Inferencia Estadística

G1: Distribución en el muestreo y principios de reducción

Gabriel Martos
Email: gmartos@utdt.edu
Matías Pérez
Email: lic.matiasdperez@gmail.com

Listado de ejercicios teórico–prácticos

1. Ejercicio de investigación: Piensa o busca ejemplos (en particular en Economía) de uso de modelos estadísticos paramétricos y noparamétricos. ¿Qué ventajas tienen los modelos noparamétricos por sobre los paramétricos? ¿Se te ocurren ventajas en el sentido contrario?

2. El soporte de $X \sim f(x;\theta)$ es el conjunto definido como:

$$\text{Soporte}(f(x;\theta)) = \{x \in \mathbb{R} \,|\, f(x;\theta) > 0\}.$$

   (a) ¿El conjunto Soporte$(f(x;\theta))$ puede depender del parámetro θ si $f(x;\theta)$ pertenece a la familia exponencial? (Formaliza tu respuesta utilizando la definición de familia Exponencial vista en clase).

   (b) Demostrar que la familia exponencial se puede escribir, de manera equivalente a la expresión que dimos en clase, como sigue:

$$f(x;\theta) = \exp\Big(w(\theta)t(x) + m(x) + d(\theta)\Big).$$

   Esta manera de escribir las distribuciones en la familia exponencial resulta práctica cuando se discuten, por ejemplo, los modelos lineales generalizados.

3. Indicar si los siguientes modelos estadísticos pertenecen a la familia exponencial y en caso afirmativo determinar las expresiones analíticas de las funciones h, c, w, t:

   (a) Poisson: $f(x,\lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, con $\lambda \in (0,\infty)$ y $x \geq 0$.

   (b) Exponencial: $f(x,\lambda) = \dfrac{1}{\lambda}e^{-\frac{1}{\lambda}x}$, con $\lambda \in (0,\infty)$ y $x \geq 0$.

   (c) Truncada en θ: $f(x;\theta) = \frac{1}{\theta}e^{1-x/\theta}$ con $0 < \theta < x$.

   (d) Laplace: $f_X(x;\mu,\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|x-\mu|}{\sigma}\right)$ con $\mu \in \mathbb{R}$, $\sigma > 0$ y $x \in \mathbb{R}$.

   (e) Loc-escala Cauchy: $f(x;\mu,\sigma) = \frac{1}{\pi}\left[\frac{\sigma}{(x-\mu)^2+\sigma^2}\right]$ con $\mu \in \mathbb{R}$, $\sigma > 0$ y $x \in \mathbb{R}$.

   (f) Gamma: $f(x;\lambda,k) = \lambda e^{-\lambda x}\dfrac{(\lambda x)^{k-1}}{\Gamma(k)}$, con $\lambda > 0$, $k > 0$ y $x > 0$.

   (g) Beta: $f(x;\beta,\alpha) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$, con $\alpha > 0$, $\beta > 0$ y $0 \leq x \leq 1$.

   Nota: En (f) y (g) $\Gamma(\cdot)$ denota la función Gamma.

4. Sea $Z \sim f(z)$; definimos el cuantíl z_p como aquel valor que verifica que:

$$P(Z \leq z_p) = \int_{-\infty}^{z_p} f(z)dz = p.$$

Demostrar que si $X \sim \sigma^{-1} f((x-\mu)/\sigma)$ (es decir, X es una variable aleatoria con una distribución en la familia de localización y escala generada por f) entonces los cuantiles de X y Z están linealmente relacionados: $x_p = \sigma z_p + \mu$ para todo $p \in (0,1)$.

5. Considere el siguiente *modelo de regresión*:

$$Y = \beta_0 + h(X) + \sigma\varepsilon,$$

donde h es una función conocida y $\varepsilon \sim N(0,1)$. Identifique el modelo de localización y escala (determine la distribución y los parámetros) que sigue $Y|X$. ¿Cómo se relaciona éste modelo con el modelo lineal habitualmente utilizado en Econometría?

6. Sabiendo que la tasa de desempleo del último trimestre en Argentina fue del $\theta = 7.2\%$ y que pretendes encuestar a 1000 personas de la población económicamente activa del país, se plantean las siguientes cuestiones:

   (a) Llamemos X_i a la v.a. que representa la condición de empleo del encuestado i-ésimo en la muestra; define el modelo, el parámetro y el soporte de la v.a. X_i.

   (b) ¿Qué representa el estadístico $T = \sum_{i=1}^{1000} X_i/1000$?

   (c) ¿Cuántas personas desempleadas esperas encontrar en la muestra aleatoria?

   (d) ¿Cuál es la varianza y el desvío estándar de T?

   (e) ¿Cuál es la probabilidad de que exactamente 60 personas en la muestra estén desempleados? ¿Cómo cambia esta probabilidad cuando θ se incrementa?

   (f) Utiliza el Teorema del Límite Central (TLC) para aproximar la probabilidad de que por lo menos 40 personas en la muestra aleatoria estén desempleados.

7. Una consultora económica quiere estimar la distribución del ingreso familiar en CABA. Suponga que en la Ciudad de Buenos Aires viven 2 millones de familias.

   (a) Si X es la variable aleatoria ingresos familiares en la población, discuta que modelo estadístico (paramétrico) resulta razonable para X y determine su(s) parámetro(s).

   (b) ¿Cómo estimaría el(los) parámetro(s) que caracterizan la distribución del ingreso? Imagine que usted le debe presentar al directorio de la consultora una justificación de su elección del estadístico a utilizar, que argumentos se le ocurre plantear.

   (c) Suponé ahora que ya realizaste una encuesta a 100 familias elegidos aleatoriamente y que de los datos se observa que el ingreso promedio de las familias es de 35 mil pesos mensuales. Indica (con esta información) tus estimaciones de los cuartiles 1, 2 y 3 de la distribución del ingreso familiar.

8. Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} X$ con $V(X) = \sigma^2 < \infty$, demuestre las siguientes propiedades:

   (a) $E(\overline{X}_n) = \mu$.

