β prefijados, puede realizarse lo siguiente:

$$\alpha = P\ \ (rechazar\ \ \ H_0 / siendo\ H_0\ verdadera) = P\ \ (\bar{x} < \bar{x}_c / \mu = \mu_0) = P\ \left(\frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \leq \frac{\bar{x}_c - \mu_0}{\sigma_{\bar{x}}} \right)$$

$$\beta = P\ \ (aceptar\ \ \ H_0 / siendo\ H_1\ verdadera) = P\ \ (\bar{x} \geq \bar{x}_c / \mu = \mu_1) = P\ \left(\frac{\bar{x} - \mu_1}{\sigma_{\bar{x}}} \leq \frac{\bar{x}_c - \mu_1}{\sigma_{\bar{x}}} \right)$$

Los cuales deberían ser iguales a α y β prefijados.

DISTINTOS CASOS DE PRUEBAS PARAMÉTRICAS

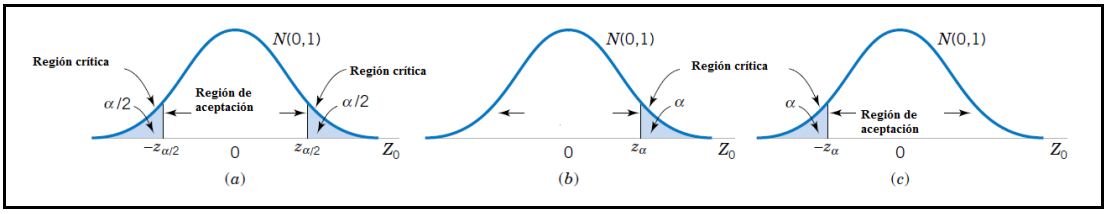
Pruebas de hipótesis referidas a la media de una distribución Normal

Desvío Poblacional Conocido

a. Formulación de las hipótesis.
b. Fijar el nivel de significación; dependiendo del problema tomar 5% o 1%
c. Elegir el estadístico de prueba: en este caso la media muestral.
d. Distribución muestral del estadístico. El estadístico de prueba será:

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

e. Establecer las zonas de aceptación y rechazo de acuerdo a como fue formulada H_1.
f. Calcular el estadístico de prueba para la muestra en consideración. De acuerdo a su valor, se toma la decisión según haya caído en una u otra zona.



Desvío Poblacional Desconocido. Muestra Grande (n>30)

El planteo general es similar al caso anterior, solo que la distribución muestral del estadístico media muestral es ahora $N(\mu, S/n^{1/2})$ ya que el desvío es desconocido, pero como la muestra es grande es posible estimarlo por el desvío muestral S.

$$Z = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$$

Desvío Poblacional Desconocido. Muestra Chica (n<30)

El planteo es nuevamente similar a los casos anteriores; la diferencia está en que el estadístico media muestral se distribuye ahora según una t de Student, ya que como la muestra es chica no se puede considerar al desvío poblacional desconocido igual a S.

$$t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n-1}}}$$

En este caso los puntos críticos corresponden a la distribución muestral t con n-1 grados de libertad por tabla o programa.

La decisión se realiza de acuerdo a las zonas de aceptación y rechazo y al valor del estadístico de prueba.

Prueba acerca de la varianza de la población

Recordar que la distribución muestral de la varianza es chi-cuadrado:

$$\chi^2 = \frac{(n-1)S'^2}{\sigma_0^2}$$

Prueba referida a una proporción poblacional

La distribución muestral de la proporción es Normal:

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Con:
p = x/n número de éxitos en la muestra / tamaño de la muestra
$\pi =$ Proporción de éxitos en la población.

PRUEBAS DE HIPÓTESIS RESPECTO A DOS PARÁMETROS

Prueba acerca de la igualdad de varianzas de dos poblaciones

Cuando es de interés comparar dos poblaciones, generalmente cuando las muestras son pequeñas, independientes, y provenientes de poblaciones normales o aproximadamente normales, se utiliza la razón de varianzas muestrales como base para esta comparación.

Este procedimiento tiene las siguientes ventajas: La razón es independiente de las unidades (siempre que ambas tengan las mismas unidades), la distribución de esta razón debido a la hipótesis de igualdad de varianza poblacionales, es independiente de los parámetros de las poblaciones.

