Inferencia Estadística

G3: Estimación por intervalos

Gabriel Martos
Email: gmartos@utdt.edu
Nicolás Ferrer
Email: nicolas.ferrer.747@gmail.com

Listado de ejercicios

1. La aerolínea Norwegian quiere determinar que porcentaje de sus clientes estarían dispuestos a pagar una tarifa plana de 5 dólares por acceso ilimitado a Internet durante los vuelos de cabotaje. De una muestra de 200 pasajeros elegidos al azar, 125 indicaron que estarían dispuestos a pagar dicha tarifa. Utilizando los datos de esta encuesta, obtenga el intervalo de confianza del 95% para estimar la proporción de clientes que estarían dispuestos a pagar por este servicio a bordo.

2. Un psicólogo quiere estimar la varianza de los puntajes de los exámenes de los empleados de una compañía. En una muestra de 18 puntajes se estimó una desviación estándar muestral de $s = 10.4$. Encuentre un intervalo de confianza del 90% para el parámetro σ^2. Indique los supuestos que necesita hacer para construir el intervalo. ¿Qué estrategia utilizaría para verificar si su supuesto es razonable?

3. La secretaría de seguridad vial de la provincia de Buenos Aires realizó un experimento para comparar los tiempos de reacción de los conductores a la luz roja y la luz verde (los colores de los semáforos). El experimento consistió en encender secuencialmente una luz roja y verde (en orden aleatorio), y a cada sujeto se le pidió que presionara un interruptor para apagar la luz inmediatamente después que esta se encendía. Los tiempos de reacción (medidos en segundos) de cada individuo en el experimento se encuentran en la siguiente tabla:

Subject	Red (x)	Green (y)	$d = x - y$
1	0.30	0.43	–0.13
2	0.23	0.32	–0.09
3	0.41	0.58	–0.17
4	0.53	0.46	0.07
5	0.24	0.27	–0.03
6	0.36	0.41	–0.05
7	0.38	0.38	0.00
8	0.51	0.61	–0.10

   (a) Enmarque el experimento en alguno de los contextos de intervalos para diferencia de medias discutidos en clase. Especifique cual es el parámetro de interés atendiendo al planteo del enunciado.

(b) Construya un intervalo de confianza con $\alpha = 0.05$ para el parámetro en cuestión e interprete los resultados.

(c) Teniendo en cuenta el tamaño de la muestra: ¿Bajo qué supuestos es válido el test propuesto? ¿Qué estrategias se le ocurren para justificar el supuesto?

4. Un empresa farmacológica necesita de tu expertise estadística para diseñar un ensayo clínico con el que cuantificar la efectividad de una droga en fase experimental que ayuda a regula el colesterol en pacientes con problemas cardiológicos. Se pretende calcular un intervalo de confianza para la reducción media de colesterol que se produce al complementar el tratamiento cardiológico con la droga en cuestión. Por experiencia pasada, testeando el mismo medicamento en pacientes con otras patologías similares, se sabe que la distribución del cambio en la cantidad de colesterol sigue una distribución normal; y que la varianza del cambio en la cantidad de colesterol es de 16mg/dL. Con esta información, la farmacéutica te pide que le indiques cuál es el tamaño de muestra mínimo con el que debería trabajar si quiere que su intervalo de confianza del 95% tenga una precisión de 5mg/dL.

5. Un grupo de sociólogos diseñó, con el objeto de medir la actitud de los economistas hacia las minorías, una encuesta con puntajes. Cuando el resultado global de la encuesta tiene puntajes elevados, entonces se evidencian actitudes negativas. Se tomaron dos muestras aleatorias independientes, una de $n_H = 151$ economistas hombres y otra de $n_M = 108$ economistas mujeres. Para el primer grupo el puntaje medio y el desvío estándar estimados fueron de $\bar{x}_H = 85.8$ y $s_H = 19.13$ respectivamente. En cambio para el segundo grupo fueron $\bar{x}_M = 71.5$ y $s_M = 18.83$. Construya un intervalo de confianza para la diferencia de medias identificando/justificando razonablemente que tipo de test de comparación de medias utiliza. Indique todos los supuestos que hace y cómo verificaría si se cumple los mismos en la práctica.

6. Se han recogido medidas de contaminación atmosférica en 10 lugares de la ciudad obteniéndose la siguiente muestra:

$$3.3; \quad 1.7; \quad 3.7; \quad 4.6; \quad 2.3; \quad 3.9; \quad 4.3; \quad 1.4; \quad 1.6; \quad 3.6$$

Hallar un intervalo de confianza al 95% para la varianza poblacional, mencionando las hipótesis estadísticas que es necesario asumir para validar el método de inferencia. ¿Es tu intervalo el de mayor precisión?