   (b) $V(\overline{X}_n) = \sigma^2/n$.

   (c) $E(S_n^2) = \sigma^2$.

9. Sabiendo que $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\mu, \sigma^2)$, definimos las variables aleatorias Y_i como:

$$Y_i = \begin{cases} 1 \text{ si } X_i \geq \mu, \\ 0 \text{ en otro caso.} \end{cases} \qquad \text{para } i = 1, \ldots, n.$$

Sea $T_n = \sum_{i=1}^n Y_i$, se pide:

(a) Computa $E(T_n)$ y $V(T_n)$.

(b) Determinar la distribución del estadístico T_n (Ayuda: Determine primero como se distribuye Y_1 y tenga en cuenta que $\{Y_1, \ldots, Y_n\} \overset{iid}{\sim} Y_1$).

10. Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Unif}(0, 10)$ (i.e. X_1 es uniforme en el intervalo $[0, 10]$):

(a) Utiliza el TCL para aproximar las siguientes probabilidades: $P(4.5 \leq \overline{X}_n \leq 5.5)$ y $P(|\overline{X}_n - 5| > 1)$ en función del tamaño de la muestra.

(b) ¿Qué ocurre con las dos probabilidades anteriores cuando $n \to \infty$?

11. Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Unif}(0, 1)$, se pide calcular los siguientes momentos:

(a) $E(X_{(1)})$ y $V(X_{(1)})$, donde $X_{(1)} = \min\{X_1, \ldots, X_n\}$ es el mínimo en la muestra.

(b) $E(X_{(n)})$ y $V(X_{(n)})$, donde $X_{(n)} = \max\{X_1, \ldots, X_n\}$ es el máximo en la muestra.

(c) ¿Cómo cambian las distribuciones del mínimo y el máximo en la muestra aleatoria si se tiene que $X \sim \text{Exp}(\theta)$?

12. Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} X$ con $E(X) = \mu \neq 0$ y $V(X) < \infty$:

(a) ¿Cuál es la media y varianza (aproximadas) de la variable aleatoria $g(\overline{X}_n) = \overline{X}_n^2$?

(b) ¿Cómo se distribuye (aproximadamente) $\overline{X}_n^2$?

(c) ¿Cómo se distribuye (aproximadamente) $e^{\overline{X}_n}$?

(d) ¿Cómo se distribuye (aproximadamente) $\overline{X}_n^2$ si $\mu = 0$? (Ayuda: Debes utilizar una aproximación de segundo orden).

13. Sea $\{X_1, \ldots, X_n\}$ una muestra aleatoria de una población en donde la variable aleatoria de interés tiene una función de densidad de parámetro $\theta > 0$ dada por

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

(a) Comprobar que el estadístico $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ es suficiente para θ determinando de manera explícita las funciones $g(T(\mathbf{x}) = t; \theta)$ y $h(\mathbf{x})$ (Fisher–Neyman).

(b) Porque es redundante utilizar el teorema de factorización para probar que T es suficiente para θ? (Ayuda: ¿A qué familia pertenece $f(x; \theta)$ y quién es T?).

14. Sea $\{X_1, \ldots, X_n\}$ una muestra aleatoria de una población cuya variable aleatoria de interés tiene una densidad como sigue

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

para $\theta \in (-\infty, 0)$. Comprueba que el estadístico $T(X_1, \ldots, X_n) = \prod_{i=1}^n X_i$ es suficiente para θ y determine las expresiones de las funciones $g(T(\mathbf{x}) = t; \theta)$ y $h(\mathbf{x})$.

15. Considere el siguiente modelo estadístico

$$f(x;\theta) = \begin{cases} \left(\frac{\theta}{2}\right)^{|x|}(1-\theta)^{1-|x|} \text{ si } x \in \{-1,0,1\};\ \text{y } 0<\theta<1, \\ 0 \text{ en otro caso,} \end{cases}$$

(a) ¿Pertenece este modelo a la familia exponencial?

(b) Dada $\{X_1,\ldots,X_n\} \overset{iid}{\sim} f(x;\theta)$ encuentre un estadístico suficiente para θ.

(c) ¿Es completo el estadístico que encontraste en el punto anterior?

16. Sea $\{X_1,\ldots,X_n\} \overset{iid}{\sim} \text{Unif}(0,\theta)$, se pide:

(a) Verificar que $T(X_1,\ldots,X_n) = \max\{X_1,\ldots,X_n\}$ es suficiente para θ.

(b) Demuestre que $E(\frac{n+1}{n}T(X_1,\ldots,X_n)) = \theta$.

(c) ¿Es $\frac{n+1}{n}T(X_1,\ldots,X_n)$ un estadístico suficiente para θ?

17. Argumentar (utilizando resultados teóricos discutidos en clase) porque cuando se trata de una población normal los estadísticos $(\overline{X}_n, S_n^2)$ son independientes.

18. Consideremos una muestra de tamaño $n = 10$ de una población Poisson (ver 3.a):

(a) Construye la función de verosimilitud.

(b) Grafica (en una misma figura) $L(\lambda \,|\, \mathbf{x})$ cuando $\sum_{i=1}^{10} x_i \in \{10, 15, 20\}$. ¿Qué cambiaría en estas gráficas, si en vez de graficar $L(\lambda \,|\, \mathbf{x})$, graficas $\ell(\lambda \,|\, \mathbf{x}) \equiv \log L(\lambda \,|\, \mathbf{x})$.

(c) ¿Si $\sum_{i=1}^{10} x_i = 10$, es más verosímil que la muestra se corresponda con una población Poisson donde $\lambda = 2$ o con una población Poisson donde $\lambda = 1$?

(d) Repite los puntos (a), (b) y (c), pero asumiendo que la población es exponencial.

Inferencia Estadística

G2: Estimación puntual

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

Listados de ejercicios teórico–prácticos

1. Siendo $\{X_1, \ldots, X_n\}$ una muestra aleatoria de una población Uniforme discreta con soporte $\{0, 1, \ldots, \theta-1, \theta\}$, siendo θ un entero mayor a 1, se pide:

   (a) Hallar el estimador de momentos del parámetro θ.

