

## 7. Estimación de Intervalos de Confianza

Estadística Computacional - San Joaquín - 2022

Ricardo Ñanculef, Manuel Goyo  
*jnancu@inf.utfsm.cl, manuel.goyo@usm.cl*

Departamento de Informática UTFSM



**Departamento de Informática**  
Universidad Técnica Federico Santa María

# Motivación

## Ejemplo

En un estudio acerca de la obesidad infantil, se ha medido el peso  $X$  de un grupo de 20 niños chilenos, obteniéndose un promedio muestral de  $\bar{X} = 40$  Kg.

- ▶ ¿Qué significa estimar el valor esperado  $X$ ?
- ▶ ¿Podemos estimar el valor esperado  $X$ ?

# Motivación

## Estimación Puntual

Hemos visto anteriormente que un *estimador puntual* del valor esperado es el promedio muestral  $\bar{X}_n$ .

- ▶ Los métodos de momentos y máxima verosimilitud permiten justamente obtener estos estimadores puntuales.
- ▶ Incerteza de la estimación: ¿Qué tanto podemos confiar en esta estimación?
- ▶ Si la muestra fuese de 400 niños en vez de 20, ¿Cómo cambia nuestra estimación?

# Desiderata & Lenguaje

## Estimación Confidencial

Dada una muestra  $S = \{X_1, X_2, \dots, X_n\}$  y un parámetro desconocido  $\theta$ , una **estimación confidencial** de  $\theta$  consiste en determinar un intervalo de la forma

$$[\hat{\theta}_1, \hat{\theta}_2]$$

que contenga el valor verdadero de  $\theta$  con alta probabilidad. Al intervalo  $[\hat{\theta}_1, \hat{\theta}_2]$  se le llama **Intervalo de Confianza (IC) para  $\theta$** . Al grado de confianza con que  $\theta$  cae dentro del intervalo

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

se le denomina **nivel de confianza** del I.C. Finalmente, al valor  $\alpha$  se le llama **nivel de significación** o **significancia** del I.C.

# Intervalos de Confianza

## Técnica del Pivoteo

Recordemos que disponemos de una muestra IID  $S = \{X_1, X_2, \dots, X_n\}$  de  $X$ .

1. Encontrar una **cantidad pivotal**, es decir, una función  $Z_S(\theta)$  que depende de la muestra  $S$ , que depende del parámetro desconocido  $\theta$ , y cuya distribución es conocida.
2. Usamos la distribución conocida de  $Z$  para construir un intervalo  $[\hat{z}_1, \hat{z}_2]$  tal que

$$P(Z \in [\hat{z}_1, \hat{z}_2]) = 1 - \alpha$$

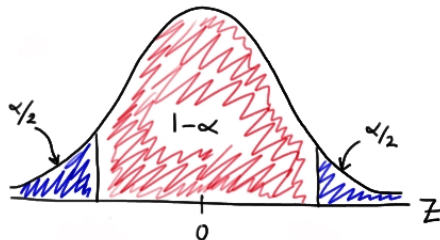
## Intervalos de Confianza

## Técnica del Pivoteo

2 Usamos la distribución conocida de  $Z$  para construir un intervalo  $[\hat{z}_1, \hat{z}_2]$  tal que

$$P(Z \in [\hat{z}_1, \hat{z}_2]) = 1 - \alpha$$

Por ejemplo, el intervalo  $\hat{z}_1 = F^{-1}(\alpha/2)$ ,  $\hat{z}_2 = F^{-1}(1 - \alpha/2)$  satisface por definición el requerimiento.



## Intervalos de Confianza

## Técnica del Pivoteo

- 3 Como  $Z$  depende de  $\theta$ , podemos partir del intervalo  $[\hat{z}_1, \hat{z}_2]$  y despejar  $\theta$  como función de  $S$ ,  $\hat{z}_1$ , y  $\hat{z}_2$ ,

$$P(\hat{z}_1 \leq Z_S(\theta) \leq \hat{z}_2) = 1 - \alpha$$

$$P\left(Z_S^{-1}(\hat{z}_1) \leq \theta \leq Z_S^{-1}(\hat{z}_2)\right) = 1 - \alpha$$

- 4 Como  $\hat{z}_1, \hat{z}_2$  son conocidos, el intervalo  $[\hat{\theta}_1, \hat{\theta}_2]$  con

$$\hat{\theta}_1 = Z_S^{-1}(\hat{z}_1) \quad \text{y} \quad \hat{\theta}_2 = Z_S^{-1}(\hat{z}_2)$$

se podrá calcular desde la muestra, y con probabilidad  $1 - \alpha$ , cubrirá el valor desconocido de  $\theta$ !

# Intervalos de Confianza

## Técnica del Pivoteo

El paso difícil es el primero:

1. Encontrar una **cantidad pivotal**, es decir, una función  $Z_S(\theta)$  que depende de la muestra  $S$ , que depende del parámetro desconocido  $\theta$ , y cuya distribución es conocida.

Revisaremos cantidades pivotaes que nos permitirán construir intervalos para los parámetros de interés más importantes en la práctica.



## Ejemplo: I.C. para $\mu = E[X]$ (asumiendo la varianza conocida)

- Supongamos que  $X \sim \mathcal{N}(\mu, \sigma^2)$  y que conocemos  $\sigma^2$ .
- Como la muestra es IID,  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  y además  $X_i$  es independiente de  $X_j$  para todo  $i \neq j$ .

### Problema

Estimar un I.C. para  $E[X] = \mu$ .

¿Qué cantidad pivotal podemos usar?

## Reproductividad de Normales

La teoría viene en nuestro auxilio:

### Teorema: Reproductividad de la Normal

Sea  $S = \{X_1, X_2, \dots, X_n\}$  una colección de variables independientes tal que  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \forall i = 1, \dots, n$ . Entonces

$$\sum_i X_i \sim \mathcal{N}\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$$

► Corolario:

$$\sum_i X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

¿Qué podemos decir de  $\bar{X}$ ?

## Cantidad Pivotal para el Problema

### Teorema Clave: Distribución de Muestreo de $\bar{X}$

Sea  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  la media muestral. Entonces, si  $X \sim \mathcal{N}(\mu, \sigma^2)$  y la muestra es IID,

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

## Cantidad Pivotal para el Problema

- Notemos que  $Z$  satisface todos los requisitos de una cantidad pivotal.