La razón de varianzas muestrales tiene una distribución F de Snedecor ya que cada una de las varianzas esta relacionada a una variable chi-cuadrado:

$$F_{\nu_1;\nu_2} = \frac{\frac{(n_1-1)\hat{S'}_1^{\,2}}{\sigma_1^2}}{\frac{(n_2-1)\hat{S'}_2^{\,2}}{\sigma_2^2}}$$

es el estadístico de la prueba.

Generalmente las pruebas son unilaterales y por derecha ya que la hipótesis nula será rechazada para valores de F muy grandes debido a que en el cálculo se coloca la varianza mayor en el numerador.

Prueba respecto a las medias de dos poblaciones Normales. Desvíos parámetros conocidos

Las hipótesis a formular pueden ser las siguientes:

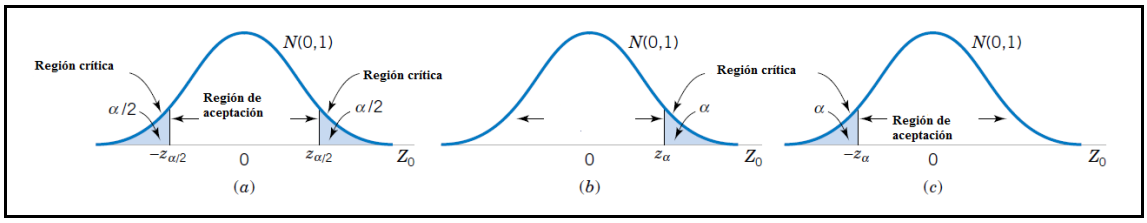
$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$

$H_0 : \mu_1 > \mu_2 \quad H_1 : \mu_1 \leq \mu_2$

$H_0 : \mu_1 < \mu_2 \quad H_1 : \mu_1 \geq \mu_2$

$H_0 : \mu_1 \leq \mu_2 \quad H_1 : \mu_1 > \mu_2$

$H_0 : \mu_1 \geq \mu_2 \quad H_1 : \mu_1 < \mu_2$



Se fijan luego las probabilidades de los errores de primero y segundo género.

El estadístico a utilizar es la diferencia de medias muestrales, cuya distribución por muestreo es para este caso:

$$\bar{x}-\bar{y} \sim N\left(\mu_x-\mu_y \; ; \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

Con lo cual el estadístico de la prueba será:

$$z=\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N\left(0 \;,1\right)$$

Este valor una vez calculado deberá compararse con el valor tabulado que surge de la ubicación de las zonas de aceptación y rechazo. Así se puede tomar la decisión de aceptar o rechazar lo propuesto.

Prueba respecto a las medias de dos poblaciones Normales
Desvío poblacionales desconocidos-Muestras grandes (n_x y/o $n_y \geq 30$)

Los pasos son similares al caso anterior, solo que la distribución muestral del estadístico diferencia de medias muestrales cambia al ser las varianzas poblacionales desconocidas y ser estimadas por las muestrales:

$$\bar{x}-\bar{y} \sim N\left(\mu_x-\mu_y \; ; \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)$$

Luego el estadístico de la prueba en base al cual se tomará una decisión será:

$$z=\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim N\left(0 \;,1\right)$$

Muestras pequeñas (n_x y/o $n_y < 30$)

Previamente debe verificarse si las varianzas que se desconocen pueden considerarse iguales o no.

Varianzas desconocidas e iguales

En este caso la distribución del estadístico diferencia de medias muestrales es una t de Student.

Debido a que las varianzas muestrales no son estimadores insesgados de las varianzas poblacionales por el tamaño de las muestras:

$$\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{S_w\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \sim \quad t_{n_x+n_y-2}$$

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

El resto de los pasos para obtener una conclusión correcta respecto a la hipótesis planteada son similares a los casos anteriores.

Varianzas desconocidas y distintas

En este caso la distribución muestral del estadístico diferencia de medias muestrales es una t de Student pero se calculan los grados de libertad:

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

Se adopta como grados de libertad el entero más próximo al valor anterior.

El resto de los pasos para tomar una decisión respecto a la hipótesis planteada es similar a los casos anteriores.

