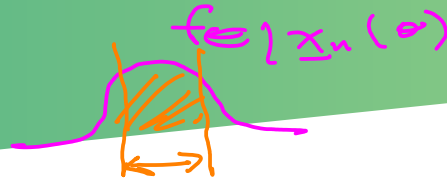


Probabilidad y estadística

Clase 6

Intervalos de confianza

Motivación

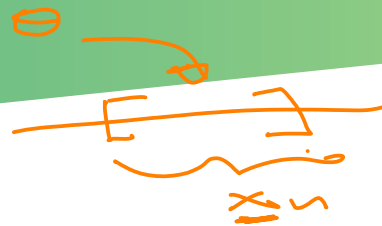


Hasta ahora habíamos visto estimadores puntuales, que, dada una muestra, nos devuelven un único valor $\hat{\theta}$ que se aproxima al valor verdadero del parámetro deseado θ .

$$\hat{\theta}_1 = \hat{\theta}(x_n^1)$$
$$\hat{\theta}_2 = \hat{\theta}(x_n^2)$$

Una forma de obtener información sobre la precisión de la estimación, en el caso de que θ sea unidimensional, es proporcionar un intervalo $[a(\underline{X}), b(\underline{X})]$ de manera que la probabilidad de que dicho intervalo contenga el verdadero valor θ sea alta, por ejemplo, 0.95.

Región de confianza



Def: Dada una m.a. \underline{X} con distribución perteneciente a una familia $F_\theta(x)$, con $\theta \in \Theta$, una **región de confianza** $S(\underline{X})$ para θ con nivel de confianza $1 - \alpha$ será un conjunto tal que

$$\mathbb{P}(\theta \in S(\underline{X})) = 1 - \alpha. (*) = 0,95 \uparrow$$

m.a. \uparrow *región*

Obs: θ **no** es aleatorio, lo aleatorio es $(*)$ es $S(\underline{X})$.

Obs: Si $S(\underline{X}) = (a(\underline{X}), b(\underline{X}))$ diremos que es un **intervalo de confianza**.

Si $S(\underline{X}) = (\min(\Theta), b(\underline{X}))$ diremos que es una **cota superior**.

Si $S(\underline{X}) = (a(\underline{X}), \max(\Theta))$ diremos que es una **cota inferior**.

↳ exp = cota de parámetro

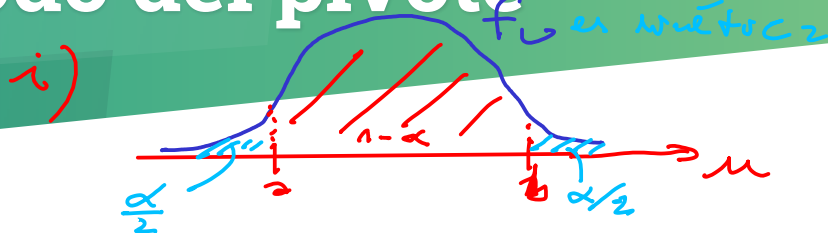
Juguemos un poquito

Usemos la siguiente api para entender mejor qué es un IC:

<http://rossmanchance.com/applets/2021/confsim/ConfSim.html>

Método del pivote

$$a = F_U^{-1}\left(\frac{\alpha}{2}\right), \quad b = F_U^{-1}\left(1 - \frac{\alpha}{2}\right)$$



$$X \sim F_\theta$$

$$U \sim F_{U, \theta} \Rightarrow P(U \leq u) = \frac{\alpha}{2}$$

Teorema: Sea \underline{X} una muestra aleatoria con distribución perteneciente a una familia $F_\theta(x)$, con $\theta \in \Theta$, y sea $U = g(\underline{X}, \theta)$ una variable cuya distribución no depende de θ . Sean a y b tales que

i) $\mathbb{P}(a \leq U \leq b) = 1 - \alpha$. Luego,

ii) $S(\underline{X}) = \{\theta : a < g(\underline{X}, \theta) \leq b\}$

es una región de confianza para θ . A U se lo llama **pivote**.

