



# Ciencia de Datos y BigData

## Análisis y Visualización de Datos - Estimación

**Dr. José Ramón Iglesias**

DSP-ASIC BUILDER GROUP  
Director Semillero TRIAC  
Ingeniería Electronica  
Universidad Popular del Cesar

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

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

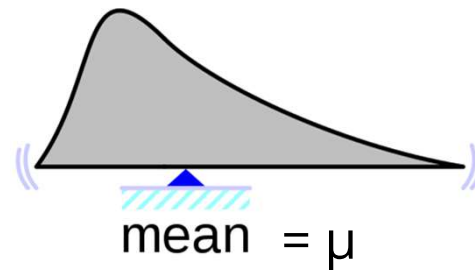
Un **estadístico** sirve para estimar un **parámetro**. Diferenciamos:

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

ejemplo: parámetro:  $\mu$ , 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  $\mu$ ,

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  $\sigma$

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

**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  $\theta$

Estimador Estadístico (v.a.)  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$

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

sesgo del estimador

Estimador es **insesgado** si  $sesgo=0$

---



# Estimadores: Eficiencia/precisión

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

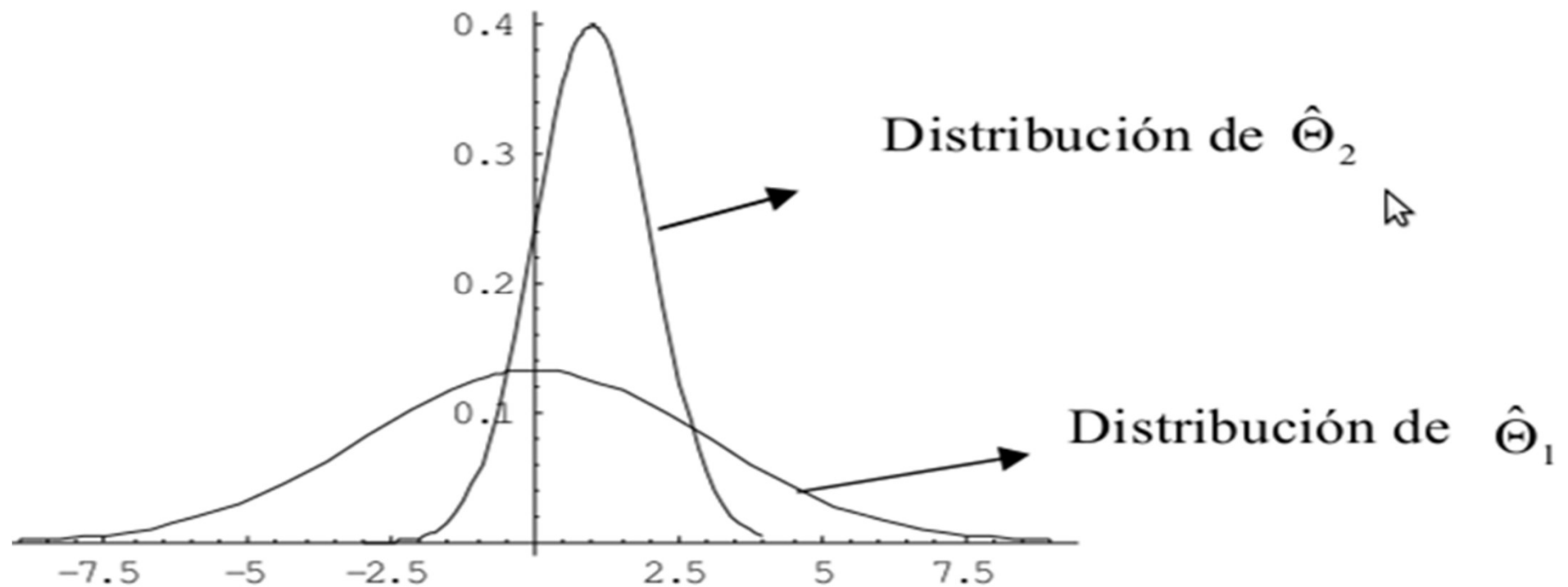
Por ejemplo,  $\hat{\theta}_1$  y  $\hat{\theta}_2$ , ambos estimadores de  $\theta$  y