   (b) De una muestra de tamaño $n = 10$ se tiene que $\sum_{i=1}^{10} x_i = 7$. Computa la estimación de momentos de θ con los datos de la muestra.

   (c) Si el soporte del modelo uniforme fuera en cambio: $\{-\theta, -\theta+1, \ldots, -1, 0, 1, \ldots, \theta-1, \theta\}$; ¿qué ocurre con el estimador que hallaste en el punto (a)? ¿Cómo redefines el estimador de momentos en este caso? Vuelve a computar la estimación de momentos en relación a la muestra dada en (b).

2. Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim}$ Binomila(θ, k) donde $(\theta; k)$ resultan desconocidos.

   (a) Hallar los estimadores de momentos de (θ, k).

   (b) Para una muestra de tamaño $n = 10$ se tiene que $\bar{x} = 1$ y $\sum_{i=1}^{10}(x_i - 1)^2 = 20$. Son las estimaciones de momentos de los dos parámetros coherentes con el modelo de probabilidad?

3. Sabiendo que $\{X_1, \ldots, X_n\} \overset{iid}{\sim}$ Gamma(α, λ), esto es:

$$f(x; \lambda, \gamma) = \begin{cases} \lambda \exp(-\lambda x) \frac{(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & \text{si }, x > 0, \\ 0 \text{ en otro caso.} \end{cases}$$

   con $\lambda > 0$ y $\alpha > 0$. Hallar los estimadores de momentos de $\boldsymbol{\theta} = (\alpha, \lambda)$. De una muestra de tamaño $n = 10$ se sabe que $\sum_{i=1}^{10} x_i = 50$ y que $\sum_{i=1}^{10} x_i^2 = 144$, computar las estimaciones de momentos de $\boldsymbol{\theta} = (\alpha, \lambda)$

4. Para los siguientes modelos de probabilidad:

   (a) Poisson: $f(x, \lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, con $\lambda \in (0, \infty)$ y $x \geq 0$.

   (b) Exponencial: $f(x, \lambda) = \dfrac{1}{\lambda} e^{-\frac{1}{\lambda}x}$, con $\lambda \in (0, \infty)$ y $x \geq 0$.

   (c) Truncada en λ: $f(x; \lambda) = \lambda^{-1} e^{1-x/\lambda}$ con $0 < \lambda < x$.

   (d) Gamma: $f(x; \lambda, k) = \lambda e^{-\lambda x} \dfrac{(\lambda x)^{k-1}}{\Gamma(k)}$, con $\lambda > 0$, $k = 2$ (conocido) y $x > 0$.

   Computa en cada caso el estimador de momentos y el estimador máximo verosímil de λ. Sabiendo que el valor del estadístico $\sum_{i=1}^{20} x_i = 25$, computa las estimaciones de momentos y máximo verosímiles relativas a λ en cada caso.

5. Demuestra que los estimadores de los puntos (a) y (b) son UMVUE y que sus ECM convergen a cero cuando el tamaño de la muestra tiende a infinito (consistencia).

6. Si $\{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta)$ donde $-1 \leq \theta \leq 1$ y

$$f(x;\theta) = \frac{1+\theta x}{2}, \text{ para } x \in [-1,1].$$

   (a) Obtener el estimador de momento del modelo.

   (b) Utiliza argumentos de la distribución en el muestreo para justificar una aproximación del riesgo cuadrático del estimador cuando $n \gg 0$ (cuando n es grande).

   (c) ¿Es el estimador de momentos consistente?

   (d) Para una muestra de tamaño $n = 4$ donde $X_1 = -0.5, X_2 = -0.1, X_3 = -0.2$, y $X_4 = 0.6$ compara el estimador de momentos contra el estimador máximo verosímil (tendrás que implementar algún método numérico para computar el segundo).

7. La información de la tabla representa una muestra realizada (de tamaño $n = 55$) de una población que sigue una distribución Poisson de parámetro λ. Con esta información se pide que halles la estimación máximo verosímil del parámetro $\psi_\lambda = P_\lambda(X = 2)$.

X:	0	1	2	3	4	5
Frecuencia	7	14	12	13	6	3

8. Considerando $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\mu, \sigma^2)$, se pide:

   (a) Hallar los estimadores máximo verosímiles de los parámetro μ y σ^2 (verifica que se cumplen las condiciones necesarias y suficientes para un máximo).

   (b) De una muestra se tiene que $\sum_{i=1}^{100} x_i = 170$ y $\sum_{i=1}^{100} x_i^2 = 810$, con esta información computa las estimaciones máximo verosímiles de μ y σ^2.

   (c) Computa el riesgo del estimador máximo verosímil de σ^2 y compara el riesgo de éste en relación al estimador insesgado S^2.

   (d) Computa la matriz de Información de Fisher y la cota CR.

   (e) Son los estimadores máximo verosímiles de μ y σ^2 eficientes.

   (f) Con la información del punto (b), construye una ellipse de confianza de nivel 0.95.

9. Sea $\{X_1, \ldots, X_n\}$ una muestra aleatoria de una población Uniforme (continua) con soporte en $[0, \theta]$, hallar el estimador máximo verosímil del parámetro θ.

10. Siendo W un estimador de θ, demuestre que $\text{ECM}(W, \theta) = \text{Sesgo}^2(W) + \text{Var}(W)$.

11. Si W es un estimador insesgado de θ, demuestre que: (1) $a + bW$ es insesgado para el parámetro $a + b\theta$ (con a y b dos constantes conocidas); y (2) W^2 es sesgado para θ^2.

12. Si $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bin}(p)$, considera los siguientes estimadores de p: $\widehat{p}_1 = \overline{X}$ (EMV) y $\widehat{p}_2 = (\sqrt{n/4} + n\overline{X})/(n + \sqrt{n})$ (estimador Bayesiano)

   (a) Computar el error cuadrático medio (ECM) de ambos estimadores.

