Estadística:
Teoría y Aplicaciones

Felipe Tobar

28 de agosto de 2019

Contenidos vistos en clases que no están en este apunte

- Clase 1: Definición de estadística, relación con probablilidades, *machine learning*, objetivo del curso.
- Clase 1: Tipos de estadísticas: frecuentistia versus bayesiana, descripción de los elementos de cada una de ellas.
- Clase 2: contexto general, intercambiabilidad, de Finetti
- Clase 3: modelo paramétrico, ejemplos, verosimilitud, condicional, posterior y contexto general (definiciones y supuestos generales del curso)

definiciones y notaciones menores

- definir borelianos de X

Capítulo 1

Estadísticos

Clase 4: 13 de agosto

1.1. Estadísticos

Un estadístico es una función de (los valores de) una variable aleatoria, definida desde el espacio muestral.

Definición 1.1.1 (Estadístico)**.** *Sea $(S, \mathcal{A}, \mu)$ un espacio de probabilidad y $X \in \mathcal{X}$ una variable aleatoria con distribución paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$. Un estadístico es una función medible de X independiente del parámetro θ.*

$$T : \mathcal{X} \to \mathcal{T} \tag{1.1}$$

$$x \mapsto T(x) \tag{1.2}$$

Es importante diferenciar el valor particular que toma $T(x)$, cuando X toma el valor específico $X = x$, de la variable aleatoria resultante de la aplicación de la función $T(\cdot)$ a la variable aleatoria X, es decir, $T(X)$. Este último tiene su propia distribución de probabilidad inducida por X y por la función T propiamente tal.

Algunos estimadores pueden ser:

$$T(x) = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad T'(x) = x, \qquad T''(x) = \text{mín}(x). \tag{1.3}$$

En términos generales, el objetivo de un estadístico es *encapsular* o *resumir* la información contenida en una muestra de datos $x = (x_1, x_2, \ldots, x_n)$ que es de utilidad

para determinar (o estimar) el parámetro de la distribución de X. Por esta razón, la función identidad o el promedio parecen cumplir, al menos intuitivamente, con esta misión. No así T'' en el ejemplo anterior.

Para formalizar esta idea, consideremos la siguiente definición

Definición 1.1.2 (Estadístico Suficiente)**.** *Sea $(S, \mathcal{A}, \mu)$ un espacio de probabilidad y $X \in \mathcal{X}$ una variable aleatoria con distribución paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$. Diremos que la función $T : \mathcal{X} \to \mathcal{T}$ es un estadístico suficiente para θ (o para X o para $\mathcal{P}$) si la ley condicional $X|T(X)$ no depende del parámetro θ, es decir,*

$$P_\theta(X \in A|T(X)),\ A \in \mathcal{B}(X), \text{no depende de } \theta. \tag{1.4}$$

Observemos entonces que si $T(X)$ es un estadístico suficiente, entonces, existe una función

$$H(\cdot,\cdot) : \mathcal{B}(X) \times \mathcal{T} \to [0,1] \tag{1.5}$$

que es una distribución de probabilidad en el primer argumento y es medible en el segundo argumento. fg

Ejemplo 1.1.1 (Estadístico suficiente trivial)**.** *Para cualquier familia paramétrica $\mathcal{P}$, el estadístico definido por*

$$T(x) = x \tag{1.6}$$

es suficiente. En efecto, $P_\theta(X \in A|X = x) = \mathbb{1}_A(x)$ no depende del parámetro de la familia.

Ejemplo 1.1.2 (Estadístico suficiente Bernoulli)**.** *Sea $x = (x_1, \ldots, x_n) \sim Ber(\theta)$, $\theta \in \Theta = [0,1]$, es decir*

$$P_\theta(X = x) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}. \tag{1.7}$$

Veamos que $T(x) = \sum x_i$ es un estadístico suficiente (por definición). En efecto

$$\begin{aligned} P(X = x|T(X) = t) &= \frac{P(T(X) = t|X = x)P(X = x)}{P(T(X) = t)} && \text{(T. Bayes)} \\ &= \frac{\mathbb{1}_{T(x)=t}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} && \text{(reemplazando modelo)} \\ &= \binom{n}{t}^{-1} && \text{(pues } T(x) = t\text{)} \end{aligned}$$

Consecuentemente, $T(x) = \sum x_i$ es estadístico suficiente.