Ejercicio 1

$$\left(\begin{array}{l} \text{Estadizar:} \\ \underbrace{\frac{X - \mu}{\sigma}} \sim N(0, 1) \end{array} \right) \quad \begin{array}{l} X \sim N(\mu, \sigma^2) \end{array}$$

Sea $\underline{X} = (X_1, \dots, X_n)$ una muestra aleatoria de tamaño n de una población con distribución normal de media μ y varianza 4. Hallar una cota inferior del 95% para μ .

Suponer $n=20$ y $\mu=3$, simular la muestra y obtener el valor de la cota.

$X \sim N(\mu, 4)$, observo una m. a. \underline{X}_n
 y $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$

$$U = g(\underline{X}, \mu) = \frac{\bar{X}_n - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$$

↑
estadizar



$$P(U < a) = 0.95 \Rightarrow a = F_U^{-1}(0.95) = 1.64$$

$$P(U < 1,64) = 0,95$$

$$P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1,64\right) = 0,95$$

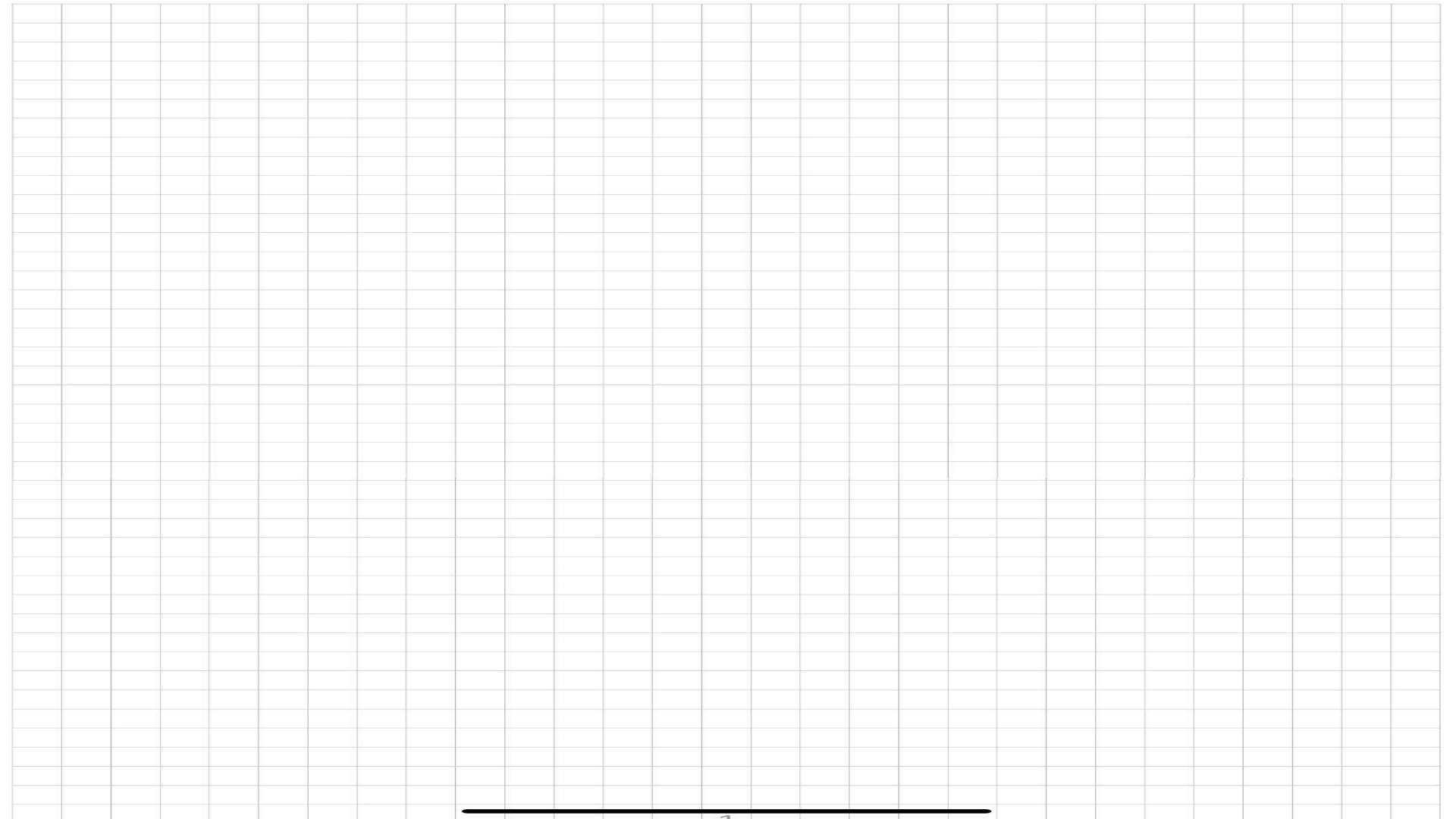
↗ es conocido (y vemos)

