



Análisis y Visualización de Datos

Diplomatura CDAAyA 2020





Teoría para aplicar



Población y muestra

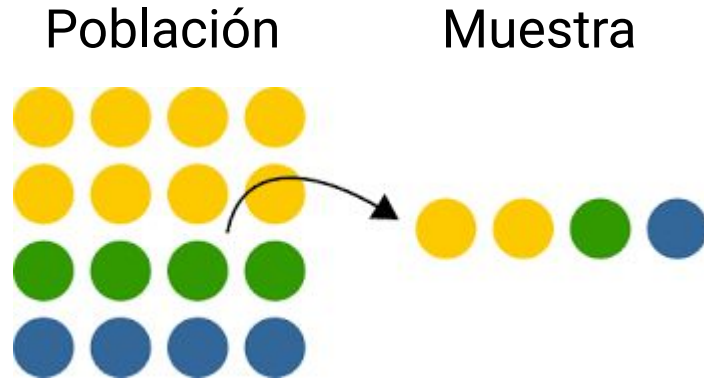
Cuando recogemos los datos muchas veces es imposible relevar la característica de interés de todo el grupo o universo, se examina una pequeña parte, llamada muestra.

Al medir una característica en una muestra, se consideran los datos x_1, x_2, \dots, x_n , como realizaciones de X_1, X_2, \dots, X_n m.a.

Notar la diferencia entre minúscula y mayúscula



Muestreo aleatorio



La muestra debe ser representativa de la población

Muestra aleatoria

Una sucesión de v.a. X_1, X_2, \dots, X_n se dice **muestra aleatoria (m.a.)** si son v. a. independientes e idénticamente distribuidas (i.i.d.). “Clones” de una misma X

Todas las medidas antes mencionadas para una muestra de datos podemos pensarlas a partir de una muestra aleatoria, también serán variable aleatorias, llamadas estadísticos. Como por ejemplo el **estadístico Media Muestral**:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Algunas propiedades teóricas: Muestra aleatoria

- Si X_1, X_2, \dots, X_n m.a. (v.a.i.i.d.) tal que $X_i \sim N(\mu, \sigma^2)$, entonces:
 - $X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2)$
 - $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n)$
- Si Z_1, Z_2, \dots, Z_n m.a. tal que $Z_i \sim N(0, 1)$, entonces:

$$V = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

LFGN: Ley Fuerte de los Grandes Números

- Dada. X_1, \dots, X_n m.a. c/u con media μ ("clones" de la misma variable con distribución cualquiera pero con esperanza $E(X_1) = \mu$, media poblacional)

$$\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mu$$

$$P(|\overline{X} - \mu| \leq \epsilon) \xrightarrow{n \rightarrow \infty} 1$$

TCL: Teorema Central del Límite

- Sea X_1, \dots, X_n m.a. c/u con media μ y varianza σ^2 . (“clones” de la misma variable con distribución cualquiera pero con media μ y varianza σ^2)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim N(0, 1)$$

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq t\right) \approx P(Z \leq t) = F(t)$$

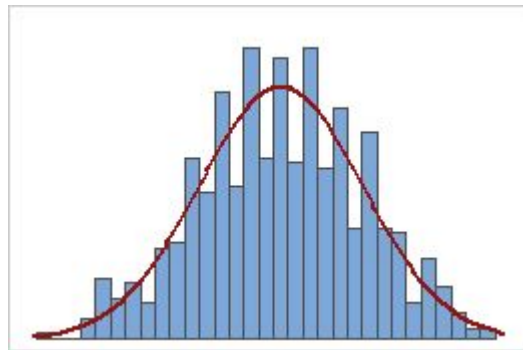
TCL: Teorema Central del Límite

- La densidad de **la v.a. media muestral** parece acampanada p/ **n grande, aprox. normal**, cualquiera sea la distribución en la población;
- La densidad de **la media muestral** crece en altura y decrece en dispersión en la medida n crece.
- La media de la distribución del promedio muestral es igual a la media de la población
- La varianza de la distribución de la media muestral es menor que la varianza de la población;

