1 - Conceptos

20 March 2020 17:58

Distribuciones de frecuencia:

- Tabla que registra el *numero de ocurrencias* de un valor observado.
- <u>Intervalos de clase:</u>
  - Posee limites de clase (separar clases *adyacentes*)
  - Punto medio
  - Fronteras de clase (La parte que *no comprende* el limite de frontera pero que *si pertenece* a esa clase)
- <u>Frecuencia acumulada:</u>
  - La *suma* de las frecuencias hasta el momento, la ultima debe ser igual al total de observaciones
- <u>Frecuencia relativa:</u>
  - La frecuencia absoluta se *divide* por las observaciones totales

Fronteras de clase: 159.5 - 179.5

Salario semanal (en pesos)	Número de Trabajadores (frecuencia)
140 – 159	07
160 – 179	20
180 – 199	33
200 – 219	25
220 – 239	11
240 – 259	04
	Total=100 (tamaño de la muestra)

Limite superior

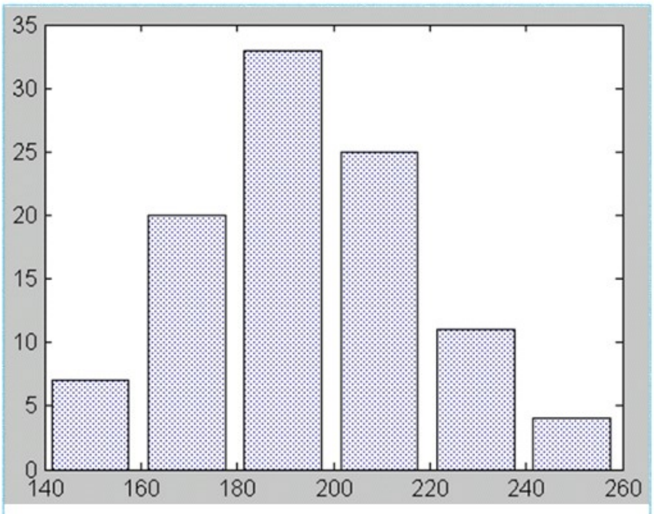
Limite inferior

$$\text{Intervalo aproximado} = \frac{\left(\begin{array}{c}\text{Valor mayor en}\\ \text{datos no agrup.}\end{array}\right) - \left(\begin{array}{c}\text{Valor menor en}\\ \text{datos no agrup.}\end{array}\right)}{\text{número deseado de clases}}$$

Punto medio = (limite inferior + limite superior) / 2

Histogramas y polígonos de frecuencia:

- Diagrama de barras de una *distribución de frecuencias*.
- <u>Ordenadas</u> -> Frecuencia
- <u>Abscisas</u> -> Frontera de clases
- Conviene que se observen mas de 30 valores