Prueba respecto a proporciones de dos poblaciones Normales

Una vez establecidas las hipótesis nulas y alternativas que corresponden, fijados α y β se busca el estadístico y su distribución muestral:

$$p_x - p_y \sim N \left(\pi_x - \pi_y \;;\; \sqrt{\frac{\pi_x(1-\pi_x)}{n_x} + \frac{\pi_y(1-\pi_y)}{n_y}} \right)$$

Debido a que las proporciones poblacionales se desconocen se toma como estimador de ellas la media ponderada de las proporciones muestrales, resultando el siguiente estadístico de prueba:

$$z = \frac{p_x - p_y - (\pi_x - \pi_y)}{\sqrt{\hat{p}(1-\hat{p}) \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \quad ; \quad con \quad \hat{p} = \frac{n_x \cdot \hat{p}_x + n_y \cdot \hat{p}_y}{n_x + n_y}$$

Ejemplo

Para determinar que influencia ejerce la temperatura del medio ambiente en el error sistemático de un instrumento que mide ángulos, se han efectuado mediciones del ángulo horizontal de un objeto, durante la mañana (t = 10° C) y durante el resto del día (t = 26° C). Los resultados de las mediciones del ángulo (en segundos angulares) son:

De mañana	38.2	36.4	37.7	36.1	37.9	37.8		
De tarde	39.5	38.7	37.8	38.6	39.2	39.1	38.9	39.2

Si los valores provienen de una distribución normal y conociendo la distribución de los errores, ¿se puede considerar que la temperatura ambiente influye en el error sistemático del instrumento?

$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$

$\sigma_1 \;\; y \;\; \sigma_2 \;\; desconocidos, \quad n_1 \;\; y \;\; n_2 \;\; < \;\; 30$

Corresponde previamente realizar el test de igualdad de varianzas, porque de este dependerá la prueba de igualdad de medias.

$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 \neq \sigma_2^2$ **α=0,10**

$\overline{x_1} = 37.35 \quad \overline{x}_2 = 38.87 \quad S_1 = 0.636 \quad S_2 = 0.628 \quad S'^2_1 = 0.76 \quad S'^2_2 = 0.72$

$$F = \frac{S'^2 \;\; mayor}{S'^2 \;\; menor} \qquad F_c = \frac{0.76}{0.72} = 1.05$$

$F_{(5;7;0.05)} = 3.97 \quad F_c < F_{tabulado}$

Según estos valores obtenidos las varianzas pueden considerarse iguales. De acuerdo a esto se plantea el test de igualdad de medias y se verifica la misma.

Prueba respecto a las medias de dos poblaciones con muestras dependientes

En los casos anteriores, se compararon medias de dos poblaciones y se suponía que las muestras que se extraían eran de poblaciones Normales e independientes. Esto significa que cada una no se relaciona con algún elemento de la otra de forma significativa.

Dependencia significa que están emparejadas en el sentido que cada observación en una se asocia con alguna observación en la otra. **Si dos muestras son dependientes deberán tener el mismo número de elementos.**

Hay situaciones en las que puede haber necesidad de comparar dos procedimientos diferentes y en las que se sospecha un cambio de nivel en la variable de respuesta, que no se puede controlar. Ya que el nivel de la medición puede variar de muestra a muestra es posible que las dos observaciones no sean independientes, por el contrario que estén relacionadas. Por ejemplo: si se tienen dos instrumentos para medir precipitación, se quiere saber si son o no equivalentes los resultados de ambos, problema muy común en el caso de comparación de datos.

En esta situación la prueba de comparación de medias se realiza de forma similar a los casos anteriores. Para derivar el estadístico de prueba adecuado se obtienen la diferencia entre las observaciones emparejadas para cada elemento d_i, que pertenecen a una nueva variable aleatoria **D**: diferencia en la población.

Podría decirse si las dos muestras no son diferentes entre sí, entonces el valor medio de **D** sería igual a 0. Un supuesto importante a realizar es que las diferencias se distribuyen normalmente.

Es posible a partir de los datos obtener la media de los n pares formados d_i; el valor del error estándar de esta media será:

$$\frac{S_d}{\sqrt{n-1}}$$

Y la razón de $\bar{d}$ a su error estándar se distribuye según una t de Student con n – 1 grados de libertad:

$$t_{n-1} = \frac{\bar{d} - \bar{D}}{\frac{S_d}{\sqrt{n-1}}}$$

con $\overline{D} = 0$.