TCL: Teorema Central del Límite

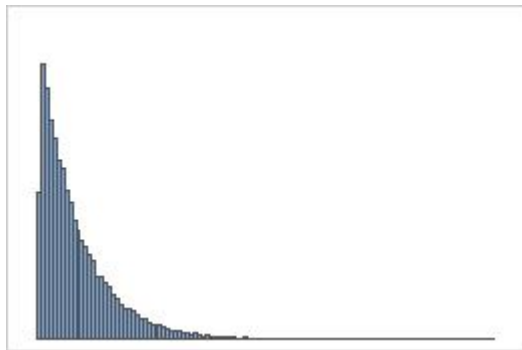


$X_i \sim \text{Uniforme}$

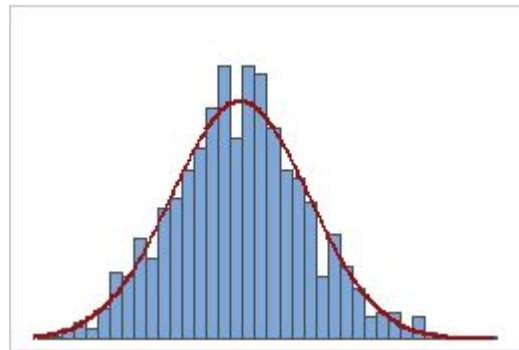


$$\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

TCL: Teorema Central del Límite



$X_i \sim \text{exponencial}$



$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Estadística Inferencial

Inferencia Estadística

Métodos utilizados para **tomar decisiones** o para **obtener conclusiones** sobre una población (usa modelos y parámetros generalmente).

Estos métodos utilizan la información contenida en una muestra de la población.

Permiten **inferir** el comportamiento de la población con un riesgo medible en términos de **probabilidad de error**.

Inferencia Estadística

Nos ubicaremos principalmente dentro de la estadística paramétrica:

Se considera una característica de interés de la **población (Ω)**. Se supone que la característica está modelada por una **variable aleatoria X** con distribución conocida y paramétrica $f_{\theta}(x)=f(x,\theta)$. (parámetro/s θ)

Se considera una **muestra aleatoria (m.a.) X_1, \dots, X_n** , con la misma distribución (paramétrica) que X .

Inferencia Estadística

Incluye dos grandes áreas:

- estimación de parámetros (estadística paramétrica)
- pruebas de hipótesis



Estimación

Estimación de parámetros

- **Estimación puntual** (por estadístico, estimador del parámetro).

$$\hat{\mu} \approx \mu \quad , \text{ donde } \hat{\mu} = \hat{\mu}(X_1, \dots, X_n), \text{ estadístico}$$

- **Estimación por intervalo**, ($I=I(X_1, \dots, X_n)$ y $S=S(X_1, \dots, X_n)$, estadísticos)

$$\mu \in [I, S] \qquad P(I \leq \mu, \mu \leq S) \approx 1$$

Estimación puntual

Estimación puntual: Estadístico

Un estadístico es una cuenta que depende de la muestra. Resulta ser una variable aleatoria también pues depende de otras. Es una **función** de la muestra aleatoria

$$Y_n = g((X_1, \dots, X_n))$$

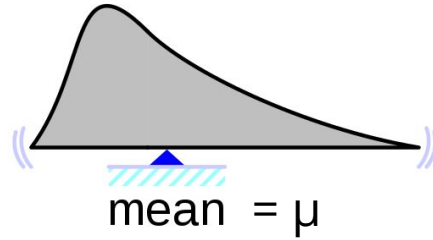
Un **estadístico** sirve para estimar un **parámetro**. Diferenciamos:

- PARÁMETRO \Rightarrow una medida resumen de la población. Valor fijo (desconocido)
- ESTADÍSTICO \Rightarrow una medida resumen de la muestra. Variable aleatoria

ejemplo: parámetro: μ , estadístico: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Estimación de parámetro por estadístico

Ejemplo: Sea X_1, \dots, X_n m.a. de



El estadístico $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ es un buen **estimador** de μ ,

Pues $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mu$ por la LFGN

$$P(|\bar{X} - \mu| \leq \epsilon) \xrightarrow{n \rightarrow \infty} 1$$

Estimadores de parámetros: Más Ejemplos

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n}$$

$$S^2 = \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{n-1}$$

estimadores de $\sigma^2 = E(X - \mu)^2$

$$s = \sqrt{S^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

estimador de σ

$$CA_F = \frac{\sum_{i=1}^N (x_i - \bar{x})^3}{N \cdot S_x^3}$$

estimador de $E(X - \mu)^3 / \sigma^3$

siendo \bar{x} la media y S_x la desviación típica

Estimadores, propiedades deseadas

Se quiere estimar el parámetro θ

Se define un estimador $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ Estadístico (v.a.)

