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.

$$\underbrace{\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)}_{c(\theta)}\underbrace{\left[x^{-1}(1-x)^{-1}\right]}_{k(x)} x^{\alpha}(1-x)^{\beta} = c(\theta)k(x)\exp\{\alpha\ln(x)+\beta\ln(1-x)\}$$

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

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

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

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

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

(c) ¿Es $\frac{n+1}{n}T(X_1,\dots,X_n)$ 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.

Recap:

m.a: $\underline{X} = \{X_1, X_2, \dots, X_n\}$

· Ingredientes de la Inferencia Estadística
- una muestra: $\{x_1, x_2, \dots, x_n\}$
- un modelo estadístico:

· 2 supuestos:

* La m.a. es iid de un modelo estadístico con función de distribución (acumulada) F desconocida ($F = F(X;\theta)$)
* $F \in \mathcal{F} = \{F(x;\theta) \mid \theta \in \Theta\}$

Terminología:

· Parametro $\longrightarrow$ θ es el valor que indexa $\mathcal{F} = \{F(\cdot|\theta) / \theta \in \Theta\}$

θ^*

· Parámetro de interés $\longrightarrow$ aspecto de la distribución de $\underline{X}$ s/ el que quiero hacer inferencia. Suele ser una función $g(\theta)$.

· Estimador de $g(\theta)$: es una función de $\underline{X}$ / $\hat{g} = \delta(\underline{X})$ $\longrightarrow$ variable aleatoria

· Estimación: estimador evaluado en los datos $\longrightarrow$ valor fijo

· Estadístico: cualquier función de los datos $T(\underline{X})$

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

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

(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:

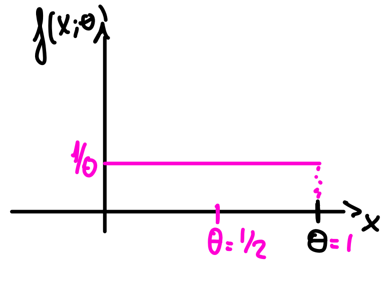
$$f(x;\theta) = \exp\Big(w(\theta)t(x) + m(x) + d(\theta)\Big).$$

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

(a)

$$f(x|\theta) = c(\theta)\,k(x).\exp(w(\theta)^T t(x))$$

$X \sim \mathcal{U}[0,\theta]$



Por definición, el soporte de $f(x;\theta)$ no puede depender de los parámetros θ $\Rightarrow$ todas las distribuciones en un modelo en particular $\mathcal{F}$ tienen el mismo soporte independiente de cual sea la configuración de los parámetros.

(b) $f(x|\theta) = h(x)\, c(\theta) \exp\left\{\sum_{i=1}^{n} w_i(\theta)\, t_i(x)\right\}, \quad x \in \mathbb{R}^n$

$$= \exp[\ln(h(x))] \exp[\ln(c(\theta))] \exp\left\{\sum_{i=1}^{n} w_i(\theta)\, t_i(x)\right\}$$

$$= \exp\left\{\sum_{i=1}^{n} w_i(\theta)\, t_i(x) + \underbrace{\ln[h(x)]}_{m(x)} + \underbrace{\ln[c(\theta)]}_{d(\theta)}\right\}$$

$$= \exp\left\{\sum_{i=1}^{n} w_i(\theta)\, t_i(x) + m(x) + d(\theta)\right\}$$

$$= \exp\left\{\sum_{i=1}^{n} w_i(\theta)\, t_i(x) + \ln[h(x)] - \ln\left[\frac{1}{c(\theta)}\right]\right\}$$

$$= \exp\left\{\sum_{i=1}^{n} w_i(\theta)\, t_i(x) + m(x) - d(\theta)\right\}$$

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 $x \in \mathbb{R}$.

(a) Poisson: $f(x;\lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}$

- $\text{supp}\{f(x;\lambda)\} = \{x \in \mathbb{Z}_0^+ \mid f(x;\theta) > 0\}$ ✓
- $h(x) = \dfrac{1}{x!} > 0, \forall x \in \mathbb{Z}_0^+ \longrightarrow$ no depende λ
- $c(\theta) = e^{-\lambda} \geq 0, \forall \lambda \in \Theta = \mathbb{R}^+$
- $\lambda^x = \exp\{\ln[\lambda^x]\} = \exp\{\underbrace{x}_{t(x)} \cdot \underbrace{\ln[\lambda]}_{w(\theta)}\}$

$f(x;\lambda) = \left(\dfrac{1}{x!}\right) \cdot e^{-\lambda} e^{\{\ln[\lambda] \cdot \underbrace{x}_{\text{estadístico suficiente}}\}} \in FE$

MLE: $\hat{\lambda}$ (tarea)

(b) $f(x;\theta) = \frac{1}{\theta} \underbrace{e^{-\frac{1}{\theta}x}}_{\text{kernel}}$, $\theta \in (0,+\infty)$ y $x \geq 0$

$\frac{1}{\sqrt{2\pi\sigma^2}}$ exp...