Este es el estadístico de la prueba. Los pasos posteriores son similares a los casos ya desarrollados; es decir una vez calculado el valor del estadístico de prueba se deberá ubicar en que región cae, si en la de aceptación o la de rechazo.

RELACIÓN ENTRE LOS INTERVALOS DE CONFIANZA Y LOS TEST DE HIPOTESIS

Los intervalos de confianza se utilizan para estimar parámetros, mientras que los test de hipótesis se utilizan para tomar decisiones en relación a parámetros de la población especificada.

En muchos casos los intervalos de confianza y los test de hipótesis se pueden usar en forma indistinta.

Si al calcular el intervalo para el parámetro, no se rechazaría la hipótesis nula planteada si el intervalo contiene el valor que se quiere probar.

TEST NO PARAMÉTRICOS

Usos de la distribución Chi-cuadrado

En lo presentado hasta este punto se desarrollaron métodos referidos a test de hipótesis paramétricos, es decir donde la hipótesis H_0 era referida al valor de un parámetro dado θ.

Existe una clase de problemas en la cual la proposición a testar se refiere a la forma de la distribución postulada F(x). Estas hipótesis se denominan *no paramétricas*.

La distribución chi-cuadrado que se ha utilizado como base en las inferencias relativas a la varianza poblacional, tiene otros usos:

- Comprobar si un conjunto de datos observados coincide o no con un conjunto de datos esperados (Prueba de frecuencia);
- Verificar la ley de distribución de una población (Prueba de la bondad del Ajuste)
- Comparar más de dos varianzas (Prueba de homogeneidad);
- Verificar la independencia de dos variables aleatorias (Pruebas e independencia)

En estos test no paramétricos la hipótesis alternativa H_1 no está especificada explícitamente. El nivel de significación α (probabilidad de cometer un error de tipo I), es elegido como una cantidad de decisión y depende de que la probabilidad de obtener el evento observado sea mayor o menor que α, para aceptar o no la hipótesis propuesta; por ésta razón es que son test unilaterales por derecha.

Un valor de α pequeño resultará en un test menos crítico o sea es muy probable que se acepte H_0.

PRUEBA DE FRECUENCIAS

En este caso se desea comprobar si un conjunto de datos coincide o no con un conjunto de datos esperados o frecuencias teóricas.

El estadístico de prueba será:

$$\chi^2_{n-k} = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}$$

De igual forma que en otros ensayos se fijará un riesgo α y se comparará el estadístico calculado con un valor tabulado.

Siendo fo la frecuencia de realización de un acontecimiento determinado, de acuerdo a H_0, y fe la frecuencia esperada.

La suma de los cocientes sigue aproximadamente la distribución chi-cuadrado si es que no hay diferencias entre las frecuencias esperadas y las observadas.

Se establece que fe = n*pi, frecuencias esperadas, deben ser mayores o iguales que 5 para que el estadístico siga una distribución chi-cuadrado, si esto no sucede, se las puede combinar hasta lograrlo.

Cuando los grados de libertad que resultan luego del agrupamiento, que son iguales a n-1, son iguales a 1, se deberá realizar una corrección restando ½ a la diferencia del denominador. Se aplica cuando **n** es pequeño, ya que si **n** es grande su efecto es despreciable:

PRUEBA DE BONDAD DE AJUSTE

En este punto se combinarán algunas ideas presentadas previamente respecto a la naturaleza de los datos observados con las técnicas de test de hipótesis ya desarrolladas. En este caso la hipótesis nula postula que x tiene la ley de distribución F(x).

Generalmente lo que se postula es la *forma* de la distribución, pero se deben estimar los parámetros a partir de los datos; el efecto de esta estimación es reducir los grados de libertad del estadístico.

La aplicación del criterio χ^2 para verificar la hipótesis nula se compone de las siguientes etapas:

a- Basándose en la muestra de observaciones de la variable aleatoria x, es necesario hallar las estimaciones de los parámetros desconocidos de la ley de distribución hipotética F(x).

b- Si x es una variable aleatoria discreta, hay que determinar las frecuencias fi con la cuales cada valor o grupo de valores se encuentra en la muestra.