Se define el $sesgo = E(\hat{\theta}) - \theta$ sesgo del estimador

Se dice que un Estimador es **insesgado** si $sesgo=0$

Estimadores, propiedades deseadas

$$sesgo = E(\hat{\theta}) - \theta$$

$$sesgo = E\left(\sum_{i=0}^n \frac{x_i}{n}\right) - \mu$$

$$sesgo = E\left(\frac{1}{n} \sum_{i=0}^n x_i\right) - \mu$$

$$sesgo = \frac{1}{n} E\left(\sum_{i=0}^n x_i\right) - \mu$$

$$E(X_i) = E(X) = \mu$$

$$sesgo = \frac{1}{n} \cdot n \cdot \mu - \mu$$
$$sesgo = 0$$

Luego:

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ es un **estimador insesgado** de μ

Estimadores, comparación

Supongamos que tomamos otro estimador para μ , lo anotamos $\hat{\mu}_2 = \frac{X_1 + X_n}{2}$

Entonces como

$$E(\hat{\mu}_2) = E\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{2}(E(X_1) + E(X_n)) = \frac{1}{2}(\mu + \mu) = \frac{1}{2}2\mu = \mu,$$

$\hat{\mu}_2 = \frac{X_1 + X_n}{2}$ es también un estimador insesgado de μ

¿Cuál de los dos estimadores es mejor?

Calculamos la varianza de cada uno utilizando las propiedades de la varianza.

Estimadores: precisión

$$V(X_i) = V(X) = \sigma^2 \quad \forall i = 1, 2, \dots, n, \text{ tenemos}$$

$$V(\bar{X}) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i),$$

Análogamente calculamos la varianza de $\hat{\mu}_2 = \frac{X_1 + X_n}{2}$:

$$V(\hat{\mu}_2) = V\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{4} (V(X_1) + V(X_n)) = \frac{1}{4} (\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}$$

Vemos que si $n > 2$ entonces $V(\hat{\mu}_1) < V(\hat{\mu}_2)$. Por lo tanto si $n > 2$ es mejor estimador $\hat{\mu}_1$

Estimadores: Eficiencia/precisión

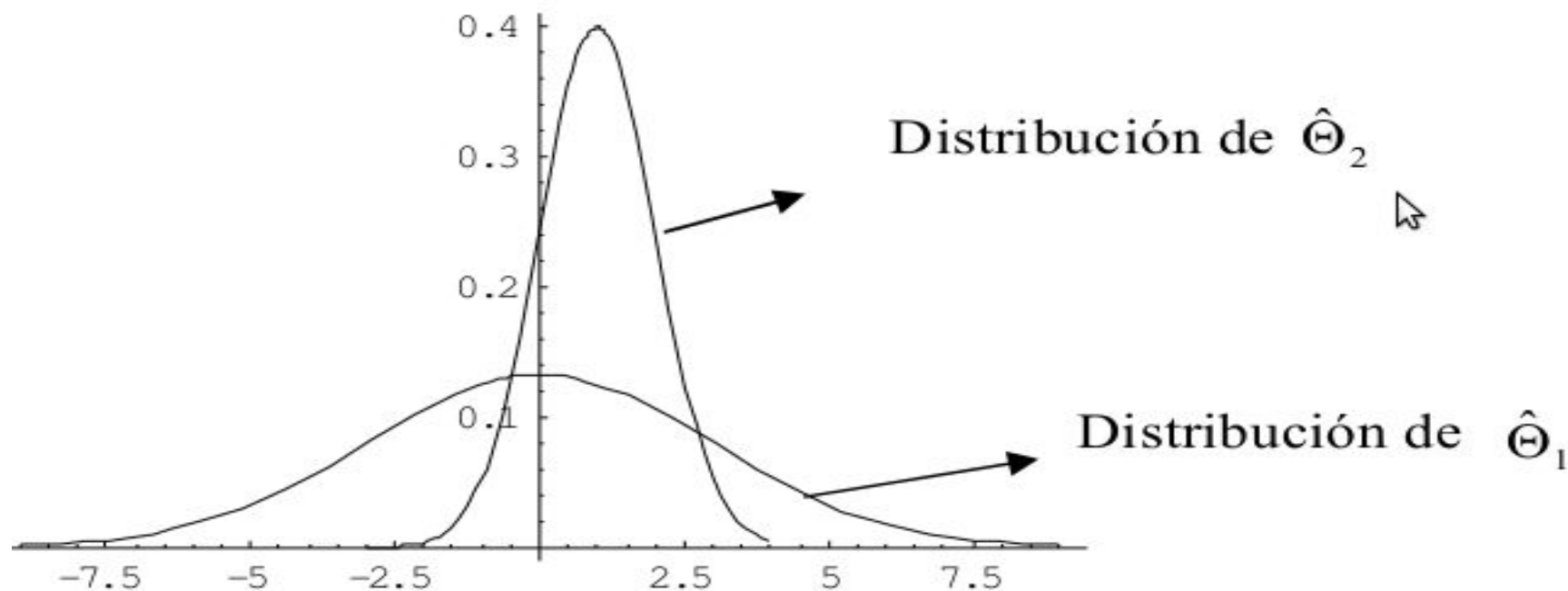
Diremos que un estimador es **más eficiente o más preciso** que otro estimador, si la varianza del primero es menor que la del segundo.