$$P\left(\mu > \bar{X} - 1,64 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

$$P\left(\mu \in \underbrace{\left(\bar{X} - 1,64 \frac{\sigma}{\sqrt{n}} ; +\infty\right)}_{S(\bar{X})}\right) = 0,95$$

“zona inferior
de confianza”

$$S(\bar{X}) = \left(\underset{\uparrow}{2,5} ; +\infty \right)$$



Algunos resultados importantes

Teorema: Sea $\underline{X} = X_1, \dots, X_n$ una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$

condición σ^2 y divide por σ $Z = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$ \rightarrow estandarizar el promedio y dividir por σ

prueba σ^2 $W = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{MLE}$$

$$X^2 \sim \chi_1^2$$

V y W son independientes

es insesgado $\text{Si } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, U = \sqrt{n} \frac{(\bar{X} - \mu)}{S} \sim t_{n-1} \rightarrow$ un pivote μ

Obs: en general vale que si $X \sim \mathcal{N}(0, 1)$ y $Y \sim \chi_n^2$, con X e Y

independientes vale que $\frac{X}{\sqrt{Y/n}} \sim t_n$

Algunos pivotes para variables normales

Dada \underline{X}_n una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$ definimos algunos pivotes:

- Para la media con varianza conocida: $U(\underline{X}, \mu) = \frac{(\bar{X} - \mu)}{\sigma} \sqrt{n} \sim \mathcal{N}(0, 1)$
- • Para la media con varianza desconocida: $U(\underline{X}, \mu) = \frac{(\bar{X} - \mu)}{S} \sqrt{n} \sim t_{n-1} \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, 1)$
- Para el desvío con media conocida: $U(\underline{X}, \sigma) = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma} \sim \chi_n^2$
- Para el desvío con media desconocida: $U(\underline{X}, \sigma) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma} \sim \chi_{n-1}^2$

Dada también \underline{Y}_m una m.a. de una distribución $\mathcal{N}(\lambda, \sigma^2)$ y sea Δ :

- Comparación de medias con varianzas conocidas: $U(\underline{X}, \Delta) = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0, 1)$
- Comparación de medias con varianzas desconocidas e iguales:

$$U(\underline{X}, \Delta) = \frac{\bar{X} - \bar{Y} - \Delta}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}, \text{ con } S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{n+m-2}$$

Ejercicio 2

Dada una muestra aleatoria $\underline{X} = (X_1, \dots, X_n)$ de una población con distribución normal con media y varianza desconocidas, hallar el intervalo de confianza de nivel 0.99 para la media de la población.

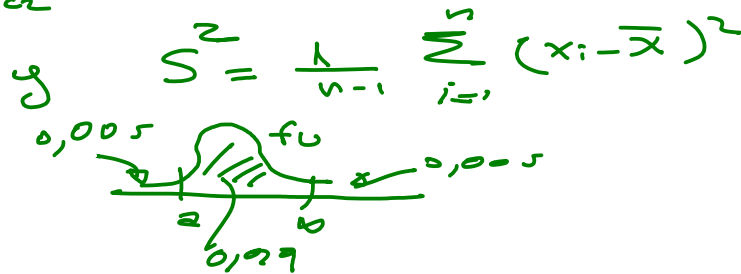
Suponer $n=50$, $\mu = 2$, $\sigma = 3$, simular la muestra y calcular el IC resultante de la misma.

$$U = \frac{\bar{x} - \mu}{s} \sqrt{n} \sim t_{n-1}$$

← "t" de Student

$$P(a < U < b) = 0,99$$

$$a = F_U^{-1}(0,005) = -3,26 \quad . \quad b = -a = 3,26$$



$$P(a < U < b) = 0,99$$

$$P\left(a < \frac{\bar{x} - \mu}{s} \sqrt{n} < b\right) = 0,99$$


$$P\left(\bar{x} - a \frac{s}{\sqrt{n}} > \mu > \bar{x} - b \frac{s}{\sqrt{n}}\right) = 0,99$$