Histograma

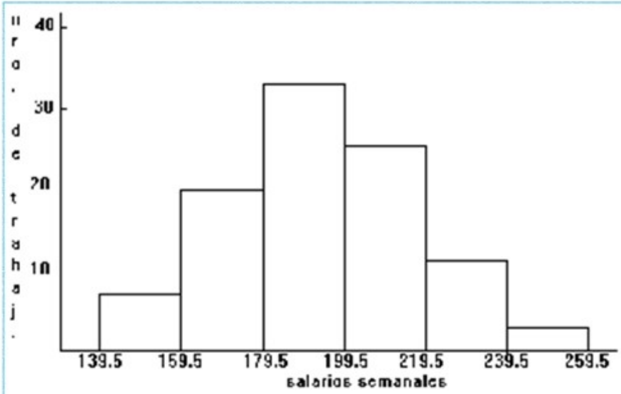


Grafico X-Y

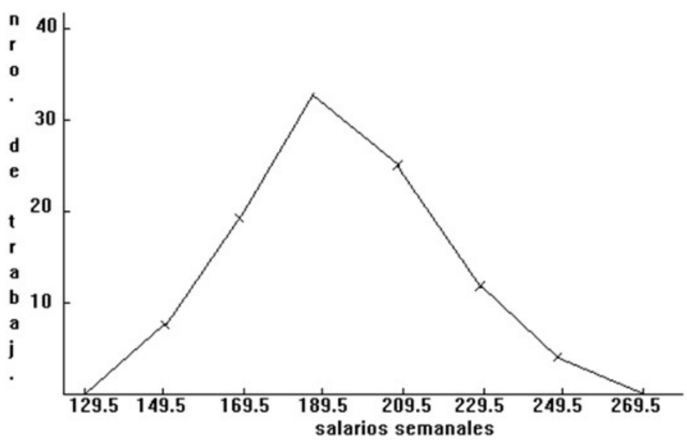
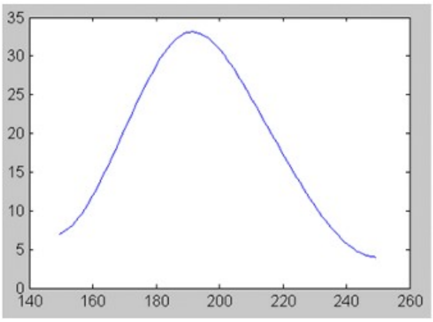
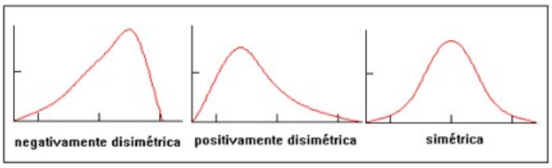


Grafico X-Y

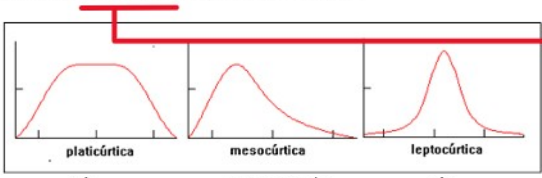
Simetría:



En términos de disimetría se las puede clasificar en:



En términos de curtosis se las puede clasificar en:



Baja concentración

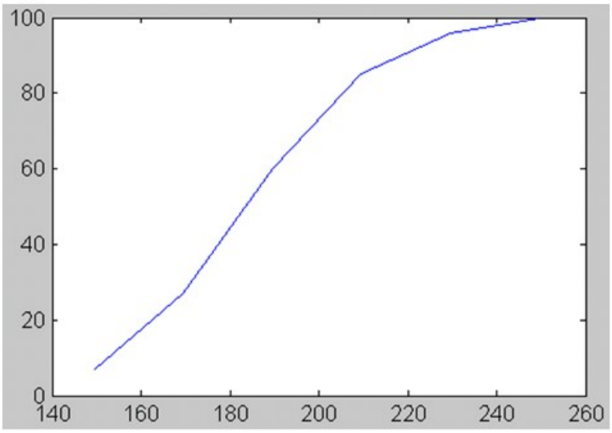
Concentración **normal**

Alta concentración

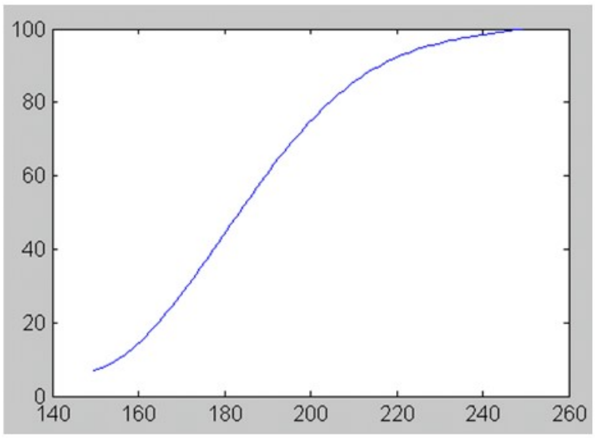
Curtosis: medida de apuntamiento o dispersión

- Determina el *grado de concentración* que presentan los valores de una variable alrededor de la *zona central* de la distribución de frecuencias.
- Determina la *dispersión* (Mayor o menor), en torno a la *media*.

Grafico de distribución de frecuencia acumulada: OJIVA



Ojiva



Ojiva suavizada

El gráfico indica la *frecuencia acumulada* debajo de cada frontera de clase.