   (b) Para n fijo, para que valores de p ocurre que $\text{ECM}(\widehat{p}_1, p) < \text{ECM}(\widehat{p}_2, p)$.

   (c) ¿Qué estimador prefieres para valores pequeños y cuál para valores grandes de n?

13. Si $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$, considera el estimador de σ^2: $\widehat{\sigma}^2_b = bS^2$ $(b > 0)$.

    (a) Computar el ECM de $\widehat{\sigma}^2_b$ (Utiliza las propiedades de S^2 en poblaciones normales).

    (b) Demuestre que para cualquier valor de σ^2, el estimador $\widehat{\sigma}^2_b$ minimiza el riesgo cuadrático cuando $b = (n-1)/(n+1)$.

14. Consideremos $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \mathrm{Cauchy}(\theta)$ donde θ es un parámetro de localización:

$$f(x; \theta) = \frac{1}{\pi}\left[\frac{1}{1 + (x-\theta)^2}\right].$$

    (a) ¿Puedes dar una solución analítica para el EMV de θ?

    (b) Determine las expresiones de las funciones $S(\theta)$ (score) y $H(\theta)$ (hessiano) relativas al método de Newton–Raphson discutidas en clase.

    (c) Experimento numérico: Considerando $\theta = 1$ (verdadero valor del parámetro), genera muestras de tamaños $n = 10, 100, 1000$ del modelo (`rcauchy(n,location = 1)`). Con estas muestras implementa el método Newton–Raphson para obtener una estimación numérica de θ. ¿Qué esperas que ocurra con tus estimaciones a medida que n crece?

15. Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \mathrm{Unif}(0, \theta)$ (modelo continuo), considera el estimador de momentos y $\frac{n+1}{n} \max(X_1, \ldots, X_n)$, como candidatos para estimar θ.

    (a) Computa el ECM de cada uno de los estimadores.

    (b) ¿Son estos estimadores consistentes?

    (c) Computa la eficiencia relativa entre los pares de estimadores.

    (d) ¿Con cuál de ellos te quedas al hacer inferencia para θ?.

16. Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \mathrm{Exp}(\theta)$, con $E(X_1) = \theta$; se proponen dos estimadores para θ: $\widehat{\theta} = \bar{X}$ (EMV) y $W = nX_{(1)}$ (donde recordemos $X_{(1)} = \min(X_1, \ldots, X_n) \sim \mathrm{Exp}(\theta/n)$).

    (a) Demuestra que $X_{(1)} \sim \mathrm{Exp}(\theta/n)$.

    (b) Computa el ECM de ambos estimadores para dirimir cuál de los dos prefieres.

    (c) ¿Son ambos estimadores consistentes?

17. Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \mathrm{Beta}(\theta, 1)$ con $\theta > 0$; esto es

$$f(x; \theta) = \theta x^{(\theta-1)}, \text{ para } x \in [0, 1].$$

    (a) Computar la información de Fisher del modelo.

    (b) Compute el estimador MV de θ: ¿Es insesgado? ¿Es eficiente?.

    (c) ¿Cómo construiría un intervalo de confianza aproximada (cuando $n \gg 0$) para θ con las cuentas del punto (a)?

    (d) Imagine que con los datos de una muestra de tamaño $n = 100$ (considere ésta una muestra grande), se tiene que $\widehat{\theta} = 7$. Indique los límites del intervalo de confianza aproximado (a un nivel $\alpha = 5\%$) para el parámetro θ.

18. Considere el siguiente modelo (lineal) $Y = \beta x + \varepsilon$, donde las x's se pueden considerar fijas (las elige quien observa los datos) y $\varepsilon \sim N(0, \sigma^2)$. Considere una muestra aleatoria $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$ de este modelo, en donde además los valores de x se eligieron de tal forma que $\sum_{i=1}^n x_i/n = 1 = 0$ y $\sum_{i=1}^n x_i^2/n = 1$. Con esta información se pide:

    (a) Identifique el modelo estadístico de Y y sus parámetros.

    (b) Hallar los estimadores máximo verosímiles de los parámetros anteriores.

    (c) Computa $I(\beta)$ y determina si el estimador $\widehat{\beta}$ es eficiente.

    (d) ¿Cómo se distribuye $\widehat{\beta}$?

    (e) Computa el ECM de $\widehat{\beta}$. ¿Es $\widehat{\beta}$ consistente?

19. Considere el siguiente modelo estadístico:

$$f(x;\theta) = \frac{1}{2}\theta^3 x^2 e^{-\theta x}, \text{ si } x, \theta > 0.$$

    (a) Computa el estimador máximo verosímil de θ.

    (b) Demuestre que cuando el tamaño de muestra es grande, la varianza del EMV es aproximadamente $\frac{\theta^2}{3n}$.

    (c) Construye un intervalo aleatorio $[L, U]$ tal que cuando la muestra es grande con una probabilidad de aproximadamente 0.95, θ pertenezca a dicho intervalo.

    (d) En una muestra de tamaño $n = 250$ se tiene que $\sum_{i=1}^{250} x_i = 432$. Obtenga el valor de la estimación máximo verosímil y determine los valores de los límites *estimados* del intervalo anterior.

20. Una empresa encuestadora quiere estimar la proporción de votantes θ en la Ciudad de Buenos Aires con intención de votar al candidato A en las próximas elecciones (suponga en todo momento que en la Ciudad de Buenos Aires hay 4 millones de votantes). Con este objetivo se tomará una muestra de $n = 1000$ votantes, preguntando *¿Votará ud a A en las próximas elecciones?*, registrando la preferencia de cada uno de los encuestados con las opciones *SI* y *NO* (no hay indecisos en esta población). La empresa necesita de su asesoramiento en lo respectivo a los siguientes puntos:

    (a) Usted propone estimar la proporción de votantes en favor de A utilizando el estadístico $\widehat{p}$ (proporción muestral) y su colega, Juan Perez, propone en cambio:

$$\widehat{p}_{JP} = \frac{n_{si} + 10}{n + 20},$$

    donde n_{si} es el número de encuestados que manifiesta intención de votar por el candidato A. ¿Es insesgado $\widehat{p}_{JP}$? Calcule su error cuadrático medio como función de p. Si el criterio para comparar estimadores es el error cuadrático medio, ¿es uno de los estimadores $\widehat{p}$ o $\widehat{p}_{JP}$ mejor que el otro cualquiera sea el valor de p en la población?