Por ejemplo, $\hat{\theta}_1$ y $\hat{\theta}_2$, ambos estimadores de θ y

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$$

diremos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$. Un estimador es más eficiente (más preciso), por tanto, cuanto menor es su varianza.

Estimadores: Sesgo y Eficiencia (presición)



Estimadores

Método de los momentos, Estimador de máxima verosimilitud (EMV), etc.

Ej.: Si X_1, \dots, X_n es una m.a. de una distrib. normal,

Entonces los estimadores máxima verosimilitud de μ y σ^2 son

$$\begin{cases} \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$

Estimación por intervalos

Intervalos de Confianza

- A veces resulta más conveniente dar un intervalo de valores posibles del parámetro desconocido, de manera tal que dicho intervalo contenga al verdadero parámetro con alta probabilidad.

Intervalos de confianza

Estimación por intervalo, ($I=I(X_1, \dots, X_n)$ y $S=S(X_1, \dots, X_n)$, estadísticos)

$$\mu \in [I, S]$$

$$P(I \leq \mu, \mu \leq S) \approx 1$$


- Un intervalo de confianza es un intervalo aleatorio (con extremos aleatorios dados por estadísticos).

Intervalos de confianza


Estimación por intervalo, se quiere estimar θ

$$P(\theta \in (\hat{\theta}_1, \hat{\theta}_2)) = 1 - \alpha$$

Parámetro
desconocido
a estimar



es un valor real
entre cero y uno
dado de antemano



Intervalos de confianza: Ejemplo

Por ejemplo si pedimos un $\alpha=0.05$ esto implica que

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 0.95$$

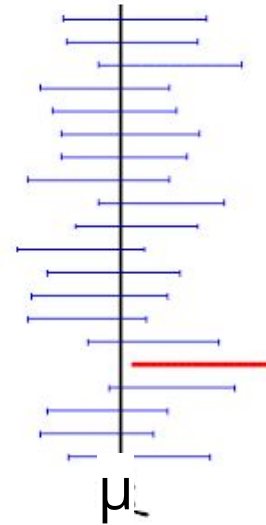
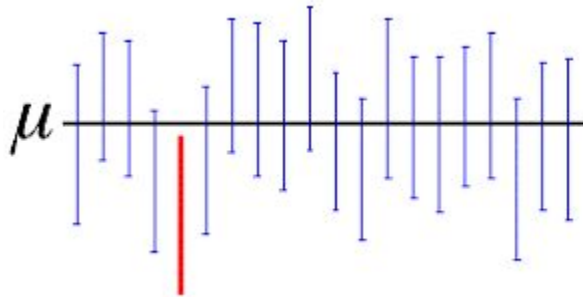
Una probabilidad del 95% que el verdadero parámetro se encuentre en el intervalo propuesto

Al valor $1 - \alpha$ o a $(1 - \alpha)100\%$ se lo llama nivel de confianza del intervalo.

Debido a su naturaleza aleatoria, es poco probable que dos muestras de una población en particular produzcan intervalos de confianza idénticos

Intervalos de Confianza

La línea negra representa el valor fijo desconocido μ . Los intervalos de confianza azules que cortan la línea negra contienen al valor μ . El intervalo de confianza rojo que está completamente por debajo de la línea horizontal no lo contiene. Un intervalo de confianza de 95% indica que 19 de 20 muestras (95%) de la misma población producirán intervalos de confianza que contendrán al parámetro.