Si x es una variable aleatoria continua, es necesario partir el dominio de sus valores en r intervalos disjuntos: Δ_1, Δ_2, Δ_3,…. Δ_r y determinar el número de elementos de la muestra, fi, que pertenecen a cada intervalo.

c- Si x es una variable discreta utilizando la ley de distribución hipotética F(x), es necesario calcular las probabilidades Pk, con las cuales la variable aleatoria x toma cada valor.
Si x es una variable aleatoria continua conviene determinar la probabilidad P_k de tomar un valor perteneciente a cada intervalo Δ.

d – Calcular el valor muestral del estadístico:

$$\chi^2_m = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}$$

e – Tomar la decisión estadística: La hipótesis nula no contradice a la muestra para el nivel de significación α, si $\chi^2_m < \chi^2_{1-\alpha/2\ (k-L-1)}$, siendo L el número de parámetros de la distribución F(x) que se estiman con ayuda de la muestra; en caso contrario se rechaza la hipótesis propuesta.

CASOS ESPECÍFICOS

Variable discreta

Modelo Binomial

Los pasos generales son:

a) H_0: f(x) = Binomial:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Basados en la muestra, será necesario estimar el parámetro p del modelo.

b) Determinar, en base a los datos, las frecuencias observadas.
c) Utilizando el modelo postulado, se obtienen las probabilidades para cada valor de la variable.
d) Se obtienen las diferencias entre las frecuencias observadas y las frecuencias esperadas, siendo estas últimas $f_e = n.p_i$ para obtener el valor del estadístico del test.

e) Se obtiene de tablas el valor de χ^2 crítico, de acuerdo a $\nu = k - L - 1$ y α. Concluir en base a estos valores, si se acepta o no que los datos se ajustan al modelo postulado.

Modelo de Poisson

Con un ejemplo, se desarrollarán los pasos para desarrollar el test.

Considere la ocurrencia de tormentas con granizo que se presentan en una región de la provincia de Santa Fe, mostradas en la siguiente tabla:

N° de tormentas por año	0	1	2	3	4	5	6
N° de ocurrencias observadas	102	144	74	28	10	2	0

La hipótesis propuesta es que las tormentas que ocurren se consideran independientes, tienen un promedio de ocurrencia, y estas ocurrencias son de tipo Poisson, de parámetro **λ**

$$f(x)= P(X = x)= \frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

N° de tormentas por año	N° de ocurrencias observadas		$P(X = x) = \frac{e^{-\lambda}\,\lambda^{x}}{x!}$	Fe=n*P		$\frac{(fo - fe)^2}{fe}$
0	102		0.3073	110.62		0.67
1	144		0.3626	130.53		1.39
2	74		0.2139	77.01		0.12
3	28		0.0841	30.29		0.17
4	10	12	0.025	8.94	11.46	0.025
5	2		0.0059	2.11		
6	0		0.0011	0.41		
				359.91		χ^2_{calc} =2.375

Se acepta la hipótesis propuesta ya que el valor calculado es menor al de tabla que es $\chi^2_{5\text{-}1\text{-}1;\ 0,05}$=7.81

Variable contínua

Modelo Normal

Los pasos a seguir son similares a los casos anteriores, sólo que para obtener las frecuencias esperadas se deberá obtener la probabilidad en un intervalo, de acuerdo a lo propuesto en H_0, en base al modelo Normal.

A partir de los datos de precipitación en 55 días de un año, se quiere comprobar la hipótesis nula que pertenecen a una población con distribución Normal.

17	19	23	18	21	15	16	13	20	18	15
20	14	20	16	14	20	19	15	19	16	19
15	22	21	12	10	21	18	14	14	17	16
13	19	18	20	24	16	20	19	17	18	18
21	17	19	17	13	17	11	18	19	19	17

Ancho de intervalo: c=2; n=55 ; $\bar{x}$ =17.84 mm
S = 2.92 mm

Intervalo	f.obs.		Li est.	Ls est.	F(Li)=a	F(Ls)=b	P=b–a	Fe= n.P	$\frac{(fo-fe)^2}{fe}$
10-12	2	6	- ∞	-2	0	0.0228	0.0228	1.254	0.145
12-14	4		-2	-1.32	0.0228	0.0934	0.0706	3.883	
14-16	8		-1.32	-0.63	0.0934	0.2643	0.1709	9.39	0.21
16-18	12		-0.63	0.055	0.2643	0.5199	0.2556	14.06	0.30
18-20	16		0.055	0.74	0.5199	0.7704	0.2505	13.78	0.36
20-22	10	13	0.74	1.42	0.7704	0.9222	0.1518	8.35	0.011
22-24	3		1.42	∞	0.9222	1	0.0778	4.28	
									χ^2_{calc} =1.026

Ya que $\chi^2_m < \chi^2_t$, la hipótesis sobre distribución de los datos no contradice los resultados observados.