Poblaciones y muestras:

Muestra aleatoria:

- Un conjunto de *observaciones* constituyen una muestra aleatoria de tamaño **n** en una población de tamaño **N** (finita), si los elementos de cada *subconjunto* **n** tiene la *misma probabilidad* de ser elegida.
- Tiene densidad de probabilidad $f_{(x)}$ si:
    - Las **n** variables aleatorias son *independientes*

- Cada X_i es un *valor* de una función aleatoria cuya *distribución* es conforme a $f_{(x)}$

Se utilizan estadísticos como $\bar{x}$ o s para *inferir* sobre los *parametros* de la poblacion μ o σ

DISTRIBUCIÓN MUESTRAL DE LA MEDIA (s conocida)

Muestra de <u>n</u>

$$\mu = \sum_{i=0}^{n} x_i \cdot p_i =$$

$$\sigma = \sum_{i=0}^{n} (x_i - \mu)^2 \cdot p_i =$$

Si una muestra aleatoria de tamaño **n** se elige de una población que tiene la media μ y varianza σ^2, entonces es un valor de una variable aleatoria cuya distribución tiene la media μ.

Demostración

Sea una <u>*muestra de tamaño n*</u> de una población con media μ y varianza σ^2 con notación (x1, x2, … , xn). La misma se puede individualizar como <u>*n valores observados de una variable aleatoria X*</u>. También se pueden considerar a estos **n** valores como <u>observaciones simples</u> de **n** <u>*variables aleatorias*</u> X1, X2, … , Xn que tienen la <u>*distribución de X*</u> (media μ y varianza σ^2) y que son <u>*independientes*</u> (ya que los valores de la muestra son independientes). Luego la media muestral es:

$$\bar{X} = \frac{X_1 + X_2 + \ ..+ X_n}{n} = \frac{X_1}{n} + \frac{X_2}{n} + ... + \frac{X_n}{n}$$

$$\mu_{\bar{X}} = E(\bar{X}) = E\left(\frac{X_1}{n}\right) + E\left(\frac{X_2}{n}\right) + ... + E\left(\frac{X_n}{n}\right) = \frac{\mu}{n} + ... + \frac{\mu}{n} = \mu$$

Para muestras tomadas de poblaciones infinitas, la varianza de esta distribución es:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

Demostración

Bajo las condiciones de la prueba anterior, la <u>*varianza de la suma*</u> de variables aleatorias independientes es la <u>*suma de las varianzas*</u> de cada una de las variables:

$$\bar{X} = \frac{X_1}{n} + \frac{X_2}{n} + ... + \frac{X_n}{n}$$

$$\operatorname{var}(\bar{X}) = \operatorname{var}\left(\frac{X_1}{n}\right) + \operatorname{var}\left(\frac{X_2}{n}\right) + ... + \operatorname{var}\left(\frac{X_n}{n}\right)$$

La varianza de una constante por una variable aleatoria es:

$$\operatorname{var}(c \cdot X) = c^2 \cdot \operatorname{var}(X)$$

$$\left(\sigma_{\bar{X}}\right)^2 = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} + ... + \frac{\sigma^2}{n} = n \cdot \frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Para muestras tomadas de poblaciones finitas de tamaño **N**, la varianza de esta distribución es:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}$$

Teorema de Chebyshev

- Muestra pequeña (n<30)
- La población no está normalmente distribuida

No es posible utilizar la distribución de probabilidad normal **ni** la t-Student para construir un intervalo de confianza

El Teorema de Chebyshev establece: La proporción de las medias en un conjunto de datos que se sitúa dentro de las k desviaciones estándar de la media no es menor de $1-1/k^2$, siendo $k \geq 1$.