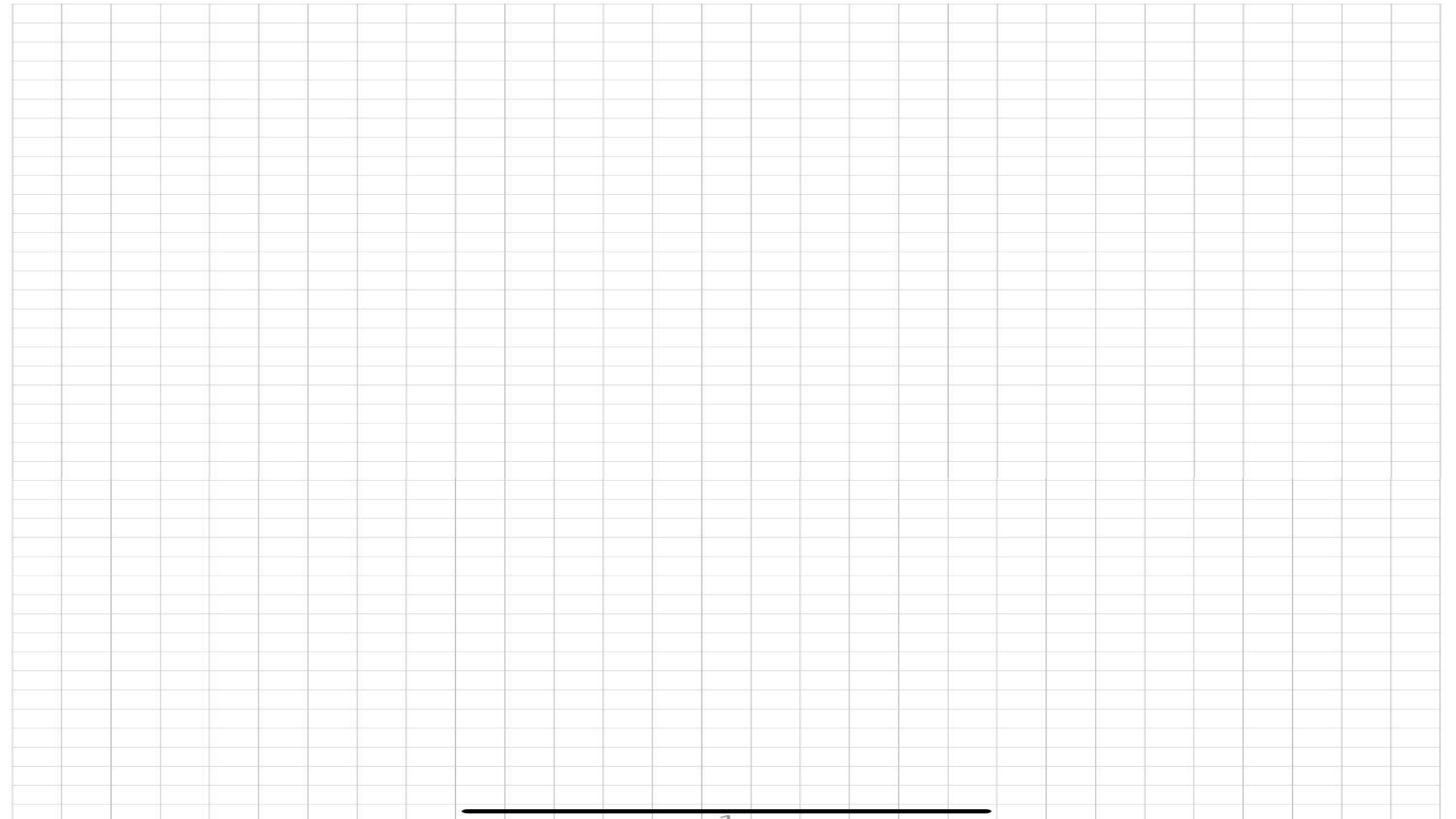
$b = 50$, genera $\mu = 2$ y obtengo

$$\bar{x} = 1,36 \quad y \quad s = 3,05$$

$$S(\underline{x}) = \left(\bar{x} - a \frac{s}{\sqrt{n}} ; \bar{x} - b \frac{s}{\sqrt{n}} \right)$$