Intuitivamente, nos gustaría poder verificar directamente de la suficiencia de un estadístico desde la distribución o densidad de una VA, o al menos verificar una condición más simple que la definición. Esto es porque verificar la no-dependencia de la distribución condicional $P(X|T)$ puede ser no trivial, engorroso o tedioso. Para esto enunciaremos el Teorema de Fisher-Neyman, el cual primero requiere revisar la siguiente definición.

Definición 1.1.3 (Familia Dominada)**.** *Una familia de modelos paramétricos* $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ *es dominada si existe una medida* μ*, tal que* $\forall \theta \in \Theta, P_\theta$ *es absolutamente continua con respecto a* μ *(denotado* $P_\theta \ll \mu$*), es decir,*

$$\forall \theta \in \Theta, A \in \mathcal{B}(X), \mu(A) = 0 \Rightarrow P_\theta(A) = 0 \tag{1.8}$$

La definición anterior puede interpretarse de la siguiente forma: si una familia de modelos paramétricos es dominada por una medida μ, entonces ninguno de sus elementos puede asignar medida (probabilidad) no nula a conjuntos que tienen medida cero bajo μ (la medida *dominante*). Una consecuencia fundamental de que la distribución P_θ esté dominada por μ está dada por el Teorema de Radon–Nikodym, el cual establece que si $P_\theta \ll \mu$, entonces la distribución P_θ tiene una densidad, es decir,

$$\forall A \in \mathcal{B}(X), P_\theta(X \in A) = \int_A p_\theta(x)\mu(\mathrm{d}x) \tag{1.9}$$

donde $p_\theta(x)$ es conocida como la densidad de P_θ con respecto a θ (o también como la derivada de Radon–Nikodym $\frac{dP_\theta}{d\mu}$).

Con la noción de Familia Dominada y de densidad de probabilidad, podemos enunciar el siguiente teorema que conecta la forma de la densidad de un modelo paramétrico con la suficiencia de su estadístico.

Clase 5: 20 de agosto

Teorema 1.1.1 (Factorización, Neyman-Fisher)**.** *Sea* $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ *una familia dominada por* μ*, entonces, T es un estadístico suficiente si y solo si existen funciones apropiadas* $g_\theta(\cdot)$ *y* $h(\cdot)$*, i.e., medibles y no-negativas, tal que la densidad de las distribuciones en* $\mathcal{P}$ *se admiten la factorización*

$$p_\theta(x) = g_\theta(T(x))h(x) \tag{1.10}$$

El Teorema de Neyman-Fisher es clave para evaluar, directamente de la densidad de un modelo, la suficiente de un estadístico. Pues al identificar la expresión de la VA que interactúa con el parámetro (en la función g_θ) es posible determinar el estadístico suficiente. Antes de ver una demostración informal del Teorema 1.1.1, revisemos un par de ejemplos.

Ejemplo 1.1.3 (Factorización Bernoulli)**.** *Notemos que la densidad de Bernoulli (que es igual a su distribución por ser un modelo discreto) factoriza tal como se describe en el Teorema 1.1.1. En efecto, consideremos $x = (x_1, \ldots, x_n) \sim$ Bernoulli(θ) y el estadístico $T(x) = \sum x_i$, entonces,*

$$p(X = x) = \underbrace{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}_{g_\theta(T(x))} \cdot \underbrace{1}_{h(x)} \tag{1.11}$$

Ejemplo 1.1.4 (Factorización Normal (varianza conocida))**.** *Consideremos ahora $x = (x_1, \ldots, x_n) \sim \mathcal{N}(\mu, \sigma^2)$, con σ^2 conocido y el estadístico $T(x) = \frac{1}{n}\sum x_i$, entonces,*

$$\begin{aligned}
p(X = x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}((x_i - \bar{x}) + (\bar{x} - \mu))^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\cancel{(x_i - \bar{x})}(\bar{x} - \mu) + (\bar{x} - \mu)^2\right) \\
&= \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}_{h(x)} \underbrace{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\bar{x} - \mu)^2\right)}_{g_\theta(T(x))}
\end{aligned}$$

A continuación, veremos la prueba del Teorema 1.1.1 para el caso discreto.