21. Se sabe que el tiempo T de respuesta de un servidor web dedicado a las apuestas online se sigue (ajusta a) una distribución Rayleigh de parámetro $\alpha > 0$, con función de densidad

$$f_T(t) = \begin{cases} \alpha t \exp(-\frac{\alpha}{2}t^2) \text{ si } t > 0; \\ 0 \text{ si } t \leq 0. \end{cases}$$

    (a) Hallar la expresión del estimador máximo verosímil del parámetro α en la población.

    (b) De una muestra de 50 tiempos de respuesta se obtuvo que $\sum_{i=1}^n t_i = 146.28$ y $\sum_{i=1}^n t_i^2 = 510.58$, cual es el valor de la estimación máximo verosímil de α?

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 $\{X_1, \ldots, X_9\} \overset{iid}{\sim} N(\mu_X, \sigma^2_X)$ y $\{Y_1, \ldots, Y_{12}\} \overset{iid}{\sim} N(\mu_Y, \sigma^2_Y)$ dos muestras aleatorias independientes de poblaciones de las que se desconoce la varianza pero se sabe que $\sigma^2_X = 3\sigma^2_Y$. 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.

Inferencia Estadística

G4: Contrastes de Hipótesis

Gabriel Martos
Nicolás Ferrer

Email: gmartos@utdt.edu
Email: nicolas.ferrer.747@gmail.com

Enunciados

1. ¿Verdadero o falso? Justifique cada una de sus respuestas convenientemente.

   (a) El nivel de significación de un test es igual a la probabilidad de que la hipótesis nula sea cierta.

   (b) Si un test con nivel de significación α rechaza la hipótesis nula, entonces la probabilidad de que la hipótesis nula sea cierta es igual a α.

   (c) La probabilidad de que un test rechace la hipótesis nula incorrectamente (es decir, cuando es cierta) es igual a la potencia del test.

   (d) Un error de tipo I ocurre cuando el test rechaza la hipótesis nula.

   (e) Si no logramos rechazar la hipótesis nula no podríamos aceptar la misma porque no estaríamos controlando la probabilidad de equivocarnos con esta decisión.

2. Considera $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\theta, \sigma^2 = 1)$. Queremos testear $H_0 : \theta \leq 0$ vs $H_1 : \theta > 0$ utilizando el estadístico $T(X_1, \ldots, X_n) = \bar{X}$ y la función de test:

$$\delta = \begin{cases} 1 \text{ si } T(X_1, \ldots, X_n) \geq Q, \\ 0 \text{ en otro caso,} \end{cases} .$$

   (a) Si $Q = 0$ calcula la probabilidad de cometer el error tipo I de este test.

   (b) Hallar probabilidad de cometer el error tipo II y su complemento (la potencia) (notar que estas cantidades dependerán de n, σ^2 y $\theta \in \Theta_1$)

   (c) ¿Cómo cambia la probabilidad de cometer el ET I y la potencia del test a medida que incrementamos Q?.

   (d) Determina el valor de Q (la región de rechazo) para que el test tenga un tamaño $\alpha = 0.05$. Para este test, imagina que tomas una muestra de tamaño $n = 10$ y observas que $\bar{x} = 0.5$, ¿cuál es el p-valor asociado a esta muestra particular? ¿Rechazamos H_0 a un nivel del 5%? ¿Cómo interpretas el p-valor?

3. Siendo $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Bern}(\theta)$, queremos testear $H_0 : \theta = 1/2$ vs $H_1 : \theta \neq 1/2$ utilizando el estadístico $T(X_1, \ldots, X_n) = \widehat{p}$ (proporción muestral) y la función de test:

$$\delta = \begin{cases} 1 \text{ si } |T(X_1, \ldots, X_n) - 1/2| \geq Q, \\ 0 \text{ en otro caso,} \end{cases} .$$

   (a) Determinar la región de rechazo para que, con n grande, la probabilidad de cometer el error tipo I del test sea de aproximadamente $\alpha = 0.05$.

(b) Hallar probabilidad (aproximada) de cometer el error tipo II y su complemento (la potencia). ¿Cómo dependen ambas cantidades respecto de n y p?

(c) Imagina que tomas una muestra de tamaño $n = 100$ y observas que $\widehat{p} = 0.54$, ¿cuál es el p-valor asociado a esta muestra particular? ¿Rechazamos H_0 a un nivel del 5%? ¿Cómo interpretas el p-valor?

4. Sea X una v.a. con función de densidad:

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

Donde $\theta \in \Theta = \{1, 2\}$. Considere el problema de testear $H_0 : \theta = 1$ vs. $H_1 : \theta = 2$ utilizando una muestra aleatoria $\{X_1, X_2\} \overset{iid}{\sim} f(x;\theta)$ y una región crítica para el test definida como: $R = \{(X_1, X_2) : X_1 X_2 \geq 3/4\}$. Hallar la probabilidad de cometer los errores tipo I y II del test (Ayuda: Notar que bajo H_0 $X \sim U(0, 1)$).

5. Sea $\{X_1, \ldots, X_8\} \overset{iid}{\sim} \text{Poiss}(\theta)$. Se quiere testear $H_0 : \theta \leq 0.5$ vs $H_1 : \theta > 0.5$. La zona de rechazo se define como $R = \{(X_1, \ldots, X_8) : T = \sum_{i=1}^{8} X_i \geq 8\}$ (Ayuda: $T \sim Pois(8\theta)$).

(a) Computar la probabilidad de cometer el error tipo I.