$$= (-1,16 ; 2,65) \quad \text{en un i.c. para } \mu$$

$\mu = 2$ 



Regiones de confianza asintóticas

Def: Sea $\underline{X}_n = X_1, \dots, X_n$ una m.a de una población con distribución perteneciente a la flía. $F_\theta(x)$, con $\theta \in \Theta$. Se dice que $S_n(\underline{X}_n)$ es una sucesión de regiones de confianza de nivel asintótico $1 - \alpha$ si:

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(\theta \in S_n(\underline{X}_n)) = 1 - \alpha$$

Teorema: Sea \underline{X}_n una m.a. de una población con distribución $F_\theta(x)$, con $\theta \in \Theta$. Supongamos que para cada n se tiene $U_n = g(\underline{X}_n, \theta)$ que converge en distribución a U , donde U es una v.a. cuya distribución no depende de θ . Entonces si a y b son tales que $\mathbb{P}(a < U < b) = 1 - \alpha$ se tiene que $S_n(\underline{X}_n) = \{\theta : a < U_n < b\}$ es una región de confianza de nivel asintótico $1 - \alpha$ para θ .

Ejercicio 3

$$\frac{3}{5} = \frac{30}{50} = \frac{1}{50} \sum_{i=1}^{50} Y_i$$

Se arroja 50 veces una moneda con probabilidad p de salir cara.
Hallar un intervalo de confianza asintótico de nivel 0.95 para p
basado en la observación $x=30$. *= 30 de veces que salió cara.*

$$X \sim \text{Bin}(n=50, p)$$

$$X = \sum_{i=1}^{50} Y_i \text{ con } Y_i \stackrel{iid}{\sim} \text{Ber}(p)$$

$$\left[\frac{\bar{Y} - E\bar{Y}}{\sqrt{V(\bar{Y})}} = \frac{\bar{Y} - p}{\sqrt{p(1-p)}} \sqrt{n} \right] \xrightarrow[n \rightarrow \infty]{\text{TCL}} N(0, 1)$$

es un pivote asintótico.

$$\xrightarrow[n \rightarrow \infty]{\text{TCL}} \bar{X} \sim N(E\bar{X}, V(\bar{X}))$$

$$\xrightarrow[n \rightarrow \infty]{\text{TCL}} \bar{X} \sim N(0, 1)$$

$$P(a < U < b) = 0,95$$

$$y \quad U \sim N(0,1)$$

$$a = -b = \frac{-1}{\sqrt{U}} (0,025) = -1,96$$

$$P\left(\underbrace{-1,96}_{ii)} < \underbrace{\frac{\bar{y} - p}{\sqrt{p(1-p)}}}_{ii)} < \underbrace{1,96}_{ii)}\right) = 0,95$$

$$S(x) = \{p : a < U < b\} \\ = \{p : p \in (a, b)\}$$

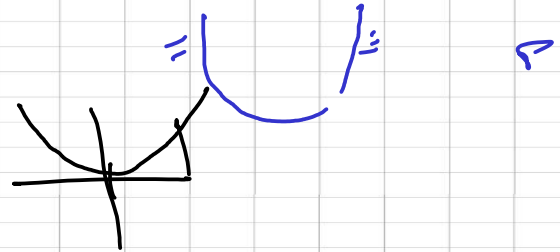
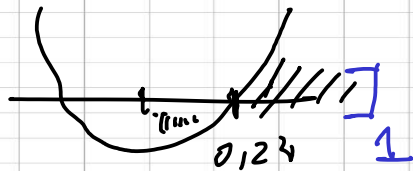
$$ii) \boxed{\bar{y} - p > 0}$$

$$\bar{y} - p < \frac{b}{\sqrt{n}} \sqrt{p(1-p)}$$

$$(\bar{y} - p)^2 < \frac{b^2}{n} p(1-p)$$

$$p^2 - 2\bar{y}p + \bar{y}^2 < \frac{b^2}{n} p - \frac{b^2}{n} p^2$$

$$0 < \left(1 - \frac{b^2}{n}\right)p^2 + \left(2\bar{y} + \frac{b^2}{n}\right)p - \bar{y}^2$$