IC: Método del pivote

Se define un **estadístico (pivote)** que **depende de la m.a. y del parámetro a estimar y cuya distribución es conocida** (o aproximada a una conocida) y no depende del parámetro.

Al conocerle la distribución se pueden establecer los límites dados por dos desigualdades donde el **estadístico pivote** tiene probabilidad **$1 - \alpha$** de valer.

Luego se despeja el **parámetro**, condicionado por dos desigualdades con probabilidad (o **nivel de confianza**) **$1 - \alpha$** .

Veamos un ejemplo sencillo para llevarlo a la práctica...

Método del Pivote: Ejemplo

Sea X_1, X_2, \dots, X_n una m.a. de una v.a. $X \sim N(\mu, \sigma^2)$, σ^2 conocido.

- se quiere construir un IC para de nivel $(1 - \alpha)$,

Notemos que $\bar{X} - \mu \sim N(0, \sigma^2/n)$, su distribución ya no depende de μ

Y luego $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, distribución conocida y no depende de ningún parámetro



Pivote

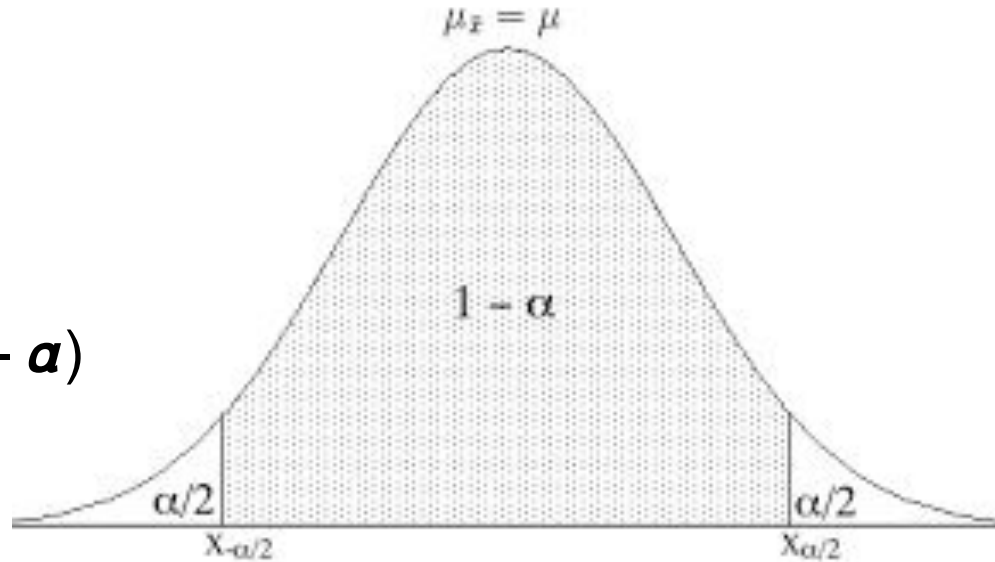
Método del Pivote: Ejemplo

Sea X_1, X_2, \dots, X_n una m.a. de una v.a. $X \sim N(\mu, \sigma^2)$, σ^2 conocido.

- se quiere construir un IC para de nivel $(1 - \alpha)$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

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



Método del Pivote: Ejemplo

Sea X_1, X_2, \dots, X_n una m.a. de una v.a. $X \sim N(\mu, \sigma^2)$, σ^2 conocido.

- se quiere construir un IC para de nivel $(1 - \alpha)$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad P \left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right) = (1 - \alpha)$$

Pivote

$$P \left(\underbrace{\bar{X} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}}_I \leq \mu \leq \underbrace{\bar{X} + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}}_S \right) = (1 - \alpha)$$

$[I, S]$ es IC de nivel $1 - \alpha$ para μ

TCL- Intervalo de confianza (asintótico, n grande)

- Sea X_1, \dots, X_n m.a. c/u con media μ y varianza σ^2 . $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

$$P\left(\underbrace{\bar{X} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}}_I \leq \mu \leq \underbrace{\bar{X} + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}}_S\right) \approx 1 - \alpha$$