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

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

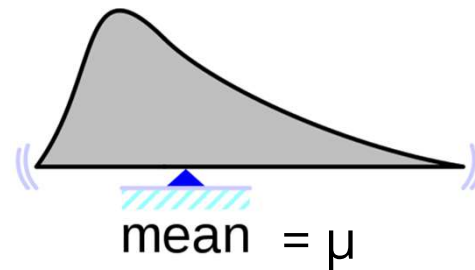
---

# Estimadores: Sesgo y Eficiencia (precisión)



# Estimador Insesgado y Eficiente (preciso)

$X_1, \dots, X_n$  m.a. de



El estadístico eficiente:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

es **estimador** insesgado de  $\mu$  y

$$\bar{X} \approx N(\mu, \sigma^2/n)$$

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

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

Demo con Notebook  
EL460\_leer\_csv\_estatura.ipynb

---

# 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  $\theta$

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

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

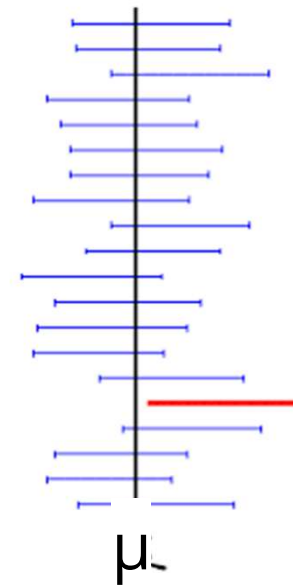
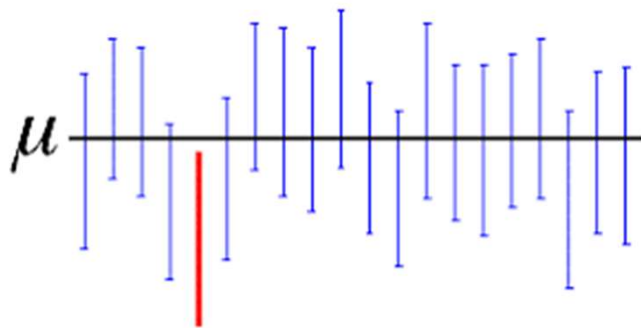
**Al valor  $1 - \alpha$  o a  $(1 - \alpha)100\%$  se lo llama nivel de confianza del intervalo.**

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  $\mu$ . Los intervalos de confianza azules que cortan la línea negra contienen al valor  $\mu$ . 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.  $X \sim N(\mu, \sigma^2)$ ,  $\sigma^2$  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  $\mu$