(b) Dejar indicada la función de potencia del test.

6. $\mathbf{X} \equiv \{X_1, \ldots, X_5\} \overset{iid}{\sim} \text{Bern}(\theta)$ y considerando $T(\mathbf{X}) = \sum_{i=1}^{5} X_i$ como el estadístico de contraste. Para el test:

$$\begin{cases} H_0 : \theta \leq \frac{1}{2}, \\ H_1 : \theta > \frac{1}{2}. \end{cases}$$

Compute la probabilidad de cometer el error tipo I y deje indicada la función de potencia del test.

7. Con $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Exp}(\lambda)$ (donde $E(X_1) = \lambda^{-1}$); considere testear: $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$. Asumiendo que $n \gg 0$, se pide que:

(a) Determine la región de rechazo asociado al LRT y al test asintótico de Wald.

(b) Dada una muestra de tamaño $n = 50$, de donde surge que $\sum_{i=1}^{50} = 150$; para $\lambda_0 = 2.5$ indique si rechazamos o no H_0 en cada uno de los casos anteriores.

8. En una encuesta a 200 directivos de empresas Argentinas sobre la ética de la gestión empresarial se les pidió responder utilizando una escala de 1 (muy en desacuerdo) a 7 (muy de acuerdo), sobre la opinión respecto de la siguiente afirmación: *Los directivos de las empresas Argentinas están interesados en la justicia social.* El promedio de los puntajes de las respuestas fue de 4,27, y el desvío estándar para la muestra fue de 1,32. Considerando que la misma encuesta, pero con empresarios Chilenos, arrojó que la media de las respuestas en ese país es de 4 puntos y queriendo argumentar que los empresarios Argentinos se preocupan más que los Chilenos por la justicia social, que test deberíamos plantear? Que conclusiones obtiene al llevar adelante el test plateado? (considere $\alpha = 0.05$ y asuma que independencia en las respuestas de los empresarios).

9. Un grupo de científicos del CONICET desarrolló un test de diagnóstico de coronavirus ultrarápido (se conocen los resultados 10 minutos después del hisopado). La ANMAT, organismo del estado que se encarga de validar la eficacia de estos test, lo contrata a ud para que diseñe un ensayo clínico con el que puedan validar la eficacia del método. Antes de comenzar a resolver el ejercicio se recomienda repase los conceptos epidemiológicos de Sensibilidad y Especificidad (los puedes encontrar en wikipedia).

(a) Ponga el problema en términos estadísticos indicando el/los modelo(s) de probabilidad involucrado(s) y sus parámetros.

- Ayuda: Un camino a seguir es pensar que la población objetivo (los argentinos) estamos divididos en dos grupos disjuntos: infectados y sanos. Si consideramos una muestra aleatoria de m pacientes contagiados $\{X_1, \ldots, X_m\} \sim \text{Bin}(\theta_X)$ y n sanos $\{Y_1, \ldots, Y_n\} \sim \text{Bin}(\theta_Y)$, los parámetros θ_X y θ_Y indicarían la sensibilidad y el complemento de la especificidad del test ($X = 1$ e $Y = 1$ indican test positivo sobre paciente contagiado y sano respectivamente) .

(b) ¿Cómo estimaría (puntualmente y haciendo estimaciones por intervalos) estas cantidades? ¿Que supuestos probabilísticos necesita hacer?

(c) Si la ANMAT se quiere asegurar que la sensibilidad del test es mayor a 990 por cada mil testeados enfermos; y una especificidad mayor al 950 por mil ¿Cómo plantearía esto en términos de un test estadístico?

(d) De un estudio sobre 500 pacientes enfermos y 300 sanos, se estimó puntualmente una sensibilidad del 0.995 y una especificidad del 0.998. Indique si le parece que este test cumple con los requerimientos del organismo.

10. Debido a las crecientes quiebras de grandes corporaciones, los auditores están cada vez más preocupados por los posibles casos de fraude y quieren evaluar si medir cuidadosamente los flujos de dinero en efectivo puede ayudar a descubrir potenciales casos. Para evaluar esta metodología, se tomaron los flujos de caja de una firma que había realizado fraude y se le pidió a 36 auditores que indicaran a partir del material la probabilidad de fraude utilizando una escala de 0 a 100. La evaluación media de estos 36 auditores fue de 66.21 y el desvío estándar de la muestra fue 22.93. Asimismo, se tomó una muestra de otros 36 auditores y se les solicitó que evaluaran la posibilidad de fraude de la firma, sin poder observar los flujos de caja. En este caso la media y el desvío estándar de la muestra fueron, respectivamente, 47.56 y 27.56. Utilice de manera conveniente alguno de los métodos de test vistos en clase para responder a la pregunta arriba planteada.

11. Una firma de energía eólica asegura que su nuevo molino de viento puede generar un promedio de 800 kilovatios de potencia por día (asuma que esto no depende de las condiciones meteorológicas ni externas al molino). Sabiendo que se puede suponer que la generación diaria de energía del molino de viento tiene distribución normal.

(a) Si ud fuera accionista de la compañía que produce los molinos; ¿con qué test apoyaría la afirmación de la firma? De una muestra de 100 días de generación en un molino, se estimó el desvío estándar en 120 kilovatios; cual es el valor más bajo de generación promedio que permite confirmar la afirmación de la firma con un nivel de confianza (aproximado) del 95%.

(b) Utilizando el valor crítico que obtuviste en el punto anterior, ¿cuál es la probabilidad de cometer el Error Tipo II si la media poblacional fuera de 820Kv/día?

(c) Si por el contrario, ud fuera un veedor externo contratado por el estado nacional (que va a comprar cientos de estos molinos en una licitación); ¿con qué test apoyaría su dictamen? De una muestra de 100 días de generación en un molino, se estimó el desvío estándar en 120 kilovatios; cual es el valor más alto de generación promedio que permite confirmar su afirmación al estado con un nivel de confianza (aproximado) del 95%.