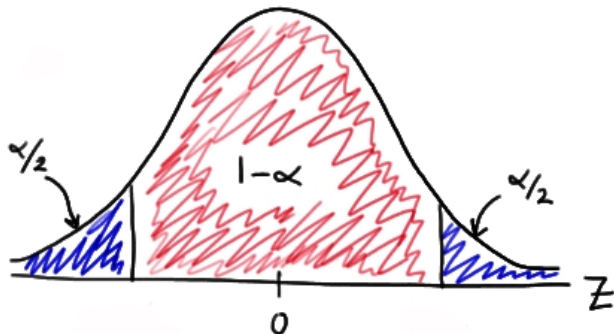
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

1. Depende de  $\mu$  y de una cantidad que podemos evaluar en la muestra  $\bar{X}$ .
  2. Asumiendo  $\sigma^2$  conocida, su distribución es conocida.
- Dado un nivel de significación  $\alpha$ . Entonces podemos encontrar  $\hat{z}_1, \hat{z}_2$  tal que

$$P(Z \in [\hat{z}_1, \hat{z}_2]) = 1 - \alpha$$

## Ejemplo: I.C. para $\mu = E[X]$ (asumiendo varianza conocida)

- En efecto, basta elegir  $\hat{z}_1 = \Phi^{-1}(\alpha/2)$ ,  $\hat{z}_2 = \Phi^{-1}(1 - \alpha/2)$  con  $\Phi$  la función de distribución de la normal estándar.



Sea  $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ . Por simetría,  $z_{1-\alpha/2} = \Phi^{-1}(\alpha/2) = -z_{\alpha/2}$ .

## Ejemplo: I.C. para $\mu = E[X]$ (asumiendo varianza conocida)

- Además, podemos despejar  $\mu$  en función de  $\bar{X}, \hat{z}_1, \hat{z}_2$ ,

$$P(\hat{z}_1 \leq Z(\mu) \leq \hat{z}_2) = 1 - \alpha$$

## Ejemplo: I.C. para $\mu = E[X]$ (asumiendo varianza conocida)

- Además, podemos despejar  $\mu$  en función de  $\bar{X}, \hat{z}_1, \hat{z}_2$ ,

$$P(\hat{z}_1 \leq Z(\mu) \leq \hat{z}_2) = 1 - \alpha$$

- En efecto,

$$\begin{aligned} P(\hat{z}_1 \leq Z(\mu) \leq \hat{z}_2) &= 1 - \alpha \\ \Rightarrow P\left(\hat{z}_1 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \hat{z}_2\right) &= 1 - \alpha \\ \Rightarrow P(\sigma\hat{z}_1/\sqrt{n} \leq (\bar{X} - \mu) \leq \hat{z}_2\sigma/\sqrt{n}) &= 1 - \alpha \\ \Rightarrow P(-\hat{z}_1\sigma/\sqrt{n} \geq (\mu - \bar{X}) \geq -\hat{z}_2\sigma/\sqrt{n}) &= 1 - \alpha \\ \Rightarrow P(\bar{X} - \hat{z}_1\sigma/\sqrt{n} \geq \mu \geq \bar{X} - \hat{z}_2\sigma/\sqrt{n}) &= 1 - \alpha \end{aligned}$$

- Reemplazando,  $\hat{z}_1 = \Phi^{-1}(\alpha/2)$ ,  $\hat{z}_2 = \Phi^{-1}(1 - \alpha/2)$

$$P(\mu \in [\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}]) = 1 - \alpha$$

## Resultado (& Recordatorio sobre Notación)

### IC para la media con varianza conocida

Asumiendo que la distribución de la población es normal, que la muestra es IID y que la **varianza es conocida**, un intervalo de confianza, a un nivel de significación  $\alpha$  para la media  $\mu = E[X]$  está dado por

$$\left[ \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right]$$

donde  $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$  y  $\Phi$  es la función de distribución acumulada (fda) de la normal estándar.



## Ejemplo: I.C. para $\mu = E[X]$ (asumiendo varianza conocida)

### Ejemplo

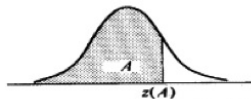
En un estudio acerca de la obesidad infantil en Chile se ha medido el peso de un grupo de 400 niños, obteniéndose un promedio muestral de  $\bar{X} = 40$  Kg. Se sabe que la variable peso  $X$  se distribuye aproximadamente Normal con varianza 1000.

- Determine un IC al 98 % de confianza para  $E[X] = \mu$ .
- ¿Cuántas observaciones adicionales se requieren para reducir a la mitad la incertidumbre alrededor de  $\bar{X}$ ? (manteniendo el nivel de significancia del I.C.).

## Ejemplo: I.C. para $\mu = E[X]$ (asumiendo varianza conocida)

### Ejemplo

- $\alpha = 0.02$ . Para obtener esa significancia basta elegir  $z_{0.01} = \Phi^{-1}(0.99)$ .
- De la Tabla de distribución acumulada para la normal estándar:  
 $z_{0.01} = \Phi^{-1}(0.99) \approx 2.33$



$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

## Ejemplo: I.C. para $\mu = E[X]$ (asumiendo varianza conocida)

### Ejemplo

- ▶  $\alpha = 0.02$ ,  $z_{0.01} = \Phi^{-1}(0.99) \approx 2.33$ ,  $n = 400$ ,  $\bar{X} = 40$ ,  $\sigma^2 = 1000$ .
- ▶ Intervalo:

$$\frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{2.33 \cdot \sqrt{1000}}{\sqrt{400}} = 3.68$$

$$\mu \in \left[ \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right] \longrightarrow [40 - 3.68, 40 + 3.68]$$

$$\mu \in [36.32, 43.68] \text{ Kg}$$

## Ejemplo: I.C. para $\mu = E[X]$ (asumiendo varianza conocida)

### Ejemplo

- ▶ ¿Cuántas observaciones adicionales se requieren para reducir a la mitad la incertidumbre alrededor de  $\bar{X}$ ?
- ▶ Incertidumbre alrededor de la media:

$$\frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{2.33 \cdot \sqrt{1000}}{\sqrt{n}} = 1.84$$

Despejando  $n$  para tener una incertidumbre de  $\pm 1.84$

$$\frac{2.33 \cdot \sqrt{1000}}{\sqrt{n}} = 1.84 \rightarrow n = (40)^2 \approx 1600 \text{ niños}$$

## Observaciones

### Poblaciones No-normales

Si la distribución de la población no es normal, el intervalo anterior sigue siendo **asintóticamente válido** bajo las condiciones del Teorema del Límite Central (capítulo anterior). Es decir,

$$\left[ \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right]$$

se acepta como un IC aproximadamente correcto para la media poblacional ( $E[X]$ ) si  $n$  es razonablemente grande.