$[I, S]$ es IC de nivel asintótico $1-\alpha$ para μ

Longitud del IC

$$P\left(\mu \in \left[\bar{X} \pm \frac{\sigma z_{\alpha/2}}{\sqrt{n}}\right]\right) \approx 1 - \alpha$$

$$L = 2 \frac{\sigma z_{\alpha/2}}{\sqrt{n}}$$

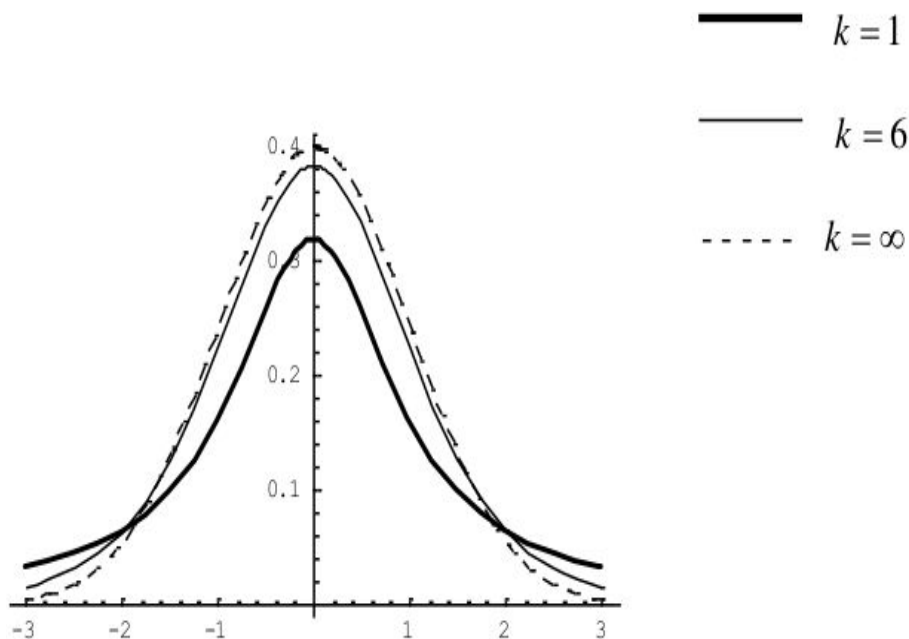
En general, si queremos hallar n tal que $L = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq l$, donde l es un valor dado, entonces despejando n

$$n \geq \left(\frac{2z_{\frac{\alpha}{2}}\sigma}{l} \right)^2$$

IC para n chico, t de Student

$$T = \frac{X - \mu}{S / \sqrt{n}}.$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$



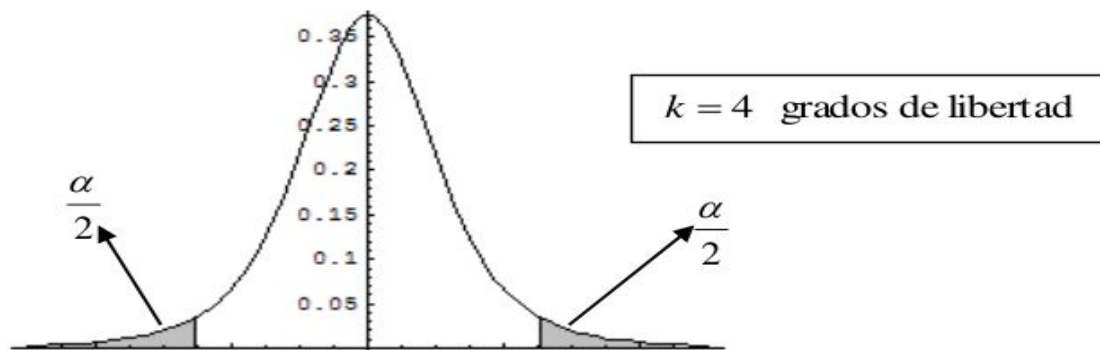
IC para n chico, t de Student con $(n-1)$ g.l.

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

Evidentemente, si definimos

$$\begin{cases} \hat{\Theta}_1 = \bar{X} - t \frac{S}{\sqrt{n}} \\ \hat{\Theta}_2 = \bar{X} + t \frac{S}{\sqrt{n}} \end{cases}, \text{ hemos construido dos estadísticos } \hat{\Theta}_1 \text{ y } \hat{\Theta}_2 \text{ tales que } P(\hat{\Theta}_1 \leq \mu \leq \hat{\Theta}_2) = 1 - \alpha,$$

veamos quien es el número t que verifica la ecuación, es decir (ver figura):



IC para μ de una normal con σ desconocido

Si (X_1, X_2, \dots, X_n) una muestra aleatoria de tamaño n de una v.a. X donde $X \sim N(\mu, \sigma^2)$, σ^2 desconocido, un intervalo de confianza para μ de nivel $1 - \alpha$ es

$$\left[\bar{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right] \quad (8.2)$$

IC para μ de una normal con σ desconocido

Ejemplo:

Se hicieron 10 mediciones sobre la resistencia de cierto tipo de alambre que dieron valores x_1, x_2, \dots, x_{10} tales que $\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 10.48$ ohms y $S = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 1.36$ ohms. Supóngase que $X \sim N(\mu, \sigma^2)$.