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



**Pivote**

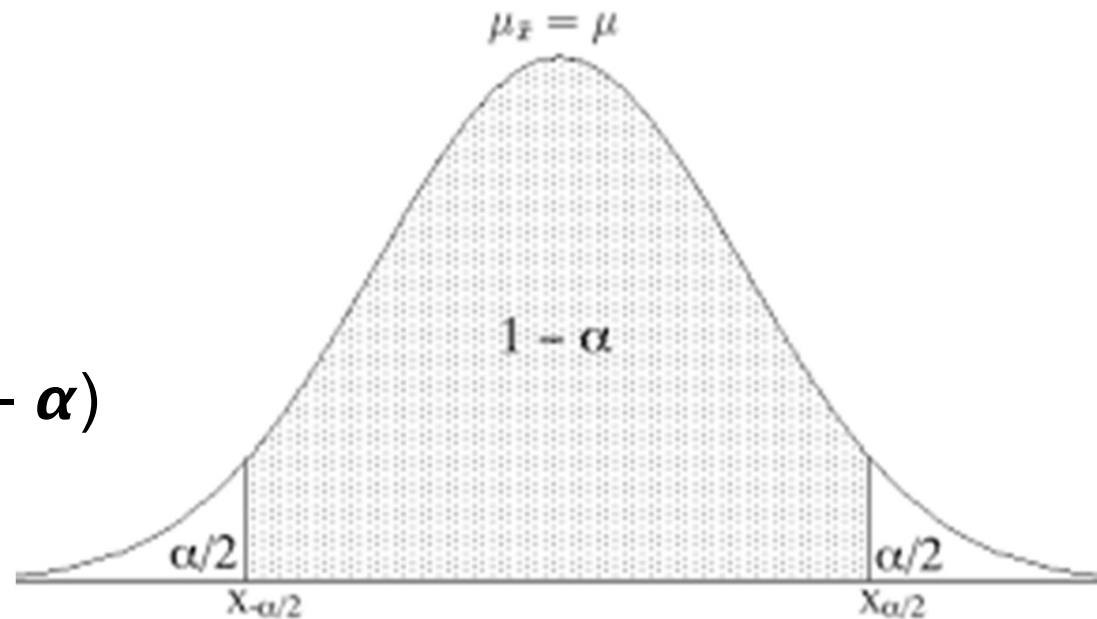
# Método del Pivote: Ejemplo

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

## TCL- Intervalo de confianza (asintótico, n grande)

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

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

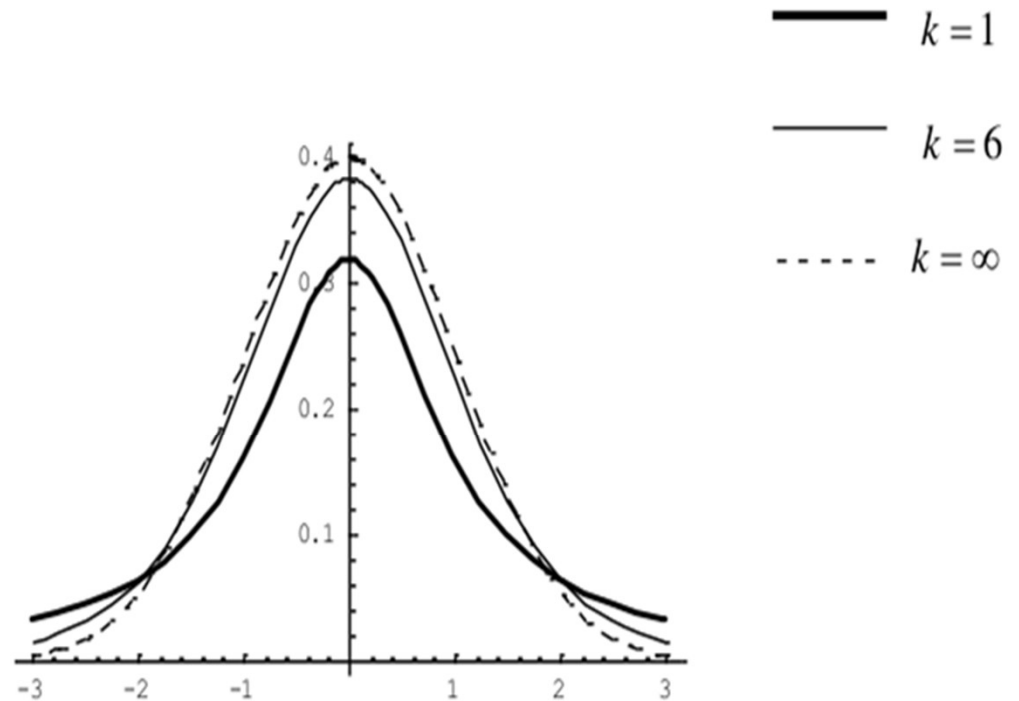
$$P\left(\underbrace{\bar{X} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}}_I \leq \mu \leq \underbrace{\bar{X} + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}}_S\right) \approx 1 - \alpha$$

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

# IC para n chico, t de Student

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

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





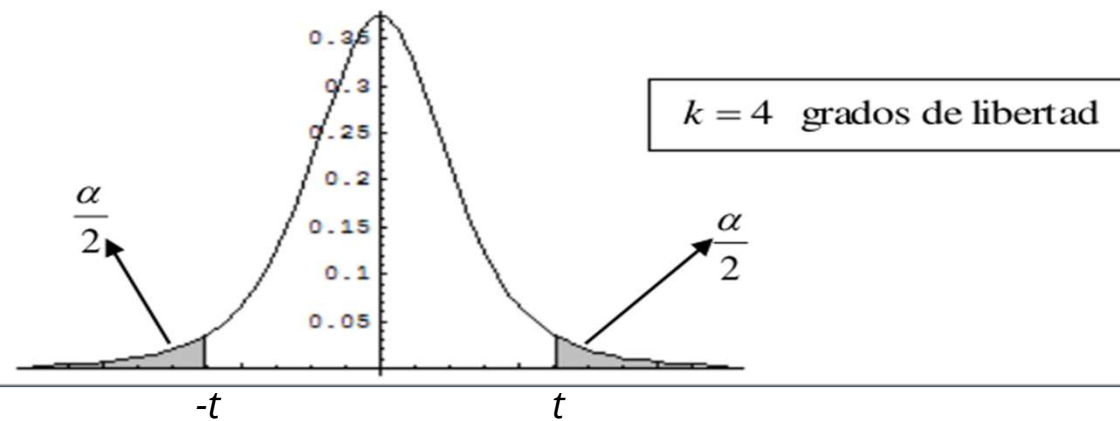
# IC para $n$ chico, $t$ de Student con $(n-1)$ g.l.

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

Evidentemente, si definimos

$\begin{cases} \hat{\Theta}_1 = \bar{X} - t \frac{S}{\sqrt{n}} \\ \hat{\Theta}_2 = \bar{X} + t \frac{S}{\sqrt{n}} \end{cases}$ , hemos construido dos estadísticos  $\hat{\Theta}_1$  y  $\hat{\Theta}_2$  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 $\mu$ de una normal con $\sigma$ desconocido

Si  $(X_1, X_2, \dots, X_n)$  una muestra aleatoria de tamaño  $n$  de una v.a.  $X$  donde  $X \sim N(\mu, \sigma^2)$ ,  $\sigma^2$  desconocido, un intervalo de confianza para  $\mu$  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 dif de medias $\mu_1 - \mu_2$

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

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

Varianzas iguales, pivote:

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

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

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.  $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\rho^*(\sigma_1 \sigma_2))$

---

# IC para $\sigma^2$ de una normal

Supongamos que se quiere hallar un intervalo de confianza para  $\sigma^2$  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  
conocida**

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

  
**Pivote con distribución**

$$\chi^2_{(n-1)}$$

# IC para $\sigma^2$

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  $\sigma$  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  $\sigma^2$  y con este un IC para **el verdadero desvío  $\sigma$**  del volumen de llenado.

---

## IC para $\sigma$

Solución:

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

Se asume que  $X \sim N(\mu, \sigma^2)$  con  $\sigma$  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  $\sigma$  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**  $\sigma < 0.15$