(d) Utilizando el valor crítico que obtuviste en el punto anterior, ¿cuál es la probabilidad de cometer el Error de Tipo II si la media poblacional fuera de 770Kv/día?

12. Una PyME local produce cables para una compañía telefónica multinacional que opera en el país. Cuando su proceso de producción está bajo control (funcionando correctamente), el diámetro de los cables fabricados sigue una distribución normal con media 1.6 centímetros y una desviación estándar de 0.05 centímetros. Dada una muestra de 16 piezas de cable, se estimó una media de 1.615 centímetros de diámetro y un desvío estándar de 0.086 centímetros.

    (a) Asumiendo que el desvío estándar de la población es de 0.05 centímetros, pruebe, con un nivel de 10%, la hipótesis nula de que la media poblacional es de 1.6 centímetros. Encuentre también el menor nivel de significación al que esta hipótesis nula puede ser rechazada a favor de la hipótesis alternativa bilateral.

    (b) Pruebe, a un nivel de 10%, la hipótesis nula de que el desvío estándar poblacional es menor o igual a 0.05 centímetros.

13. Un productor de vino afirma que la proporción de sus clientes que no pueden distinguir su producto de un jugo de uva congelado es, como máximo, 0,09. El productor decide poner a prueba esta hipótesis nula frente a la alternativa de que la verdadera proporción es de más de 0,09. La regla de decisión elegida es rechazar la hipótesis nula si la proporción de la muestra de las personas que no pueden distinguir entre estos dos sabores excede 0,14.

    (a) Si se elige una muestra aleatoria de 100 clientes y se usa esta regla de decisión ¿Cuál es la probabilidad de realizar un Error de Tipo I?

    (b) Si se selecciona una muestra de 400 clientes y se usa la misma regla de decisión ¿Cuál es la probabilidad de realizar un Error de Tipo I? Explicar, con palabras y gráficamente, ¿Por qué su respuesta difiere de la respuesta de la parte a)?

    (c) Supongamos que la verdadera proporción de clientes que no pueden distinguir entre estos sabores es 0.20. Si se selecciona una muestra aleatoria de 100 clientes, ¿cuál es la probabilidad de realizar un Error de Tipo II?

    (d) Supongamos que, en lugar de utilizar la regla de decisión anterior, se decide rechazar la hipótesis nula si en la muestra la proporción de clientes que no pueden distinguir entre los dos sabores excede 0,16. Dada una muestra de 100 clientes:

        (i) Sin hacer cálculos, ¿la probabilidad de realizar un Error de Tipo I será mayor, menor, o igual que la misma probabilidad calculada en la parte a)?.

        (ii) Si la proporción real es 0.20, ¿la probabilidad de realizar un Error Tipo II ser mayor, menor, o igual que la misma probabilidad calculada en la parte c)?

14. Una empresa intenta comercializar paquetes con una mezcla de ceniza pulverizada de combustible y cemento. Esta mezcla, que se usa en la construcción, para poder ser comercializada debe ser sometida primero a un proceso de evaluación realizado por el ente regulador estatal para constatar que cumple con ciertas condiciones que aseguran que el material es confiable. En particular, el ente regulador requiere que la resistencia a la compresión de la mezcla comercializada en cada paquete sea mayor que 1300 KN/m. En el proceso de evaluación, el ente regulador toma una muestra aleatoria de 20 paquetes de la mezcla y mide la resistencia a la compresión de cada paquete. Teniendo en cuenta que las mediciones de la mezcla no son perfectas y que en el proceso de fabricación de la mezcla existen pequeñas variaciones en el mezclado, el ente regulador plantea un modelo que establece que la resistencia a la compresión para mediciones de paquetes elegidos al azar está distribuida normalmente. A su vez, dadas las características físicas del proceso de fabricación de la mezcla y del proceso de medición de la resistencia a la compresión, se sabe fehacientemente que el desvío estándar de las mediciones de resistencia a la

compresión de paquetes elegidos al azar es $\sigma = 60$ KN/m. Denotemos con μ la media de la verdadera resistencia de la mezcla en la población de paquetes fabricados por una empresa que presenta su producto para ser evaluado por el regulador.

(a) Para decidir si un producto cumple con las especificaciones el ente regulador plantea como hipótesis nula $H_0 : \mu \leq 1300$ y como hipótesis alternativa $H_1 : \mu > 1300$. ¿Considera usted que es correcta la elección de hipótesis nula y alternativa? Comente sobre las consecuencias de cometer errores de tipo I y II.

(b) Denote con $\overline{X}$ al promedio de la resistencia a la compresión de la muestra de 20 paquetes seleccionados al azar. Suponga que el ente regulador permite la comercialización de la mezcla evaluada sólo si $\overline{X} > 1331.26$. De acuerdo a las hipótesis nula y alternativa planteadas en el inciso anterior, ¿cuál es el nivel del test de esta regla de decisión?

(c) Suponga que una empresa que presenta su nuevo producto para su evaluación en realidad cumple con los requerimientos porque para su mezcla, $\mu = 1350$ (esto, por supuesto, es desconocido para el ente regulador). ¿Cuál es la probabilidad de que, con la regla usada por el ente regulador, la empresa pueda obtener el permiso para comercializar su producto? Si debiera explicar el sentido de esta probabilidad en una reunión de empresarios de la construcción que no saben nada sobre Estadística, ¿cómo lo explicaría con palabras sencillas?

(d) Dadas las hipótesis planteadas en el inciso (a), indique si la probabilidad calculada en el inciso (c) coincide con

- La probabilidad de error de tipo I del test.
- El nivel del test.
- La potencia del test en el valor $\mu = 1350$.
- La probabilidad de error de tipo II del test en $\mu = 1350$.

(e) Suponga que habiendo quedado insatisfecha porque el cálculo de la probabilidad del inciso (c) arrojó un valor muy bajo, la empresa presenta una queja formal ante el ente regulador solicitándole que reconsidere el procedimiento de evaluación para que el 95% de las mezclas que tengan un $\mu = 1350$ pasen satisfactoriamente la prueba del ente regulador. Suponga que el ente regulador desea mantener el nivel del test arrojado por el cálculo del inciso (b). ¿Qué cambios en el proceso de evaluación debe hacer el ente regulador para satisfacer del pedido de la empresa?