# Observaciones

## Colas del IC y Simetría

- ▶ En un cierto sentido, el IC anterior es arbitrario, porque es posible elegir infinitos  $\hat{\theta}_1$  y  $\hat{\theta}_2$  que satisfagan la propiedad de “cobertura” del parámetro  $\theta$  requerida.
- ▶ Por ejemplo, podemos elegir ...

$$\left( -\infty, \bar{X} + \frac{z_{\alpha}\sigma}{\sqrt{n}} \right]$$

## Observaciones

### Colas del IC y Simetría

- ▶ En un cierto sentido, el IC anterior es arbitrario, porque es posible elegir infinitos  $\hat{\theta}_1$  y  $\hat{\theta}_2$  que satisfagan la propiedad de “cobertura” del parámetro  $\theta$  requerida.
- ▶ Por ejemplo, podemos elegir ...

$$\left( -\infty, \bar{X} + \frac{z_{\alpha}\sigma}{\sqrt{n}} \right]$$

- ▶ El I.C. anterior es asimétrico y se ocasionalmente se denominará **intervalo de 1-cola superior** o one-sided (upper) confidence interval.
- ▶ Análogamente  $\left[ \bar{X} - \frac{z_{\alpha}\sigma}{\sqrt{n}}, \infty \right)$  se denominará **intervalo de 1-cola inferior** o one-sided (lower) confidence interval.
- ▶ Cuando sea necesario, el I.C. simétrico derivado anteriormente se llamará **intervalo de 2-colas** o two-sided confidence interval.

## Observaciones

- El **largo** del intervalo,  $L = \hat{\theta}_2 - \hat{\theta}_1$ , es inversamente proporcional a la precisión en la estimación.
- Si queremos disminuir el largo del intervalo, por ejemplo, aumentando  $\alpha$ , aumentaremos la precisión de la estimación, pero reduciremos la confianza que el parámetro esté en ese intervalo:  $P(\theta \in [\hat{\theta}_1, \hat{\theta}_2]) = 1 - \alpha$ .

¿ $\alpha = 1$  ó  $\alpha = 0$ ?



## I.C. en casos sin conocer la varianza.

- ▶ ¿Qué ocurre si no conocemos la varianza poblacional  $\sigma^2$ ?
- ▶ Podríamos estimarla también. Necesitaremos una cantidad Pivotal.

## I.C. en casos sin conocer la varianza.

- ▶ ¿Qué ocurre si no conocemos la varianza poblacional  $\sigma^2$ ?
- ▶ Podríamos estimarla también. Necesitaremos una cantidad Pivotal.

### Teorema Clave: Distribución de Muestreo de $S^2$

Sea

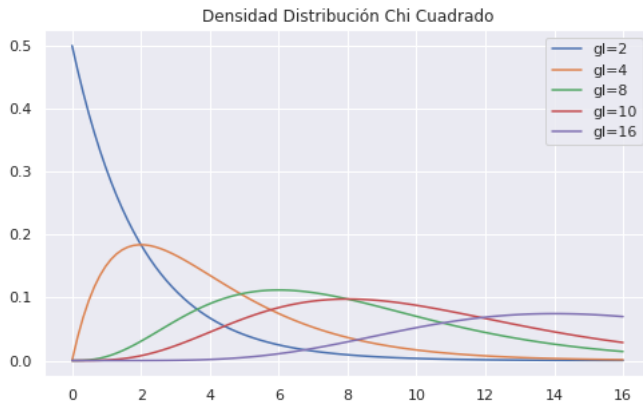
$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

Si  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  y  $X_i$  es independiente de  $X_j$ ,  $\forall i \neq j$ ,

$$Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Recordemos que  $\chi_\nu^2 \equiv \text{Gamma}(\nu/2, 2)$ .

# Distribución Chi Cuadrado



## I.C. para $E[X]$ con varianza desconocida

- ▶ Supongamos que  $X \sim \mathcal{N}(\mu, \sigma^2)$  y que **NO** conocemos  $\sigma^2$ .
- ▶ Supongamos que la muestra es IID.
- ▶ ¿Cómo estimamos un I.C. para  $E[X] = \mu$ ?
- ▶ ¿Tenemos cantidades pivotaes para  $\mu$  y  $\sigma$ ?

## I.C. para $E[X]$ con varianza desconocida

- ▶ Supongamos que  $X \sim \mathcal{N}(\mu, \sigma^2)$  y que **NO** conocemos  $\sigma^2$ .
- ▶ Supongamos que la muestra es IID.
- ▶ ¿Cómo estimamos un I.C. para  $E[X] = \mu$ ?
- ▶ ¿Tenemos cantidades pivotaes para  $\mu$  y  $\sigma$ ?

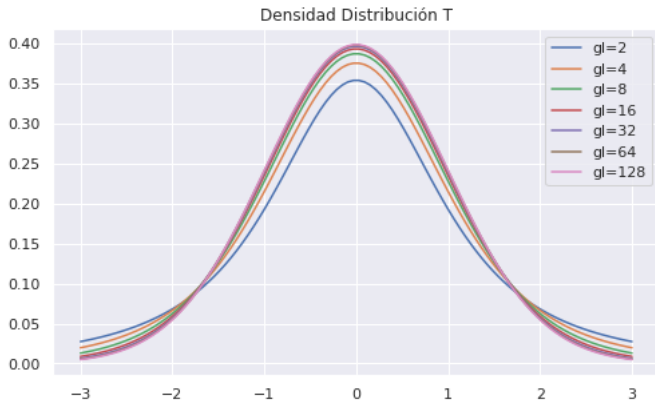
### Teorema: Cociente de una Normal Estándar y una Raíz de Chi-cuadrado Normalizada

Sea  $W \sim \mathcal{N}(0, 1)$  y  $V \sim \chi_\nu^2$ . Entonces,

$$T = \frac{W}{\sqrt{V/\nu}} \sim \mathcal{T}_\nu$$

donde  $\mathcal{T}_\nu$  denota la distribución de t-student con  $\nu$  grados de libertad.