Al aplicarlo a la distribución de muestreo de una media, la *probabilidad* de que una media muestral se sitúe dentro de **k** unidades de error estándar ($\sigma_{\bar{x}} = \sigma/\sqrt{n}$) a partir de la media de la población es:

$$P\left(\left|\bar{x} - \mu\right| \leq k \cdot \sigma_{\bar{x}}\right) \geq 1 - \frac{1}{k^2}$$

haciendo $k \cdot \sigma_{\bar{x}} = k \cdot \sigma/\sqrt{n} = \varepsilon$, queda:

$$P\left(\left|\bar{x} - \mu\right| \leq k \cdot \varepsilon\right) \geq 1 - \frac{\sigma^2}{n \cdot \varepsilon^2}$$

Para que $|\bar{x} - \mu|$ sea pequeño (menor o igual que ε) basta con hacer n grande.

Si la media proviene de una *población grande* (n > 30, aún si no se conoce la varianza de la misma) es posible definir una *variable aleatoria* llamada media estandarizada cuyos valores están dados por:

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \longrightarrow$$ tiene una distribución *normal estándar*

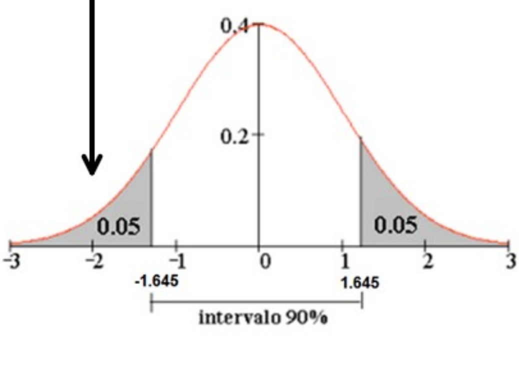
Con una confianza del 90%, a ambos lados de la "campana" deben quedar colas con áreas de 5%.

$$-z_{5\,\%} \leq \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \leq z_{5\,\%}$$

$$-z_{5\,\%} \cdot \frac{s}{\sqrt{n}} - \bar{x} \leq -\mu \leq z_{5\,\%} \cdot \frac{s}{\sqrt{n}} - \bar{x}$$

$$z_{5\,\%} \cdot \frac{s}{\sqrt{n}} + \bar{x} \geq \mu \geq -z_{5\,\%} \cdot \frac{s}{\sqrt{n}} + \bar{x}$$

$$\bar{x} - z_{5\,\%} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{5\,\%} \cdot \frac{s}{\sqrt{n}}$$



En este caso se puede destacar:

- El *incremento* del tamaño de la muestra.
- La *aplicación* del estadístico z, en vez de usar el Teorema de Chebishev.

Teorema central del límite

Sea la variable aleatoria:

$Y = X1 + X2 + \dots + Xn$,

donde X1, X2, … , Xn son variables aleatorias distribuidas idénticamente, cada una con media μ y varianza finita σ^2.
Luego la distribución del estadístico **Z** es:

$$z_n = \frac{Y - n \cdot \mu}{\sqrt{n} \cdot \sigma}$$

el cual se *aproxima* a una distribución normal estándar cuando **n** tiende a infinito.

El teorema central del limite establece que:

La suma de un *número grande de variables aleatorias* tendrá una *distribución normal*, *independientemente* de la distribución individual de la variables sumandos.

Además:

$$\frac{Y - n \cdot \mu}{\sqrt{n} \cdot \sigma} = \frac{\frac{Y}{n} - n \cdot \frac{\mu}{n}}{\sqrt{n} \cdot \frac{\sigma}{n}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

O sea:
La media de **n** variables aleatorias independientes, idénticamente distribuidas (la media de una muestra aleatoria), tendrá aproximadamente una *distribución normal*.

Este teorema se cumple aún cuando las variables sumandos no tengan idéntica distribución, solo si las variables aleatorias individuales hacen una *contribución* "relativamente pequeña" respecto a la suma total

Distribución muestral de la media (σ desconocida)

Como se ha dicho anteriormente, cuando **n** es grande (mayor de 30) se puede reemplazar a σ por s (desviación estándar de la muestra).
Si se supone que la muestra **no** es grande (**n** menor que 30) pero que proviene de una población normal, se puede probar el siguiente teorema:

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño **n** tomada de una *población normal* que tiene media μ y varianza σ^2 , entonces:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Donde **t** es el valor de una *variable aleatoria* con distribución t-Student y parámetro $\nu = n - 1$ (grados de libertad).