$$u) \frac{\bar{y} - p < 0}{\sqrt{p(1-p)}} < 0$$

$$\frac{\bar{y} - p}{\sqrt{p(1-p)}} < 0$$

$$\frac{\frac{\partial}{\partial p} \sqrt{p(1-p)}}{\frac{\partial}{\partial p} (\bar{y} - p)} < 0$$



$$\frac{\partial}{\partial p} p(1-p) > (\bar{y} - p)^2$$

$$\frac{\partial^2}{\partial p^2} p - \frac{\partial^2}{\partial p^2} p^2 > \bar{y}^2 - 2p\bar{y} + p^2$$

$$\left(1 + \frac{\partial^2}{\partial p^2}\right) p^2 + \left(2\bar{y} + \frac{\partial^2}{\partial p^2}\right) p - \bar{y}^2 > 0$$



$$\frac{3}{5} = 0,6$$

$$S = (0,24; +\infty) \cap (0,4; 0,98) = (0,4; 0,98)$$

IC para la media de una población desconocida

En general, dada una m.a \underline{X}_n de una población desconocida, una buena forma de aproximarse a la media de dicha población es considerar el promedio de las muestras (\bar{X}_n).

Por TCL, sabemos que \bar{X}_n tiende en distribución a una v.a. normal. En particular,

? ¿queremos la IC para EX

$$\frac{\bar{X}_n - \mathbb{E}[X]}{\sqrt{\text{var}(X)/n}} \stackrel{(a)}{\sim} \mathcal{N}(0, 1)$$

Se puede probar que si se desconoce también la varianza de la población (que es lo más común) vale que

Resolven EJ 3 usando este:

$$\frac{\bar{X}_n - \mathbb{E}[X]}{S/\sqrt{n}} \stackrel{(a)}{\sim} \mathcal{N}(0, 1)$$

30

$$= \frac{1}{n-1} \sum_{i=1}^{30} (y_i - \bar{y})^2$$

$$S = \frac{1}{49} \left(30 \cdot \left(1 - \frac{3}{5} \right) - 20 \cdot \frac{3}{5} \right)$$