# Distribución T “de Student”



## Distribución T “de Student”

### Definición

Una v.a. sigue una distribución T con parámetro  $\nu$  si su fdp está dada por

$$f(x) = \left( \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \right) \left( 1 + \frac{x^2}{\nu} \right)^{-(\nu+1)/2}.$$

- ▶  $E[X] = 0$  para  $\nu > 1$ .
- ▶ Mediana  $\tilde{X} = 0$  para  $\nu > 1$ .
- ▶  $Var[X] = \frac{\nu}{\nu-2}$  para  $\nu > 2$ .
- ▶ A  $\nu$  se le denomina el *número de grados de libertad* de la variable.
- ▶ Sus percentiles no son analíticos, pero se encuentran tabulados en múltiples fuentes.

## Resultado

### I.C. para $E[X]$ con varianza desconocida

Asumiendo normalidad y que la muestra es IID, un intervalo de confianza a un nivel de significación  $\alpha$  para la media poblacional ( $\mu = E[X]$ ) está dado por

$$\left[ \bar{X} - \frac{t_{\alpha/2}^{n-1} S}{\sqrt{n}}, \bar{X} + \frac{t_{\alpha/2}^{n-1} S}{\sqrt{n}} \right]$$

donde  $t_{\alpha/2}^{n-1} = F_{T_{n-1}}^{-1}(1 - \alpha/2)$ ,  $F_{T_{n-1}}$  es la función de distribución acumulada (fda) de la distribución de t-student con  $n - 1$  grados de libertad y

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

- Observación: Cuando  $n > 50$ , la  $t$ -student es prácticamente indistinguible de la normal estándar.



## Ejemplo: I.C. para $E[X]$ con varianza desconocida

### Ejemplo

En un estudio acerca de la obesidad infantil en Chile se ha medido el peso de un grupo de 20 niños, obteniéndose un promedio muestra de  $\bar{X} = 40$  Kg y un valor de  $S^2 = 400\text{Kg}^2$ . Se sabe que la variable peso  $X$  **con media y varianza desconocidas**,

- Determine un IC al 90 % de confianza para  $E[X] = \mu$ .
- ¿Cuántas observaciones adicionales se requieren para reducir a la mitad la incertidumbre alrededor de  $\bar{X}$ ?

Ejemplo: I.C. para  $E[X]$  con varianza desconocida

## Ejemplo

- $\alpha = 0.10$ ,  $n = 20$ . Basta elegir  $t_{0.05}^{19} = F_{T_{19}}^{-1}(0.95)$ .
- De una tabla de percentiles para la  $t$ -student obtenemos  $t_{0.05}^{19} = F_{T_{19}}^{-1}(0.95) \approx 1.729$ :

TABLA F (continuación) Valores de cuantiles de la distribución  $t$  de Student

$\nu$	$t_{0.800}$	$t_{0.900}$	$t_{0.950}$	$t_{0.975}$	$t_{0.990}$	$t_{0.995}$	$t_{0.999}$
10	0.879	1.372	1.812	2.228	2.764	3.169	4.144
11	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.870	1.350	1.771	2.160	2.650	3.012	3.852
14	0.868	1.345	1.761	2.145	2.624	2.977	3.787
15	0.866	1.341	1.753	2.131	2.602	2.947	3.733
16	0.865	1.337	1.746	2.120	2.583	2.921	3.686
17	0.863	1.333	1.740	2.110	2.567	2.898	3.646
18	0.862	1.330	1.734	2.101	2.552	2.878	3.610
19	0.861	1.328	1.729	2.093	2.539	2.861	3.579
20	0.860	1.325	1.725	2.086	2.528	2.845	3.552

## Ejemplo I.C. para $E[X]$ con varianza desconocida

### Ejemplo

►  $\alpha = 0.10$ ,  $n = 20$ ,  $t_{0.05}^{19} \approx 1.729$ ,  $\bar{X} = 40$ ,  $S^2 = 400$ .

► Intervalo:

$$\frac{t_{\alpha/2}^{n-1} S}{\sqrt{n}} = \frac{1.729 \cdot 20}{\sqrt{20}} \approx 7.73$$

$$\mu \in \left[ \bar{X} - \frac{t_{\alpha/2}^{n-1} S}{\sqrt{n}}, \bar{X} + \frac{t_{\alpha/2}^{n-1} S}{\sqrt{n}} \right] \longrightarrow [40 - 7.73, 40 + 7.73]$$

$$\mu \in [32.27, 47.73] \text{ Kg}$$

## I.C. para la Varianza

- ▶ Dada  $X \sim \mathcal{N}(\mu, \sigma^2)$  con  $\mu, \sigma^2$  desconocidos.
- ▶ Dada una muestra es IID de  $X$ .

### Problema

Estimar un I.C. para  $\sigma^2 = \text{Var}[X]$ .

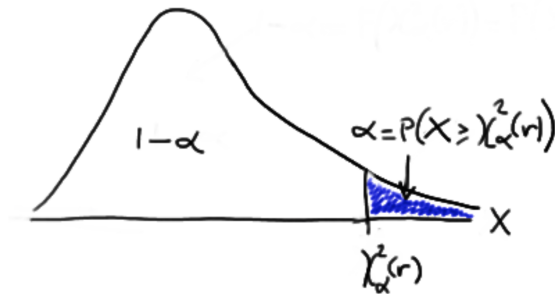
- ▶ Ojo: con frecuencia se denotará por  $\sigma^2$  la varianza de una v.a. aún si la población no es normal.

## I.C. para la Varianza

► Como

$$Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

- Basta encontrar los percentiles  $\chi_{\alpha/2}^{n-1}$  y  $\chi_{1-\alpha/2}^{n-1}$  de la distribución chi cuadrado con  $n-1$  grados de libertad.

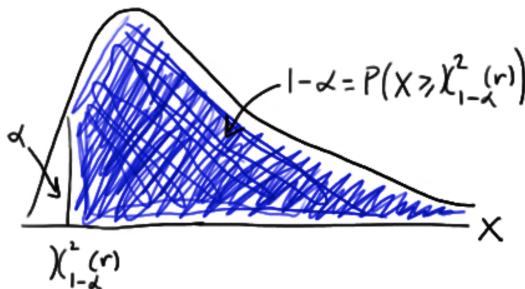


## I.C. para la Varianza

► Como

$$Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

- Basta encontrar los percentiles  $\chi_{\alpha/2}^{n-1}$  y  $\chi_{1-\alpha/2}^{n-1}$  de la distribución chi cuadrado con  $n-1$  grados de libertad.