15. Durante la discusión parlamentaria por la promulgación de la ILE (2018), surgió la controversia de que la proporción de mujeres en favor de esta ley era menor a la de los hombres. Indique el modelo estadístico que utilizaría para reflejar las opiniones de hombres y mujeres al respecto (establezca los supuestos que considere necesarios); y justifique como plantearía un test para dirimir la cuestión de manera estadística. En una encuesta llevada a cabo por la consultora Synopsis (ver nota de prensa), se cuantifica este argumento utilizando la estimación puntual de las diferencias de proporciones:

> *... entre los 1485 casos el apoyo es levemente mayor entre los varones que entre las mujeres: el 55,6% por ciento de los hombres avaló la sanción del proyecto (el 28,3% estuvo en contra), mientras que hizo lo mismo el 51,6% de las mujeres (el 36,7% estuvo en contra).*

Utilice el test apropiado para dirimir si estas diferencias son significativas a un nivel del 5%, mencionando los supuestos que hace para poder correr el test (en la nota de prensa no se menciona como se compone la encuesta).

Inferencia Estadística

G5: Inferencia Bayesiana

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

Enunciados

1. Si $X|\theta \sim \text{Poiss}(\theta)$, y además podemos asumir que $P(\theta = 2) = 1/3$ y $P(\theta = 3) = 2/3$ (esto es $\Theta = \{2, 3\}$). Dada la información $x_1 = 2$ y $x_2 = 4$, compute la probabilidad a–posteriori[1] para θ. ¿Qué inferencia puede hacer respecto del parámetro θ?

2. Una de las ventajas del enfoque Bayesiano reside en que el teorema de Bayes se puede utilizar de forma secuencial; y esto es particularmente útil cuando necesitamos 'refrescar el modelo' con información nueva. Imagina que de manera secuencial recibes la información (datos) $\mathbf{x}_1$ y luego $\mathbf{x}_2$, argumenta porque es cierto que:

$$\pi(\theta|\mathbf{x}_1, \mathbf{x}_2) \propto L(\theta|\mathbf{x}_2)\pi(\theta|\mathbf{x}_1).$$

3. Si $X|\theta \sim \text{Poiss}(\theta)$ y $\theta \sim \text{Gamma}(\alpha, \beta)$[2], esto es:

$$\pi(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad \alpha > 0,\ \beta > 0,\ \theta > 0,$$

   donde α y β son dos hiperparámetros conocidos (elegidos por quien modela el problema).

   (a) Comprobar que la distribución aposteriori de θ es Gamma de parámetros: $\alpha_n = \sum_{i=1}^{n} x_i + \alpha$ y $\beta_n = \beta/(n\beta + 1)$.

   (b) ¿A dónde convergen la media y la varianza a posteriori cuando $n \to \infty$?

   (c) Imagine que la variable aleatoria X_i da cuenta de la cantidad de delitos registrados en la ciudad en el día i; y que de una muestra de 10 días se tiene que $\sum_{i=1}^{10} x_i = 140$. Justificando su elección de los parámetros α y β (encuentre algún argumento razonable para elegirlos), obtenga la distribución a posteriori de θ.

   (d) Reporte la media y varianza a posteriori.

   (e) Construya un HPD al 95% y 99% e interprete los resultados.

4. El modelo Beta–Bernoulli, asume una prior Beta para el parámetro θ. Obtener la distribución a posteriori si en vez de una prior Beta utilizáramos una distribución uniforme en el intervalo $(0, 1)$ para θ:

   (a) ¿Cómo interpretas el uso de una prior uniforme en términos prácticos?

   (b) Calcula en este contexto $E(\theta \,|\, \mathbf{x})$ y $V(\theta \,|\, \mathbf{x})$.

[1] Recuerda que por probabilidad total: $P(X_1 = 2, X_2 = 4) = \sum_{\theta \in \Theta} P(X_1 = 2, X_2 = 4|\Theta = \theta)P(\Theta = \theta)$.
[2] Bajo esta parametrización la media a priori de θ es $\alpha\beta$ y la varianza $\alpha\beta^2$

5. En clase discutimos el modelo Normal–Normal. Utiliza la fórmula de Bayes y las propiedades de los modelos conjugados para construír de forma detallada la distribución a posteriori $\pi(\theta \,|\, \mathbf{x}) = N(\mu_n, \sigma_n^2)$, que recordemos tiene parámetros:

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_0^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}} \text{ y } \sigma_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}.$$

   (a) ¿A dónde convergen los parámetros de la posteriori cuando $n \to \infty$?

   (b) ¿Cómo interpretas este resultado?

   (c) Determina la estructura que tendría un intervalo de confianza creíble a posteriori de probabilidad 0.95. ¿Es tu intervalo el HPD?

6. Imagina que trabajas para la consultora económica XYZ y se te encarga hacer inferencia bayesiana para el parámetro $\theta =$ "tasa de desempleo en CABA". Tomas una muestra de tamaño $n = 100$ de la población relevante y observas que la variable $y =$ Número de desempleados en la muestra $= 18$. Se pide respondas a lo siguiente:

   (a) ¿Cómo propondrías elegir la prior sobre θ?

   (b) Computa la distribución a–posteriori (para tu elección de prior).

   (c) Computa la esperanza, moda y varianza aposteriori de θ.

   (d) Computa la HPD para $\alpha = 5\%$.

   (e) Un economista amigo, con una visión diametralmente opuesta a la tuya en cuanto a la situación económica actual, presenta estimaciones diferentes utilizando los mismos datos de la encuesta anterior. ¿Cómo es esto posible?

   (f) ¿Qué crees que ocurriría con la "distancia" entre tus conclusiones y la de tu amigo economista si el tamaño de la muestra fuera 10 veces más grande?