La varianza *depende* de los grados de libertad ν. Cuando este valor tiende a infinito ó cuando **n** es grande, la varianza de la distribución *tiende* a 1 y la t_Student se convierte en normal estándar.

TABLA DE RESUMEN PARA ESTIMACION DE INTERVALOS DE CONFIANZA PARA LA MEDIA DE UNA POBLACI ÓN.

Población	Tamaño de la muestra	σ conocida	σ desconocida
Normalmente Distribuida	Grande ($n \geq 30$)	$\bar{x} \pm z_{\alpha} \cdot \sigma_{\bar{x}}$	$\bar{x} \pm z_{\alpha} \cdot s_{\bar{x}}$
	Chica ($n < 30$)	$\bar{x} \pm z_{\alpha} \cdot \sigma_{\bar{x}}$	$\bar{x} \pm t_{\alpha} \cdot s_{\bar{x}}$
No Normalmente Distribuida	Grande ($n \geq 30$)	$\bar{x} \pm z_{\alpha} \cdot \sigma_{\bar{x}}$	$\bar{x} \pm z_{\alpha} \cdot s_{\bar{x}}$
	Chica ($n < 30$)	$\bar{x} \pm k \cdot \sigma_{\bar{x}}$ donde $1\text{-}1/k^2$ se define por medio del Teorema de Chebyshev	$\bar{x} \pm k \cdot s_{\bar{x}}$ donde $1\text{-}1/k^2$ se define por medio del Teorema de Chebyshev

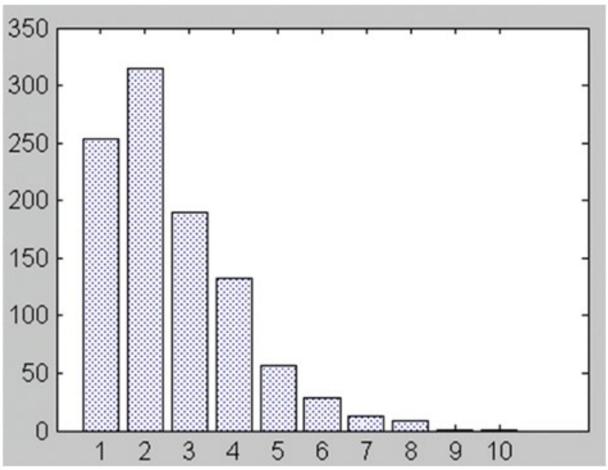
Cuando **n** tiende a ∞ se puede tomar como población <u>normalmente distribuida</u>.

Distribución muestral de la varianza

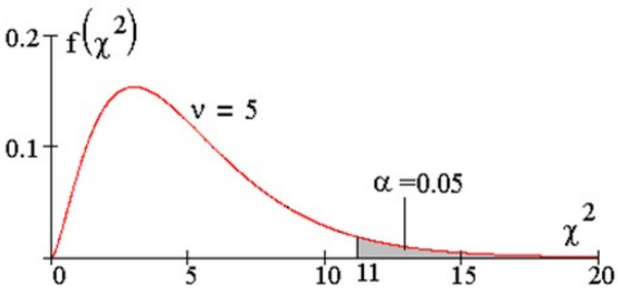
<u>La distribución muestral de la varianza **no** puede</u>:

- Ser *negativa*
- Responder a una distribución *normal*

Sino que responde a la distribución *Chi-cuadrado* (χ^2). La cual esta tabulada para $\chi^2{}_{\alpha}$. Con los para metros "α" y "ν", los cuales representan el *area* debajo de la curva de la distribucion hacia la derecha y los *grados de libertad*, respectivamente.



Simulación en MAthCAD



Usando una tabla

Si s^2 es la varianza de una muestra aleatoria de tamaño **n** tomada de una *población normal* cuya varianza es σ^2 , entonces:

$$\chi^2 = (n-1) \cdot \frac{s^2}{\sigma^2}$$

es un valor de una *variable aleatoria* que tiene la distribución chi-cuadrado con parámetro $\nu = n - 1$.