## I.C. para la Varianza

### Solución

Asumiendo que la población es normal y que la muestra es IID, un I.C. a un nivel de significación  $\alpha$  para  $\sigma^2 = \text{Var}[X]$  está dado por

$$\left[ \frac{(n-1)S^2}{\chi_{\alpha/2}^{n-1}}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}^{n-1}} \right]$$

donde  $\chi_{\alpha/2}^{n-1}$  y  $\chi_{1-\alpha/2}^{n-1}$  son los percentiles  $(1 - \alpha/2)100$ -ésimo y  $(\alpha/2)100$ -ésimo de la distribución chi cuadrado con  $n - 1$  grados de libertad.

► **Warning:** La distribución chi cuadrado NO ES SIMÉTRICA, i.e.,  $\chi_{\alpha/2} \neq \chi_{1-\alpha/2}$

## I.C. para $E[X] - E[Y]$ (asumiendo varianzas conocidas)

### Problema

Supongamos que tenemos dos v.a.  $X$  e  $Y$  y queremos un I.C. para  $E[X - Y] = E[X] - E[Y]$ .

- Recogemos una muestra para cada variable aleatoria  $\underline{X} = (X_1, X_2, \dots, X_{n_X})$ ,  $\underline{Y} = (Y_1, Y_2, \dots, Y_{n_Y})$  (podemos tener  $n_X \neq n_Y$ )
- Supongamos que  $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$  y que conocemos  $\sigma_X^2$ .
- Supongamos que  $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$  y que conocemos  $\sigma_Y^2$ .
- Supongamos que las muestras son ambas IID.
- Sabemos que  $\mu_X = E[X]$ ,  $\mu_Y = E[Y]$ .
- Un I.C. para  $E[X] - E[Y]$  es entonces un I.C. para  $\mu_X - \mu_Y$ .



I.C. para  $E[X] - E[Y]$  (asumiendo varianzas conocidas)

## Teorema Clave: Distribución de Muestreo de la Diferencia de dos Promedios

Sea  $\bar{X}, \bar{Y}$  las medias muestrales. Entonces, bajo los supuestos anteriores,

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0, 1)$$

- La cantidad anterior nos sirve como cantidad pivotal para despejar  $\mu_X - \mu_Y$  en función de cantidades muestrales, tal como lo hicimos con  $\mu$  en el caso 1.

## Resultado

### I.C. para $E[X] - E[Y]$ (asumiendo varianzas conocidas)

Asumiendo la normalidad de ambas poblaciones, que las muestras son IID y que las varianzas poblacionales son conocidas, un intervalo de confianza con nivel de significación  $\alpha$  para  $E[X] - E[Y] = \mu_X - \mu_Y$  está dado por

$$\left[ (\bar{X} - \bar{Y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right]$$

donde  $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$  y  $\Phi$  es la función de distribución acumulada (fda) de la normal estándar.

I.C. para  $E[X] - E[Y]$  (asumiendo varianzas desconocidas, pero iguales)

- ¿Qué sucede si desconocemos  $\sigma_X^2$  y  $\sigma_Y^2$  pero sabemos que son idénticas?
- Consideremos el estimador combinado de la varianza muestral,

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2};$$

con  $S_X^2 = \frac{\sum_{i=1}^{n_X} (X_i - \bar{X})^2}{n_X - 1}$ ,  $S_Y^2 = \frac{\sum_{j=1}^{n_Y} (Y_j - \bar{Y})^2}{n_Y - 1}$ ,

# Reproductividad de la Chi Cuadrado

## Teorema

Sean  $V \sim \chi^2_{\nu_1}$  y  $W \sim \chi^2_{\nu_2}$  dos v.a. independientes. Entonces

$$V + W \sim \chi^2_{\nu_1 + \nu_2}$$

## Cantidad Pivotal bajo Homocedasticidad

### Teorema Clave: Distribución de Muestreo de la Diferencia de dos Medias bajo Homocedasticidad

Sea  $\bar{X}$ ,  $\bar{Y}$  las medias muestrales. Bajo los supuestos detallados anteriormente, y asumiendo  $\sigma_X^2 = \sigma_Y^2$ , tenemos que

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim \mathcal{T}_{n_X + n_Y - 2}$$

donde  $\mathcal{T}_{n_X + n_Y - 2}$  denota la distribución de t-student con  $n_X + n_Y - 2$  grados de libertad.

## Resultado

### I.C. para $E[X] - E[Y]$ bajo Homocedasticidad

Bajo los supuestos detallados anteriormente, en particular asumiendo  $\sigma_X^2 = \sigma_Y^2$ , un intervalo de confianza a un nivel de significación  $\alpha$  para  $E[X] - E[Y] = \mu_X - \mu_Y$  está dado por

$$\left[ (\bar{X} - \bar{Y}) \pm t_{\alpha/2}^{n_X+n_Y-2} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \right]$$

donde  $S_p$  es el estimador combinado de la varianza,  $t_{\alpha/2}^{n_X+n_Y-2} = F_{T_{n_X+n_Y-2}}^{-1}(1 - \alpha/2)$  y  $F_{T_{n_X+n_Y-2}}$  es la función de distribución acumulada (fda) de la distribución de  $t$ -student con  $n_X + n_Y - 2$  grados de libertad.

- Observación: Cuando  $n_X + n_Y - 2 > 50$ , la  $t$ -student es prácticamente indistinguible de la normal estándar.

## Ejemplo: estimación I.C. para diferencia de medias

### Ejemplo

Estudiando el desempeño de dos maquinarias en una fábrica, asumiendo que tienen la misma varianza poblacional, se recolectaron dos muestras con la siguiente información:  $S_X^2 = 6.30$ ,  $\bar{X} = 10.26$ , con  $n_X = 10$ , y  $S_Y^2 = 3.61$ ,  $\bar{Y} = 9.02$ , con  $n_Y = 10$ .

- ▶ Construya un intervalo de confianza para la diferencia entre las medias del desempeño de las máquinas a un nivel significancia de  $\alpha = 0.05$ .
- ▶ ¿Es posible concluir que la media del desempeño de una maquinaria será mayor que la otra?