7. Sean L y U dos variables aleatorias que verifican que: $L \leq U$, $P(L \leq \theta) = 1 - \alpha_L$, $P(U \geq \theta) = 1 - \alpha_U$. Demostrar que: $P(L \leq \theta \leq U) = 1 - \alpha_L - \alpha_U$.

8. Considere una muestra aleatoria de tamaño $n = 1$ de los modelos de probabilidad:

$$f_1(x;\theta) = \begin{cases} 1 \text{ si } \theta - 1/2 < x < 1/2 + \theta, \\ 0 \text{ otro caso.} \end{cases} \quad \text{y } f_2(x;\theta) = \begin{cases} 2x/\theta^2 \text{ si } 0 < x < \theta, \text{ con } \theta > 0, \\ 0 \text{ otro caso.} \end{cases}$$

(a) Hallar las expresiones de los pivotes y sus distribuciones.

(b) Dar una expresión para los intervalos de confianza de nivel $1 - \alpha$.

9. Considere una muestra $\{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;p)$, donde:

$$f(x;p) = (1-p)^{x-1}p, x = 1, 2, 3, \ldots$$

(a) Hallar la expresión del pivote aproximado.

(b) Dar una expresión general para el intervalos de confianza de nivel $1-\alpha$.

(c) Dada una muestra de tamaño $n = 100$ de donde surge que $\bar{x} = 50$, computar el intervalo de confianza (de Wald) del 95%.

10. Sea $f(x)$ una densidad conocida, encuentre los pivotes y mencione como computaría los respectivos intervalos para:

(a) μ en el modelo de localización $f(x-\mu)$.

(b) σ en el modelo de escala $f(x/\sigma)/\sigma$.

11. Sea $X_1, \ldots, X_n$ una muestra iid de un modelo estadístico con parámetro $\theta > -1$

$$f(x;\theta) = \begin{cases} (\theta+1)x^\theta \text{ si } 0 \leq x \leq 1, \\ 0 \text{ en otro caso.} \end{cases}$$

(a) Computa el estimador máximo verosímil de θ.

(b) Determina un intervalo aleatorio de nivel de confianza aproximado $1-\alpha$.

(c) De una muestra de tamaño $n = 500$ se sabe que $\sum_{i=1}^{500} \log(x_i) = -450$, construye el intervalo de confianza con $\alpha = 0.05$. Recuerde que cuando $n \gg 0$:

$$\text{Var}(\hat{\theta}) \approx -\frac{1}{\frac{\partial^2}{\partial p^2}\ell(\theta|\mathbf{x})\big|_{\theta=\hat{\theta}}}.$$

12. Si $\mathbf{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} \Gamma(1,\theta)$, se puede demostrar entonces que el pivote $g(\mathbf{X},\theta) = (2/\theta)\sum_{i=1}^n X_i$ tiene una distribución $\chi^2_{(2n)}$. Sabiendo esto se pide:

(a) Construya la expresión general del intervalo de confianza de nivel $1-\alpha$ para el parámetro θ.

(b) Con los datos de la siguiente muestra (que provienen del modelo anterior):

1 58 4 67 5 95 21 124 22 124 28

160 40 202 42 260 51 303 53 363

compute el intervalo de confianza de nivel 95%.

13. **Intervalos predictivos**: Considere $\{X_1, \ldots, X_n, X_{n+1}\} \overset{iid}{\sim} N(\mu, \sigma^2)$ y llamemos $\bar{X} = \sum_{i=1}^n X_i/n$ y $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ (media y cuasi-varianzas muestrales con los primeros n datos de la muestra aleatoria). Se pide:

(a) ¿Cómo se distribuye la variable aleatoria $\bar{X} - X_{n+1}$?

(b) Hallar la constante c tal que el estadístico $c(\bar{X} - X_{n+1})/S \sim t_{n-1}$.

(c) Para $n = 8$, determine la constante k tal que:

$$P(\bar{X} - kS < X_9 < \bar{X} + kS) = 0.80,$$

el intervalo $(\bar{x} - ks, \bar{x} + ks)$ se conoce con el nombre de intervalo *predictivo* de X.

14. Sean $\{X_1, \ldots, X_9\} \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$ y $\{Y_1, \ldots, Y_{12}\} \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ dos muestras aleatorias independientes de poblaciones de las que se desconoce la varianza pero se sabe que $\sigma_X^2 = 3\sigma_Y^2$. Construya un pivote para el parámetro de interés $\Delta = \mu_X - \mu_Y$, determine su distribución y establezca la forma general del intervalo de confianza del 95%.