Se desea obtener un intervalo de confianza para la esperanza poblacional μ al 90 %.

Tenemos que $1 - \alpha = 0.90 \rightarrow \alpha = 0.1 \rightarrow \alpha / 2 = 0.05$

De la Tabla de la t de Student tenemos que $t_{0.05,9} = 1.8331$. Entonces el intervalo de confianza buscado es:

$$\left[\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \right] = \left[10.48 - 1.8331 \frac{1.36}{\sqrt{10}}, 10.48 + 1.8331 \frac{1.36}{\sqrt{10}} \right]$$

Esto es: $[9.69, 11.27]$.

IC para dif de medias $\mu_1 - \mu_2$

X_1, X_2, \dots, X_{n_1} una m.a. de una v.a. $X \sim N(\mu_1, \sigma^2)$

Y_1, Y_2, \dots, Y_{n_2} una m.a. de una v.a. $Y \sim N(\mu_2, \sigma^2)$, Varianzas iguales, pivote:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S \sqrt{(n_1 + n_2) / n_1 n_2}} \sim t_{n_1 + n_2 - 2}$$

$$\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \quad S^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

Varianzas distintas: Método de Welch (usa t de student con k grados de libertad)

IC para dif de medias $\mu_1 - \mu_2$

Muestras apareadas:

X_1, X_2, \dots, X_n una m.a. de una v.a. $X \sim N(\mu_1, \sigma_1^2)$

Y_1, Y_2, \dots, Y_n una m.a. de una v.a. $Y \sim N(\mu_2, \sigma_2^2)$

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ una m.a. de normal bivariada $N((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho))$

$$Z_i = X_i - Y_i$$

Z_1, Z_2, \dots, Z_n una m.a. de una v.a. $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\rho/(\sigma_1 \sigma_2)) = N(\mu, \sigma^2)$

IC para σ^2 de una normal

Supongamos que se quiere hallar un intervalo de confianza para σ^2 de una distribución normal.

Sea X_1, X_2, \dots, X_n una muestra aleatoria de una v.a. X , donde $X \sim N(\mu, \sigma^2)$.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Estimador

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

Pivote con distribución $\chi^2_{(n-1)}$ **conocida**

IC para σ^2

Un fabricante de detergente líquido está interesado en la **uniformidad** de la máquina utilizada para llenar las botellas. De manera específica, es deseable que la desviación estándar σ del proceso de llenado sea menor que 0.15 onzas de líquido

Supongamos que la distribución del volumen de llenado es aproximadamente normal. Al tomar una muestra aleatoria de 20 botellas, se obtiene una varianza muestral $S^2 = 0.0153$. Hallar un intervalo de confianza de nivel 0.95 para la verdadera varianza σ^2 y con este un IC para **el verdadero desvío σ** del volumen de llenado.

IC para σ

Solución:

La v.a. de interés es X : “volumen de llenado de una botella”

Se asume que $X \sim N(\mu, \sigma^2)$ con σ desconocido.

Estamos en las condiciones para aplicar (8.8)

Tenemos que $1 - \alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \chi^2_{1-\frac{\alpha}{2}, n-1} = \chi^2_{0.975, 19} = 8.91$ y $\chi^2_{\frac{\alpha}{2}, n-1} = \chi^2_{0.025, 19} = 32.85$

Además $S^2 = 0.0153 \Rightarrow S = 0.1237$

Por lo tanto el intervalo es

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}}; \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}} \right) = \left(\frac{(20-1) \times 0.0153}{32.85}; \frac{(20-1) \times 0.0153}{8.91} \right) = (0.00884; 0.0326)$$

Y un intervalo para σ es $(\sqrt{0.00884}; \sqrt{0.0326}) = (0.09; 0.1805)$

Por lo tanto con un nivel de 0.95 los datos **no apoyan la afirmación que** $\sigma < 0.15$