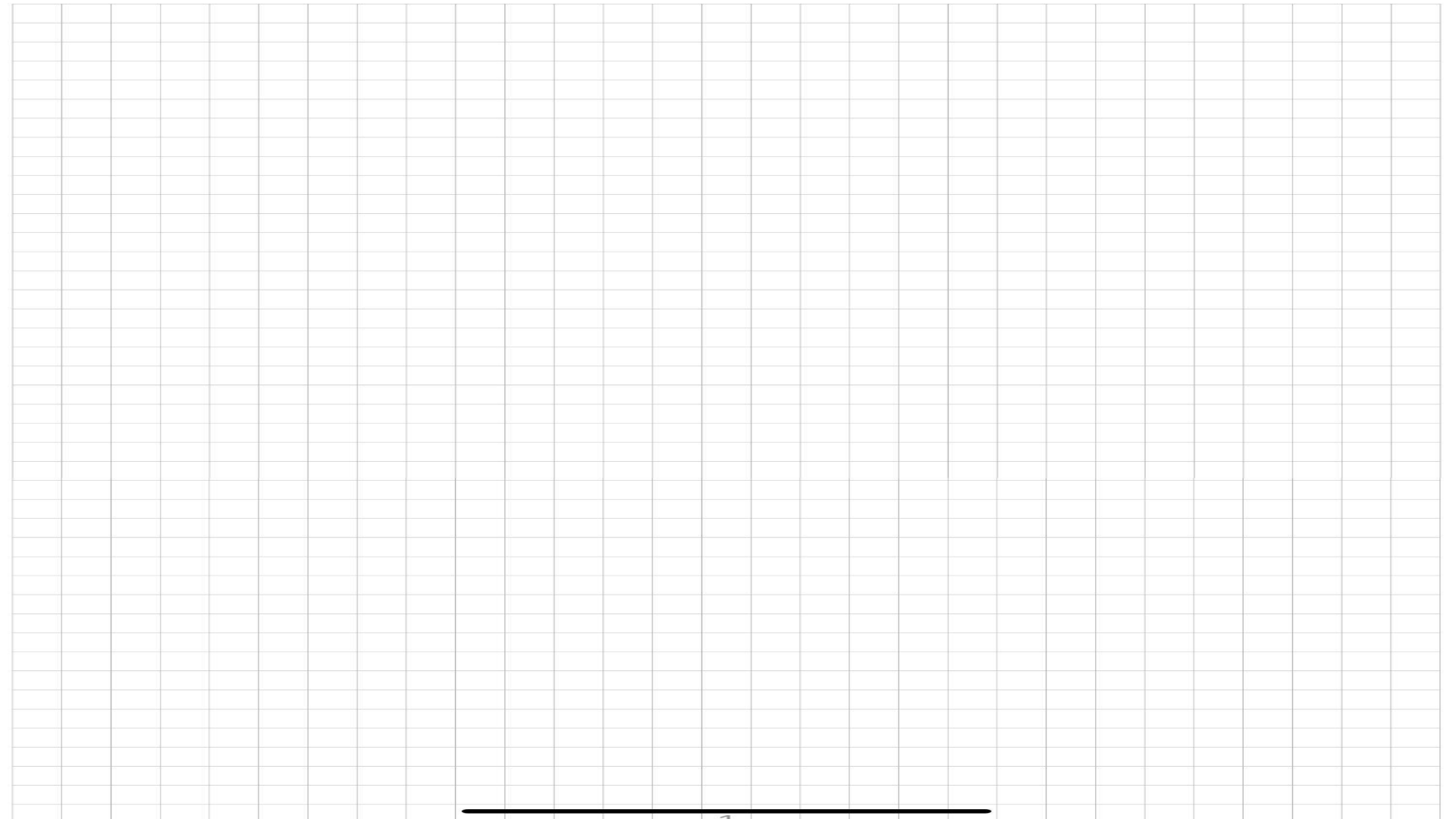
Ejercicio 4

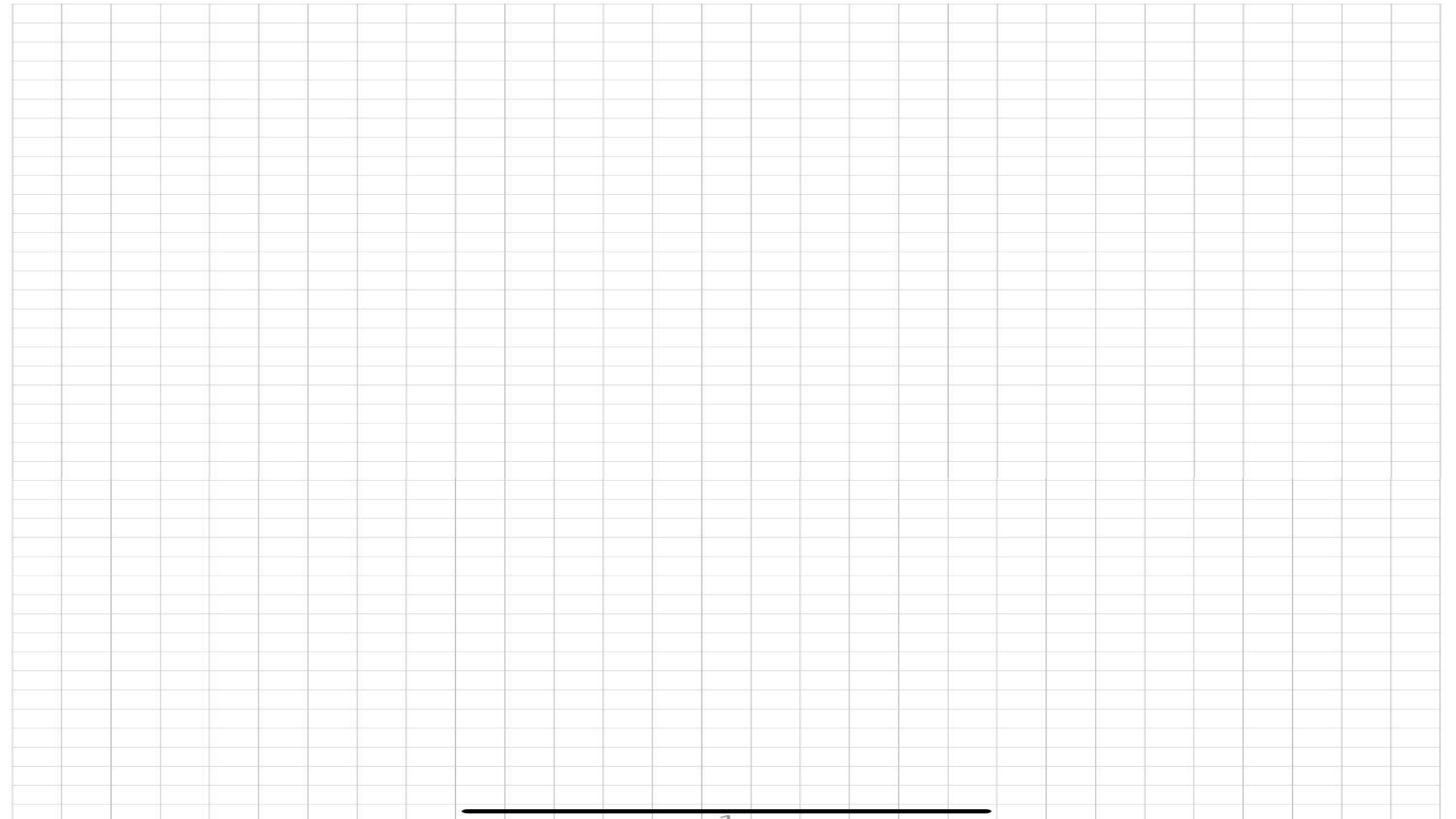
De un experimento en los efectos de un medicamento para la ansiedad se midió el puntaje en un test de memoria antes y después de tomar el medicamento. A partir de los datos que se encuentran en el archivo `Islander_data.csv` hallar un IC para la media del tiempo de respuesta después de consumir el medicamento.