15. Nos interesa estudiar la homogeneidad de los rendimientos de los estudiantes utilizando datos del programa PISA. Trabajamos con dos poblaciones–digamos A = CABA y B = Resto del país– de alumnos evaluados y podemos asumir que la distribución de la variable de interés (las notas de la evaluación) en ambas poblaciones es normal: digamos $X_A \sim N(\mu_A, \sigma_A^2)$ y $X_B \sim N(\mu_B, \sigma_B^2)$. Se puede demostrar que el siguiente pivote:

$$g(S_A^2, S_B^2, \sigma_A^2, \sigma_B^2) = \frac{S_A^2/\sigma_A^2}{S_B^2/\sigma_B^2} \sim F_{n_B-1}^{n_A-1},$$

donde $n_A - 1$ son los grados de libertad en el numerador y $n_B - 1$ grados de libertad en el denominador de una F de Snedecor.

(a) Con la información anterior, construye un intervalo de confianza de nivel $1 - \alpha$ para el parámetro de interés $\tau \equiv \sigma_B^2/\sigma_A^2$.

(b) ¿Es tu intervalo del punto anterior único? ¿Es este intervalo el de máxima precisión?

(c) Con muestras de tamaño $n_A = 100$ y $n_B = 80$, se estimó que $s_A = 1.5$ y $s_B = 2.3$. Construya el intervalo de confianza relativo a $\alpha = 0.05$.

(d) ¿Cómo interpreta el intervalo estimado en términos del problema práctico planteado?

16. Simula $n = 20$ datos de (X_1, X_2); un vector aleatorio que sigue una distribución normal bi-variante de parámetros: $(\mu_1 = 1, \mu_2 = 2)$, $\sigma_1^2 = \sigma_2^2 = 1$ y la covarianza $\rho_{1,2} = 0.25$. Con los datos generados, computa la región de confianza para el vector de parámetros: (μ_1, μ_2) al nivel de confianza 95% y representarla gráficamente. Repite tu experimento pero ahora considerando las situaciones:

(a) Para $\alpha = 0.05$ y los parámetros del modelo fijos, considera muestreos y estimaciones con tamaños muestrales progresivamente mayores, por ejemplo: $n = 100, 500, 1000$. ¿Qué le ocurre a las regiones de confianza estimadas?

(b) Para $n = 100$ y los parámetros del modelo fijos, considera niveles de confianza progresivamente mayores, por ejemplo: $\alpha = 0.05, 0.01, 0.001$. ¿Qué le ocurre a las regiones de confianza estimadas?

(c) Para $n = 100$ y $\alpha = 0.05$ fijos, considera niveles de *ruido* en los datos mayores, por ejemplo: $\sigma_1^2 = \sigma_2^2 = 2, 5, 10$ y $\rho_{1,2} = 0.25\sigma_1^2$. ¿Qué le ocurre a las regiones de confianza estimadas?

17. Compute los intervalos asintóticos de Wald para los modelos: Exponencial, Poisson y Binomial. Compare estos intervalos con los que obtenemos con el método de la verosimilitud, es decir:

$$\text{IC}_{1-\alpha}(\theta) = \{\theta : \ell(\theta|\mathbf{x}) \geq \ell(\widehat{\theta}|\mathbf{x}) - \frac{1}{2}c_1(\alpha)\},$$

donde $c_1(\alpha)$ es el cuantil de una chi con 1 grado de libertad.