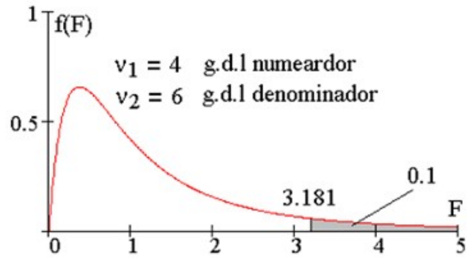
$$\chi^2 = (n-1) * \frac{s^2}{\sigma^2} = \nu * \frac{s^2}{\sigma^2}$$

Si se toma la razón de dos muestras tomadas aleatoriamente, sirve como prueba para determinar si dos muestras *provienen* de poblaciones con varianzas iguales, en dicho caso, esta debe ser *cercana* a 1.

Si s_1^2 y s_2^2 son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 , respectivamente, tomadas de dos *poblaciones normales* que tienen la *misma varianza*, entonces:

$$F = \frac{s_1^{\ 2}}{s_2^{\ 2}}$$

es una *variable aleatoria* que tiene la distribución F con parámetros $\nu_1 = n_1 - 1$ (grados de libertad del numerador) y $\nu_2 = n_2 - 1$ (grados de libertad del denominador).
La distribución esta tabulada para los valores F_α con parámetros ν_1 y ν_2 . Donde α es el área bajo la distribución hacia la derecha.



Una propiedad de esta distribución es que:

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)}$$

esto se puede apreciar, a partir del trabajo con tablas, para α=0.95, ν_1=7 y ν_2= 13, resulta:

$F_{0.95}(7, 13) = 2.832$ $\qquad$ $F_{0.05}(13, 7) = 0.353$

dichos números son evidentemente uno el recíproco del otro.

Inferencias relativas a medias

Estimación puntual:	*Estimador insesgado:*
Se refiere a la elección de un estadístico, (un número calculado a partir de los datos muestrales) respecto al cual tenemos alguna esperanza o seguridad de que esté "razonablemente cerca" del parámetro que se ha de estimar.	Un estadístico es un estimador insesgado, si y sólo si la media de la distribución de estimados es igual a θ.

Si se comparan las distribuciones muestrales de la *media* y la *mediana* de **muestras aleatorias** de tamaño **n** de la **misma población normal**.
Las dos distribuciones:

- Tienen la misma media μ
- Ambas son simétricas
- Ambas tienen forma acampanada
- Sus varianzas difieren.
  - Varianza para la **media** es: σ^2/n
  - Varianza para la **mediana** es: $1.5708*\sigma^2/n$ → para poblaciones infinitas

Un estadístico $\hat{\theta}_1$ es un *estimador insesgado* más eficiente del parámetro θ que el estadístico si:

- $\hat{\theta}_1$ y $\hat{\theta}_2$ son ambos *estimadores insesgados* de θ.
- La varianza de la distribución muestral del primer estimador es menor que la del segundo.

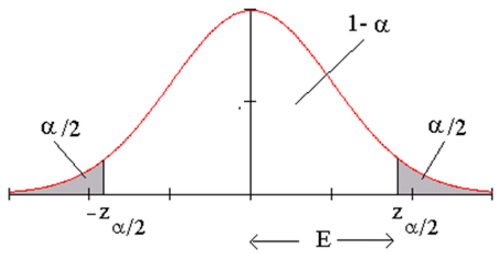
Es más probable que la media esté más cerca de μ que la mediana, por lo que, en la práctica, la media muestral es un estadístico aceptable para estimar la media de la población μ.

Hay una baja probabilidad de acertar exactamente a μ, por lo que conviene acompañar la *estimación puntual* con una afirmación de cuan cerca se puede encontrar la estimación.

Para un **n** grande se puede asegurar que con una probabilidad de **1 - α**, se cumple con la desigualdad:

$$-z_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}} \quad \text{o bien} \quad \frac{|\bar{x}-\mu|}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}$$

Tal que:



Donde el error es $|\bar{x}-\mu|$
Donde E es el *Error Máximo de Estimación*

Siempre la mitad del intervalo de confianza

$$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \geq |\bar{x}-\mu| \qquad \textit{Con probabilidad } 1-\alpha$$

Teniendo esto en cuanta eso, $\bar{x} \pm E$ son los limites de confianza.