X : "tiempo de respuesta"

$$U = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \stackrel{TCL}{\sim} N(0,1)$$

$$P\left(\underset{-1,96}{a} < U < \underset{1,96}{b}\right) = 0,95 \rightarrow$$





Intervalos de confianza aproximados (plug-in)

Fun de distrib. empírica



Intervalos basados en la Normal:

$$\underline{T(\hat{F}_n)} \approx N(T(F), \hat{se}^2)$$



Boots trap

Entonces, un intervalo aproximado de nivel $1 - \alpha$ está dado por

$$T(\hat{F}_n) \pm z_{\alpha/2} \hat{se}.$$

extensión lineal

Ejemplo: un intervalo para la media

está dado por $\bar{X}_n \pm z_{\alpha/2} \hat{se}.$ donde

$$\mu = T(F) = \int x dF(x) = \int x f(x) dx$$

$$se = \sqrt{V(\bar{X}_n)} = \sigma / \sqrt{n}$$

$$\hat{F}_n(1 - \frac{\alpha}{2})$$

$$\hat{se} = \frac{s}{\sqrt{n}}$$

Bootstrap: estimar la varianza

La idea es estimar $\mathbb{V}_{\hat{F}_n}(T_n)$ mediante una simulación:

$$\begin{array}{lcl}
 \text{Real world} & F & \Rightarrow X_1, \dots, X_n \Rightarrow T_n = g(X_1, \dots, X_n) \\
 \text{Bootstrap world} & \hat{F}_n & \Rightarrow X_1^*, \dots, X_n^* \Rightarrow T_n^* = g(X_1^*, \dots, X_n^*)
 \end{array}$$

$\underbrace{\quad}_{\text{muestras con reemplazo}}$
 Bootstrap Variance Estimation

1. Draw $X_1^*, \dots, X_n^* \sim \hat{F}_n$. (muestras de m. z.)
2. Compute $T_n^* = g(X_1^*, \dots, X_n^*)$.
3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \dots, T_{n,B}^*$.
4. Let

$$v_{\text{boot}} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2 \Rightarrow \hat{SE} = \sqrt{v_{\text{boot}}}$$

Bibliografía

- "Notas de Estadística", Graciela Boente y Víctor Yohai, FCEyN, UBA.
- "All of Statistic: A concise Course in Statistical Inference", Larry Wasserman