Método Pivotal

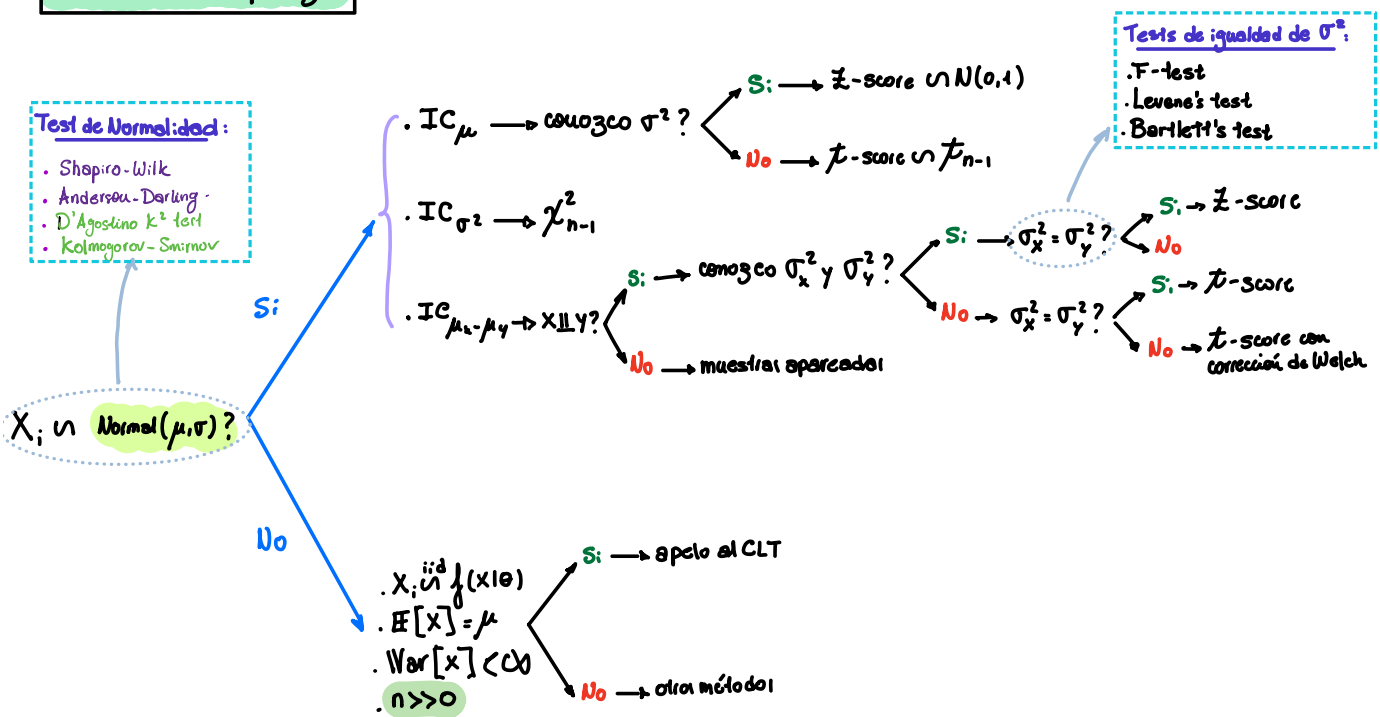
Sea $\underline{X} = \{X_1, X_2, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta)$, $g_n : \underline{X} \times \Theta \to \mathbb{R}^n$ es un **pivote** si y solo si:

- $\forall \theta \in \Theta$ fijo, $g_n(\underline{X}, \theta)$ es **continua** → no estadístico, depende de θ
- $P[g_n(\underline{X}, \theta) \leq c]$ no depende de θ $\Rightarrow$ F_{g_n} no depende de θ

↳ $IC_{1-\alpha}(\theta) = \{\theta \in \Theta \mid q_{1-\alpha/2} \leq g_n(\underline{X}, \theta) \leq q_{\alpha/2}\}$

Si F_{g_n} no es conocida con exactitud y solo de manera aproximada (asintótica), el IC será asintótico

Intervalos de confianza



---

1. La aerolínea Norwegian quiere determinar que porcentaje de sus clientes estarían dispuestos a pagar una tarifa plana de 5 dólares por acceso ilimitado a Internet durante los vuelos de cabotaje. De una muestra de 200 pasajeros elegidos al azar, 125 indicaron que estarían dispuestos a pagar dicha tarifa. Utilizando los datos de esta encuesta, obtenga el intervalo de confianza del 95% para estimar la proporción de clientes que estarían dispuestos a pagar por este servicio a bordo.

X_i : la persona i-ésima está dispuesto a pagar USD 5. $\Rightarrow X_i \sim$ Bernoulli(p)

$\underline{X} = \{X_1, X_2, \ldots, X_{200}\} \overset{iid}{\sim}$ Bernoulli(p)

Parámetro de interés: $g(\theta) = p \xrightarrow[MM]{MLE} \hat{p} = \bar{X} = \frac{1}{200}\sum_{i=1}^{200} X_i = \frac{125}{200} = \frac{5}{8} = 0.625$ → estimación puntual

. Por LLN, $\bar{x} \xrightarrow{p} \mathbb{E}[X] = p$. Además, X_i son iid, $\exists\, \mathbb{E}[X] = p$ y $\text{Var}[x] = p(1-p) < \infty$, puedo aplicar CLT:

$$\frac{\sqrt{n}\,(\bar{x} - \mathbb{E}[x])}{\sqrt{\text{Var}[x]}} \xrightarrow{D} N(0,1) \Rightarrow \underbrace{\frac{\sqrt{n}\,(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}}}_{g(T(\underline{x}),\theta)} \xrightarrow{D} N(0,1)$$