TEST DE BONDAD DE AJUSTE DE KOLMOGOROFF

Otro test cuantitativo que se utiliza para verificar si los datos se ajustan a un modelo propuesto es el de Kolmogoroff.

Primero se ordena la muestra:

$x(1) \leq x(2) \leq \ldots.. \leq x(n)$

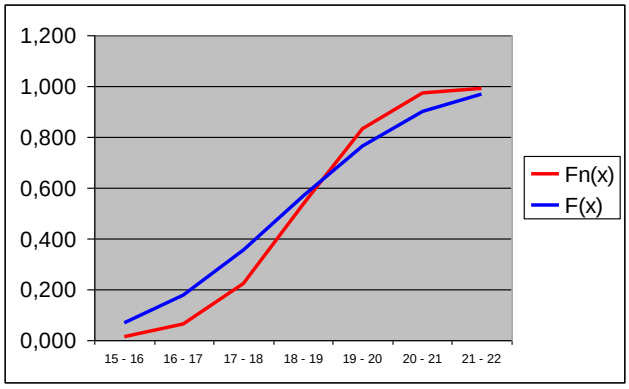
Luego puede definirse la función de cuantía muestral considerando la proporción de valores que no exceden a un valor dado x, $G_n(x)$: r/n.

Ya que r s el número de observaciones que son menores o iguales que x_r, r puede ser considerado como la suma de variables Bernoulli con $p = F(x_r)$.

El Teorema del Límite Central implica que la distribución límite de r/n es Normal, con valor esperado p.

Es así que el test para el ajuste puede ser construido sobre una medida apropiada de la desviación de $G_n(x)$ de F(x). Esta medida es el valor absoluto de la máxima diferencia entre $G_n(x)$ y F(x), y es lo que se define como el *estadístico de Kolmogoroff*.

$$D_n = máx\left|G_n(x) - F(x)\right|$$



La distribución de este estadístico no depende de F(x):

$$\lim_{n \to \infty} \qquad P\left(D_n > \frac{C}{\sqrt{n}}\right) = 2\sum (-1)^{m-1} e^{-2m^2C^2}$$

este test es independiente de la distribución.

Han sido tabuladas, para distintos valores de n, las probabilidades exactas
P ($D_n > C/n^{1/2}$) o su complemento, así como los valores críticos $\mathbf{D_n}$ para los cuales la probabilidad es 0.01 ó 0.05. Dos valores críticos asintóticos muy importantes son:

$$\alpha = 0.01 \qquad d_n = \frac{1.63}{\sqrt{n}}$$

$$\alpha = 0.05 \qquad d_n = \frac{1.36}{\sqrt{n}}$$

Los cálculos para desarrollar el test son muy simples, casi siempre un examen preliminar de los datos puede revelar áreas para las cuales las desviaciones son probablemente mayores. Una vez obtenido el valor absoluto de la máxima diferencia, se obtiene el valor crítico del estadístico tabulado para tomar una decisión: aceptar o no el modelo propuesto.

La comparación entre los dos test permite analizar los siguientes puntos:

- El test de Kolmogoroff es más sencillo en cuanto a cálculos.
- No pierde información por agrupamiento y puede ser aplicado a muestras pequeñas.
- El test de Chi-cuadrado puede ser aplicado a datos discretos y continuos, y también a hipótesis compuestas, es decir, cuando además de probar un modelo,

deban estimarse sus parámetros a partir de los datos, mientras que el de Kolmogoroff se aplica sólo a datos continuos.
- Los dos test son comparables en términos de potencia, y existe alguna evidencia de que el test de Kolmogoroff es algo más potente que el de Chi-cuadrado.