- $\operatorname{supp}\{f(x;\theta)\} = \mathbb{R}^+$ y no depende de θ ✓

- $f(x;\theta) = \mathbb{1}(x \geq 0) \cdot \frac{1}{\theta} e^{-\frac{1}{\theta}x}$

  - $h(x) = \mathbb{1}(x \geq 0)$ → no depende de θ y ≥ 0
  - $\boxed{c(\theta) = \frac{1}{\theta}}$ → no depende de x y es ≥ 0
  - $w(\theta) = -1/\theta$
  - $t(x) = x$

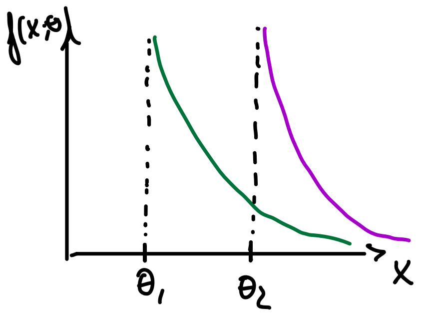
$$\int_0^{+\infty} e^{-\frac{1}{\theta}x}\, dx = -\theta e^{-\frac{1}{\theta}x} \Big|_0^{+\infty} = 0 - (-\theta) = \theta$$

$$\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} dx = \sqrt{2\pi\sigma^2}$$

$$\int_0^{+\infty} \frac{1}{\theta} e^{-\frac{1}{\theta}x}\, dx = 1$$

(c) Truncada: $f(x;\theta) = \frac{1}{\theta}\exp\{1 - x/\theta\}$ con $0 < \theta < x$

$\operatorname{supp}\{f(x;\theta)\} = \{x \in \mathbb{R}^+ \text{ y } x > \theta\}$ => el soporte depende de θ => $\notin$ FE



e) Log-scaled Cauchy: $f(x;\mu,\sigma) = \frac{1}{\pi}\left[\frac{\sigma}{(x-\mu)^2 + \sigma^2}\right]$

- $\operatorname{Supp}\{f(x;\mu,\sigma)\} = \mathbb{R}$ => no depende de $\theta = \begin{bmatrix}\mu \\ \sigma\end{bmatrix}$

- $h(x) = \mathbb{1}\{x \in \mathbb{R}\} = 1$
- $c(\theta) = \frac{\sigma}{\pi} \cdot \mu^0$
- $\exp\left\{\ln\left[\frac{1}{(x-\mu)^2+\sigma^2}\right]\right\} = \exp\left\{-\ln\left[(x-\mu)^2+\sigma^2\right]\right\}$

$$= \exp\left\{-\ln\left[x^2+\mu^2-2x\mu+\sigma^2\right]\right\}$$

no puede descomponerse como $w(\theta)t(x)$

$\therefore f(x;\mu,\sigma) \notin$ FE

$$\begin{bmatrix} t_1(\underline{X}) & t_2(\underline{X}) \end{bmatrix} \begin{bmatrix} w_1(\theta) \\ w_2(\theta) \end{bmatrix} = t_1(\underline{X})\, w_1(\theta) + t_2(\underline{X})\, w_2(\theta)$$

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?

- $\mathbb{E}_{Y|X}[Y] = \mathbb{E}_{Y|X}[\beta_0 + h(x) + \sigma\varepsilon] = \beta_0 + h(x) + \mathbb{E}_{Y|X}[\sigma\varepsilon] = \beta_0 + h(x)$
- $\mathbb{V}\mathrm{ar}_{Y|X}[Y] = \mathbb{V}\mathrm{ar}_{Y|X}[\underbrace{\beta_0 + h(x)}_{\text{Fijo}} + \sigma\varepsilon] = \mathbb{V}\mathrm{ar}_{Y|X}[\sigma\varepsilon] = \sigma^2\, \mathbb{V}\mathrm{ar}_{Y|X}[\varepsilon] = \sigma^2$

$$\frac{Y - \beta_0 - h(x)}{\sigma} = \varepsilon \Rightarrow \phi(y) \sim N(0,1)$$

$$F_\varepsilon(e) = P(\varepsilon \leq e) = P\left(\frac{Y - \beta_0 - h(x)}{\sigma} \leq e\right) = P(Y \leq \beta_0 + h(x) + \sigma e)$$

$$= F_Y(\beta_0 + h(x) + \sigma e)$$

$$f_\varepsilon(e) = \frac{\partial F_\varepsilon}{\partial e} = \frac{\partial F_Y}{\partial e} = \frac{\partial F_Y}{\partial Y} \cdot \frac{\partial Y}{\partial e} = \ldots$$