Construyo un IC asintótico con $\alpha = 0.05$:

$$P\left(z_{\frac{\alpha}{2}} \leq g(T(\underline{x}),p) \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-1.96 \leq \frac{\sqrt{n}\,(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}} \leq 1.96\right) = 0.95$$

$$P\left(-1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq \hat{p} - p \leq 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) = 0.95$$

$$P\left(-\hat{p} - 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq -p \leq -\hat{p} + 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) = 0.95$$

$$P\left(\hat{p} + 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \geq p \geq \hat{p} - 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) = 0.95$$

$$P\left(\hat{p} - 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p \leq \hat{p} + 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) = 0.95$$

depende de α

$$IC_{p,95\%} = [0.558, 0.692]$$

3. La secretaría de seguridad vial de la provincia de Buenos Aires realizó un experimento para comparar los tiempos de reacción de los conductores a la luz roja y la luz verde (los colores de los semáforos). El experimento consistió en encender secuencialmente una luz roja y verde (en orden aleatorio), y a cada sujeto se le pidió que presionara un interruptor para apagar la luz inmediatamente después que esta se encendía. Los tiempos de reacción (medidos en segundos) de cada individuo en el experimento se encuentran en la siguiente tabla:

$n=8$

Subject	Red (x)	Green (y)	$d = x - y$
1	0.30	0.43	–0.13
2	0.23	0.32	–0.09
3	0.41	0.58	–0.17
4	0.53	0.46	0.07
5	0.24	0.27	–0.03
6	0.36	0.41	–0.05
7	0.38	0.38	0.00
8	0.51	0.61	–0.10

(a) Enmarque el experimento en alguno de los contextos de intervalos para diferencia de medias discutidos en clase. Especifique cual es el parámetro de interés atendiendo al planteo del enunciado.

X_i = tiempo de respuesta a la luz roja del i-ésimo individuo

Y_i = tiempo de respuesta a la luz verde del i-ésimo individuo

$D_i = X_i - Y_i$

Parámetro de interés: $\mu_D = \mathbb{E}[D] = \mu_X - \mu_Y$

(1) No asumo normalidad:

- D_i son i.i.d
- $\mathbb{E}[D_i] = \mu_D$
- $\mathrm{Var}[D_i] < \infty$

} entonces apelo al CLT y construyo un pivote aproximado

$$g(T(\underline{X}), \theta) = \frac{\sqrt{n}(\bar{D} - \mu_D)}{\sqrt{\mathrm{Var}[D]}} \xrightarrow{D} N(0,1)$$

$$P\left(z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{D} - \mu_D)}{\sqrt{\mathrm{Var}[D]}} \leq z_{1-\alpha/2}\right) \overset{a}{=} 1 - \alpha = 0.95$$

$$P\left(\bar{D} - 1.96 \cdot \frac{S_D}{\sqrt{n}} \leq \mu_D \leq \bar{D} + 1.96 \cdot \frac{S_D}{\sqrt{n}}\right) \overset{a}{=} 0.95$$

$\bar{D} = -0.06$
$S_D = 0.07648$
$\Big\}$ $IC_{\mu_D, 95\%} = [-0.1129; -0.007]$ → validez asintótica

(2) Probar normalidad:

Debo verificar que $D_i \overset{iid}{\sim} N(\mu_D, \sigma_D^2)$ => aplicar test de normalidad

. Los 4 tests aplicados no permiten sostener la hipótesis de normalidad de los D_i => puedo construir un pivote exacto.

(b) Conozco $Var[D_i]$? No lo conozco => lo estimo S_D^2

$$g(T(\underline{X}), \theta) = \sqrt{n}\,\frac{(\bar{D}_n - \mu_D)}{S_D} \sim t_{n-1=7}$$

$$\Rightarrow P\left(t_{\alpha/2} \leq \sqrt{n}\,\frac{(\bar{D}_n - \mu_D)}{S_D} \leq t_{1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\bar{D}_n - t_{\alpha/2}\cdot\frac{S_D}{\sqrt{n}} \leq \mu_D \leq \bar{D}_n + t_{\alpha/2}\cdot\frac{S_D}{\sqrt{n}}\right) = 1-\alpha$$

$t_{\alpha/2} = 2.3646$

$IC_{\mu_D, 95\%} = [-0.12644, 0.00144]$