## Ejemplo - Resolución

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2} = \frac{(10 - 1) \cdot 6.30 + (10 - 1) \cdot 3.61}{10 + 10 - 2} = 4.955$$

### Ejemplo

- ▶ Tenemos  $S_p = \sqrt{4.955} = 2.226$ .
- ▶ Con  $\alpha = 0.05$ , tenemos  $t_{0.025}^{18} = F_{T18}^{-1}(0.975)$ . De la tabla  $t$ -student:  $t_{0.025}^{18} = 2.101$ .
- ▶ Despejamos:

$$t_{\alpha/2}^{n_X + n_Y - 2} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} = 2.101 \cdot 2.226 \sqrt{\frac{1}{10} + \frac{1}{10}} = 2.092$$

Entonces el intervalo corresponde a :

$$\mu_X - \mu_Y \in [(\bar{X} - \bar{Y}) \pm 2.092] = [1.24 \pm 2.092] = (-0.852, 3.332)$$

- ▶ ¿Es posible concluir que la media del desempeño de una será mayor que la otra?
  - **NO**, ya que incluye el 0.



## I.C. para $E[X] - E[Y]$ con varianzas desconocidas arbitrarias

- El caso general en el cuál no conocemos las varianzas de las poblaciones y éstas son arbitrarias, i.e. no necesariamente iguales, es más complicado.
- Se suele aceptar el I.C. de Satterthwaite como una aproximación razonable.

### I.C. para $E[X] - E[Y]$ bajo Heterocedasticidad

Un intervalo de confianza aproximado para  $E[X] - E[Y] = \mu_X - \mu_Y$ , a un nivel de significación  $\alpha$ , está dado por

$$\left[ (\bar{X} - \bar{Y}) \pm t_{\alpha/2}^{\nu} \sqrt{\frac{S_X}{n_X} + \frac{S_Y}{n_Y}} \right]$$

donde  $t_{\alpha/2}^{\nu} = F_{T\nu}^{-1}(1 - \alpha/2)$ ,  $F_{T\nu}$  es la función de distribución acumulada (fda) de la distribución de t-student con  $\nu$  grados de libertad donde

$$\nu = \frac{\left( \frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y} \right)^2}{\frac{\left( \frac{S_X^2}{n_X} \right)^2}{n_X - 1} + \frac{\left( \frac{S_Y^2}{n_Y} \right)^2}{n_Y - 1}}$$

## I.C. para el Cuociente de Dos Varianzas

### Problema

Tenemos dos v.a.  $X$  e  $Y$  y queremos un I.C. para  $\text{Var}[Y]/\text{Var}[X]$ .

- Recogemos una muestra para cada variable aleatoria  $\underline{X} = (X_1, X_2, \dots, X_{n_X})$ ,  $\underline{Y} = (Y_1, Y_2, \dots, Y_{n_Y})$  (podemos tener  $n_X \neq n_Y$ ).
- Supongamos que  $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$  y que no conocemos ni  $\mu_X$  ni  $\sigma_X^2$ .
- Supongamos que  $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$  y que no conocemos ni  $\mu_Y$  ni  $\sigma_Y^2$ .
- Un I.C. para  $\text{Var}[Y]/\text{Var}[X]$ . es entonces un I.C. para  $\sigma_Y^2/\sigma_X^2$ .
- Supongamos que las muestras son IID.

## Cuociente de Chi Cuadrados

### Teorema

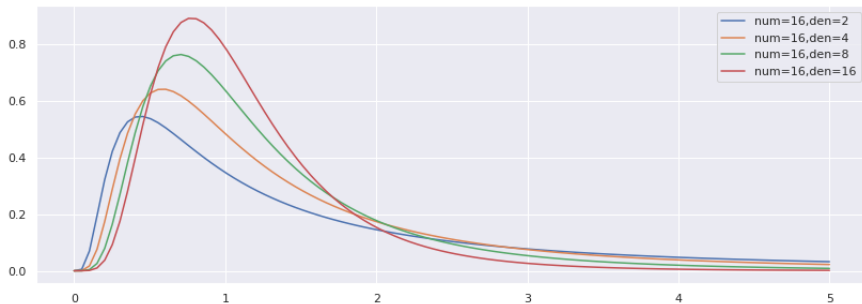
Sean  $V \sim \chi_{\nu_1}^2$  y  $W \sim \chi_{\nu_2}^2$  dos v.a. independientes. Entonces

$$\frac{V/\nu_1}{W/\nu_2} \sim F_{\nu_1, \nu_2}$$

donde  $F_{\nu_1, \nu_2}$  denota la *Distribución F de Fisher-Snedecor* con parámetros  $\nu_1, \nu_2$ . Los parámetros se leen:  $\nu_1$  grados de libertad en el numerador y  $\nu_2$  grados de libertad en el denominador.

# Distribución de Fisher

Densidad Distribución F



# La Distribución F de Fisher

## Definición

Una v.a.  $X$  sigue una distribución F con parámetros  $\nu_1$  y  $\nu_2$  si su fdp está dada por

$$f(x) = \begin{cases} \left( \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \right) \left( \frac{\nu_1 x}{\nu_1 x + \nu_2} \right)^{\nu_1/2} \left( 1 - \frac{\nu_1 x}{\nu_1 x + \nu_2} \right)^{\nu_2/2} x^{-1} & x > 0 \\ 0 & \text{en otro caso} \end{cases}$$

- ▶  $E[X] = \nu_2/(\nu_2 - 2)$  para  $\nu_2 > 2$ .
- ▶  $Var[X] = \frac{\nu_2^2(2\nu_1+2\nu_2-4)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$  para  $\nu_2 > 4$ .
- ▶ A  $\nu_1$  y  $\nu_2$  se les denomina el *número de grados de libertad del numerador* y *número de grados de libertad del denominador* respectivamente.

## Cantidad Pivotal para el Cuociente de Varianzas

### Teorema Clave: Distribución del Cuociente de Dos Varianzas Muestrales

Si tenemos dos muestras IID independientes entre sí  $\underline{X} = (X_1, X_2, \dots, X_{n_X})$ ,  $\underline{Y} = (Y_1, Y_2, \dots, Y_{n_Y})$ , y la población es normal, tenemos que

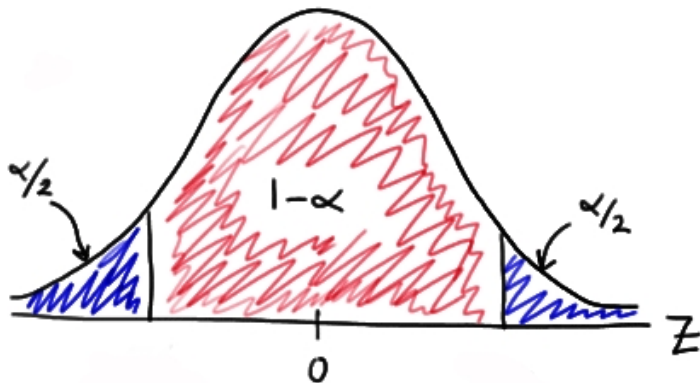
$$F = \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} \sim F(n_X - 1, n_Y - 1)$$

con

$$S_X^2 = \frac{\sum_{i=1}^{n_X} (X_i - \bar{X})^2}{n_X - 1}, \quad S_Y^2 = \frac{\sum_{j=1}^{n_Y} (Y_j - \bar{Y})^2}{n_Y - 1},$$

## La Distribución F de Fisher

- Un intervalo que contenga  $F$  con probabilidad  $1 - \alpha$  se puede encontrar buscando los percentiles  $f_{\alpha/2}^{n_X-1, n_Y-1}$  y  $f_{1-\alpha/2}^{n_X-1, n_Y-1}$  que acumulan probabilidades  $1 - \alpha/2$  y  $\alpha/2$  respectivamente.