Por lo que se puede asegurar con una confianza del $1-\alpha$ (probabilidad) de que el *error de estimacion* para una media $\bar{x}$ (media muestral) es a lo sumo E (Error maximo de estimación).

Se hacen afirmaciones de:

- Probabilidad, acerca de valores futuros de variables aleatorias (digamos error potencial de una estimación).
- Confianza, una vez que los datos han sido obtenidos.

Un Intervalo de Confianza para la media es un *intervalo estimado* construido con respecto a la *media de la muestra*, por el cual puede especificarse la *probabilidad que el intervalo incluya el valor de la media poblacional*.
El Grado de Confianza, asociado con un intervalo de confianza, indica el *porcentaje de los intervalos que incluirán el parámetro* que se está estimando.

Si se desea que el error sea una *cantidad predeterminada*, se debe tomar un tamaño muestral tal que:

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \longrightarrow n = \left(z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{E} \right)^2$$

n redondeado al entero superior

Esto es aplicable a muestras donde se conoce su σ (desviación estándar poblacional) o s (desviación estándar muestral) en el caso de que la muestra sea lo suficientemente grande.
En el caso de que se este muestreando una población normal pero con un tamaño de muestra pequeño y desviación estándar poblacional desconocida, se debe tomar:

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

Donde **t** es una variable aleatoria con distribución t-Student con ν = n - 1 grados de libertad.
Entonces su error máximo de estimación es:

$$E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Estimación por intervalos

Dado que la *probabilidad* de hacer una *estimación puntual* es de 0, se hacen estimaciones por *intervalos*.

Si se conocen σ^2 (o s en una muestra grande) y μ , se puede crear un intervalo de confianza contiene a μ con un nivel de confianza $1 - \alpha$

$$-z_{\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}} \longrightarrow -z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \longrightarrow \boxed{\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}}$$

Esto expresado como intervalo quedaría:

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Estimación Bayesiana

Regla de Bayes

$$P\left(A_k / B\right) = \frac{P\left(B / A_k\right) \cdot P\left(A_k\right)}{\sum_j P\left(B / A_j\right) \cdot P\left(A_j\right)}$$

"causas"

Probabilidad de que el evento B, sea *resultado* de la causa A_k.
Método para calcular la probabilidad de una causa dado un efecto.

Este método es considerado un método "a posteriori", donde se conoce el valor de los parámetros.

Hay métodos de inferencia que consideran a los parámetros como *variables aleatorias* en los que se valoran conceptos de *probabilidad subjetiva*.

Se presenta un método Bayesiano para estimar la media de una población considerando a μ como una *variable aleatoria* con distribución subjetiva.
En la distribución a priori se tienen una media μ_0 y una desviacion estadar σ_0.
Si $f_{(x)}$ es la funcion de la distribucion, se toma la probabilidad entre los numeros a y b deseados, tal que $P = \int_a^b f_{(x)} dx$,dando un analisis a priori.
Luego se toma una muestra y se registran x' y σ. Se pueden calcular la probabilidad con estos numeros, pero se tiene una *mejor estimacion* con:

$\mu 1 := \frac{n \cdot x' \cdot \sigma 0^2 + \mu 0 \cdot \sigma^2}{\ldots}$ $\sigma 1 := \sqrt{\frac{\sigma 0^2 \cdot \sigma^2}{\ldots}}$

$$\mu 1 := \frac{n \cdot x' \cdot \sigma 0^2 + \mu 0 \cdot \sigma^2}{n \cdot \sigma 0^2 + \sigma^2} \qquad \sigma 1 := \sqrt{\frac{\sigma 0^2 \cdot \sigma^2}{n \cdot \sigma 0^2 + \sigma^2}}$$

Calculando la probabilidad con estos parámetros que *relacionan* el análisis a priori y a posteriori se obtiene una mejor estimación