## I.C. para el Cuociente de Dos Varianzas

- Usando la misma metodología que hemos venido empleando

$$P\left(f_{1-\alpha/2}^{n_X-1, n_Y-1} \leq \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \leq f_{\alpha/2}^{n_X-1, n_Y-1}\right) = 1 - \alpha$$
$$P\left(f_{1-\alpha/2}^{n_X-1, n_Y-1} \frac{S_Y^2}{S_X^2} \leq \left(\frac{\sigma_Y^2}{\sigma_X^2}\right) \leq f_{\alpha/2}^{n_X-1, n_Y-1} \frac{S_Y^2}{S_X^2}\right) = 1 - \alpha$$



## I.C. para el Cuociente de Dos Varianzas

### Solución

Asumiendo poblaciones normales y que las muestras son IID, un intervalo de confianza a un nivel de significación  $\alpha$  para  $\sigma_Y^2/\sigma_X^2 = \text{Var}[Y]/\text{Var}[X]$  está dado por

$$\left[ f_{1-\alpha/2}^{n_X-1, n_Y-1} \frac{S_Y^2}{S_X^2}, f_{\alpha/2}^{n_X-1, n_Y-1} \frac{S_Y^2}{S_X^2} \right]$$

con

$$S_X^2 = \frac{\sum_{i=1}^{n_X} (X_i - \bar{X})^2}{n_X - 1}, \quad S_Y^2 = \frac{\sum_{j=1}^{n_Y} (Y_j - \bar{Y})^2}{n_Y - 1},$$

y donde  $f_{\gamma}^{\nu_1, \nu_2}$  es el percentil de la distribución F de Fisher-Snedecor con  $\nu_1, \nu_2$  grados de libertad que acumula probabilidad  $1 - \gamma$ .

I.C. para el cuociente  $\sigma_Y^2/\sigma_X^2$

## WARNING

- La distribución  $F$  no es simétrica, pero satisface la siguiente propiedad:

$$f_{\alpha/2}^{n_X-1, n_Y-1} = \frac{1}{f_{1-\alpha/2}^{n_Y-1, n_X-1}}$$

## I.C. para una Proporción

### Problema

Supongamos que queremos estimar la proporción de una población que satisface un cierto criterio, por ejemplo, **¿Qué porcentaje de la población votará por nuestro candidato favorito?.**

- Sea  $X$  la variable aleatoria que representa la satisfacción de nuestro criterio en un individuo aleatorio de la población. Por ejemplo

$$X(\omega) = \begin{cases} 1 & \text{si el individuo } \omega \text{ vota por mi candidato} \\ 0 & \text{en otro caso} \end{cases}$$

- Claramente  $X$  es Bernoulli con parámetro desconocido  $p$ .
- ¿Cómo estimar  $p$ ?

## I.C. para una Proporción

- ▶ Si recogemos una muestra de  $X$  en la población,  $\underline{X} = (X_1 = x_1, \dots, X_n = x_n)$ .
- ▶ Sea  $S^+ = \sum_{i=1}^n X_i$ . ¿Qué podemos decir de la siguiente cantidad?

$$\hat{p} = \frac{1}{n} S^+ = \frac{1}{n} \sum_{i=1}^n X_i$$

## I.C. para una Proporción

- Si recogemos una muestra de  $X$  en la población,  $\underline{X} = (X_1 = x_1, \dots, X_n = x_n)$ .
- Sea  $S^+ = \sum_{i=1}^n X_i$ . ¿Qué podemos decir de la siguiente cantidad?

$$\hat{p} = \frac{1}{n} S^+ = \frac{1}{n} \sum_{i=1}^n X_i$$

- Por el TLC

$$\hat{p} = \frac{1}{n} S^+ = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}(p, p(1-p)/n)$$

- De este modo,

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, 1)$$

## I.C. para una Proporción

Una motivación alternativa del resultado es la siguiente:

- ▶ Si la población es Bernoulli, tenemos que  $S^+ = \sum_{i=1}^n X_i$  es binomial  $\text{Bin}(n, p)$ .
- ▶ Es posible mostrar que  $\text{Bin}(n, p)$  se puede aproximar mediante una normal  $\mathcal{N}(np, np(1-p))$ .
- ▶ Entonces  $S^+$  admitiría una aproximación normal. Más formalmente, tenemos

### Teorema Clave: Distribución de Muestreo aproximada para una Proporción

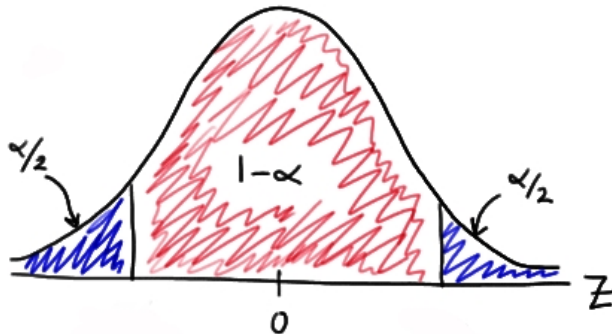
Consideremos la frecuencia empírica del suceso  $\hat{p} = \frac{1}{n} S^+ = \frac{1}{n} \sum_{i=1}^n X_i$ . Entonces,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, 1)$$

donde  $\xrightarrow[n \rightarrow \infty]{\mathcal{D}}$  denota convergencia en distribución.

## I.C. para una Proporción

- Notemos que  $Z$  sirve como cantidad pivotal, ya que tiene distribución conocida y nos permite despejar  $p$  en función de estadísticas muestrales.
- Para contener  $Z$  con probabilidad  $1 - \alpha$  basta elegir  $\hat{z}_1 = \Phi^{-1}(\alpha/2)$ ,  $\hat{z}_2 = \Phi^{-1}(1 - \alpha/2)$  con  $\Phi$  la función de distribución acumulada de la normal estándar. Por simetría,  $z_{1-\alpha/2} = \Phi^{-1}(\alpha/2) = -z_{\alpha/2}$ .



## I.C. para una Proporción

► Despejando  $p$ ,

$$P(\hat{z}_1 \leq Z \leq \hat{z}_2) = 1 - \alpha$$

$$\Rightarrow P\left(\hat{z}_1 \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq \hat{z}_2\right) = 1 - \alpha$$

$$\Rightarrow P\left(\hat{z}_1 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq (\hat{p} - p) \leq \hat{z}_2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

$$\Rightarrow P\left(-\hat{z}_2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq (p - \hat{p}) \leq -\hat{z}_1 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\hat{p} - \hat{z}_2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} - \hat{z}_1 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$



## Resultado

### I.C. para una Proporción

Un intervalo de confianza aproximado para una proporción  $p$  a un nivel de significación  $\alpha$  esta dado por

$$\left[ \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

donde  $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ ,  $\Phi$  es la función de distribución de la normal estándar y  $\hat{p}$  es el valor muestral (empírico) de la proporción.

## Ejemplo: I.C. para una Proporción

### Ejemplo

En una muestra aleatoria simple de la población, con  $n = 1000$ , 550 personas manifestaron la decisión de votar por mi candidato favorito. Determine un I.C. del 99 % de confianza para la probabilidad de que mi candidato gane las elecciones.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, 1)$$

- $\alpha = 0.01$ . Para obtener esa significancia basta elegir  $z_{0.005} = \Phi^{-1}(0.995)$ .
- De la Tabla de distribución acumulada para la normal estándar:  
 $z_{0.005} = \Phi^{-1}(0.995) \approx 2.58$

## Ejemplo: I.C. para una Proporción

$$\hat{p} = 0.55$$

$$z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 2.58 \sqrt{\frac{0.55(1 - 0.55)}{1000}} = 0.0158$$

► Intervalo aproximado para  $p$ ,

$$p \in [0.55 - 0.0158, 0.55 + 0.0158] = [0.5342, 0.5658]$$

## I.C. para la Diferencia de dos Proporciones

### Problema

- ▶ Supongamos ahora que en vez de estimar una proporción  $p$  queremos estimar la diferencia entre dos proporciones.
- ▶ Por ejemplo: la diferencia entre la proporción de personas que votarán por cierto candidato en Santiago versus la proporción que lo hará en Valparaíso.
- ▶ Sea  $X(\omega_1) \sim \text{Ber}(p_1)$  la v.a. que toma el valor 1 si un individuo  $\omega_1$  de Santiago votará por el candidato.
- ▶ Sea  $X(\omega_2) \sim \text{Ber}(p_2)$  la v.a. que toma el valor 1 si un individuo  $\omega_2$  de Valparaíso votará por el candidato.
- ▶ Recogemos sendas muestras IID  $\underline{X} = \{X_1, \dots, X_{n_1}\}$  y de  $\underline{Y} = \{Y_1, \dots, Y_{n_2}\}$ .
- ▶ Si medimos  $\hat{p}_1 = \bar{X}$  y  $\hat{p}_2 = \bar{Y}$ , un estimador puntual de  $p_1 - p_2$  es  $\hat{p}_1 - \hat{p}_2$ .
- ▶ ¿Cómo estimar un IC para  $p_1 - p_2$ ?
- ▶ Supongamos que  $\underline{X}$  y  $\underline{Y}$  son independientes ...

# I.C. para la Diferencia de dos Proporciones

## Teorema Clave

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, 1)$$

donde  $\xrightarrow[n \rightarrow \infty]{\mathcal{D}}$  denota convergencia en distribución.

## Resultado

### I.C. para la Diferencia de dos Proporciones

Un intervalo de confianza aproximado para la diferencia de dos proporciones  $p_1 - p_2$  a un nivel de significación  $\alpha$  está dado por

$$\left[ (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right]$$

donde  $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ ,  $\Phi$  es la función de distribución de la normal estándar y  $\hat{p}_1, \hat{p}_2$  son las proporciones muestrales (empíricas).

► Ojo: El supuesto más fuerte acá es que  $\underline{X}$  y  $\underline{Y}$  son independientes ...

## Teorema del Límite Central

- Recordemos: prácticamente, todo nuestro análisis está basado en suponer que las variables de interés se distribuyen normalmente  $\mathcal{N}(\mu, \sigma^2)$ .
- Eso permite conocer la distribución de  $\bar{X}$  o de  $\text{Var}[X]$ .
- El resultado que nos permite considerar esto una aproximación razonable a la realidad es el T.L.C.

### TLC (De Moivre, Laplace, Lyapunov y otros)

Sean  $X_1, X_2, \dots, X_n$   $n$  variables aleatorias independientes e idénticamente distribuidas con  $E[X_i] = \mu < \infty$  y  $\text{Var}[X_i] = \sigma^2 < \infty$ . Sea  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Entonces,

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, 1)$$

donde  $\xrightarrow[n \rightarrow \infty]{\mathcal{D}}$  denota convergencia en distribución.

## Teorema del Límite Central

- El análisis de  $p$  es una aplicación directa del TLC.
- El resultado anterior NO depende de la distribución de las  $X_i$ .
- El TLC implica que con  $n$  muy grande

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx \mathcal{N}(0, 1)$$

aún si las observaciones  $X_i$  no son normales.

- Esto implica que todos los I.C. estudiados anteriormente son asintóticamente válidos (es decir válidos para  $n$  muy grande) aún si el supuesto de distribución normal para la variables de interés no se cumple.



# Temas

## 1 Introducción

## 2 Intervalos de Confianza

- I.C. para la Media con Varianza Conocida.
- I.C. cuando la Varianza es desconocida
- I.C. para la Media con Varianza Desconocida.
- I.C. para la Varianza
- I.C. para la Diferencia de Medias con Varianzas Conocidas
- I.C. para la Diferencia de Medias bajo Homocedasticidad
- I.C. para la Diferencia de Medias bajo Heterocedasticidad
- I.C. para el Cuociente de Varianzas.
- I.C. para una Proporción
- I.C. para la Diferencia de dos Proporciones

## 3 Teorema del Límite Central