Demostración de Teorema Neyman-Fisher, caso discreto. Primero probamos la implicancia hacia la derecha ($\Rightarrow$), es decir, asumiendo que $T(X)$ es un estadístico su-

ficiente, tenemos,

$$\begin{aligned} p_\theta(X = x) &= P_\theta(X = x, T(X) = T(x)) \\ &= \underbrace{P_\theta(X = x|T(X) = T(x))}_{h(x), \text{ no depende de } \theta \text{ por hipótesis}} \underbrace{P_\theta(T(X) = T(x))}_{g_\theta(T(x))} \end{aligned}$$

es decir, la factorización deseada.

Ahora probamos la implicancia hacia la izquierda ($\Leftarrow$), es decir, asumiendo la factorización en la ecuación (1.10), tenemos que el modelo se puede escribir como

$$p_\theta(X = x|T(X) = t) = \frac{p_\theta(T(X) = t|X = x)p_\theta(X = x)}{p_\theta(T(X) = t)}$$

Donde $p_\theta(T(X) = t|X = x) = \mathbb{1}_{T(x)=t}$ y la hipótesis nos permite escribir

$$\begin{aligned} p_\theta(X = x) &= g_\theta(T(x))h(x) \\ p_\theta(T(X) = t) &= \sum_{x';T(x')=t} p_\theta(X = x') = \sum_{x';T(x')=t} g_\theta(T(x'))h(x') \end{aligned}$$

Incluyendo estas últimas dos expresiones en eq.(1.1), tenemos

$$p_\theta(X = x|T(X) = t) = \frac{\mathbb{1}_{T(x)=t}\cancel{g_\theta(T(x))}h(x)}{\sum_{x';T(x')=t}\cancel{g_\theta(T(x'))}h(x')} = \frac{\mathbb{1}_{T(x)=t}h(x)}{\sum_{x';T(x')=t}h(x')} \tag{1.12}$$

donde los términos que se cancelan son todos iguales a $g_\theta(t)$.

Finalmente, como el lado derecho de la ecuación (1.12) no depende de θ, se concluye la demostración. □

La idea de suficiencia del estadístico dice relación, coloquialmente, con la *información* contenida en el estadístico que permite *descubrir* el parámetro θ. En ese sentido, se tiene la intuición que un estadístico es suficiente si tiene la información *suficiente*. En el extremo de esta intuición, el estadístico puede ser simplemente todos los datos, i.e, $T(X) = X$, en cuyo caso la suficiencia es directa como se vio en el Ejemplo 1.1.1, sin embargo, estaremos interesado en estadísticos que son suficientes pero que contienen la mínima cantidad de información.

Sin una definición formal de *información* aún, recordemos que los estadísticos representan un resumen o una compresión de los datos mediante una función, i.e., la función $T(\cdot)$. Usando el mismo concepto, en el cual la aplicación de una función *quita información desde la preimagen a la imagen*, podemos definir el siguiente concepto.

Definición 1.1.4 (Estadístico Suficiente Minimal). *Un estadístico $T : \mathcal{X} \to \mathcal{T}$ es suficiente minimal si*

- *$T(X)$ es suficiente, y*
- *$\forall T'(X)$ estadístico suficiente, existe una función f tal que $T(X) = f(T'(X))$.*

FALTA: Ejemplo estadístico minimal, particiones suficientes y comentarios sobre particiones

Clase 6: 22 de agosto

Los estadísticos suficiente minimales están claramente definidos pero dicha definición no es útil para encontrar o construir estadístico suficiente minimales. El siguiente Teorema establece una condición que permite evaluar si un estadístico es suficiente minimal

Teorema 1.1.2 (Suficiencia minimal). *Sea $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ una familia dominada con densidades $\{p_\theta \text{ t.q. } \theta \in \Theta\}$ y asuma que existe un estadístico $T(X)$ tal que para cada $x, y \in \mathcal{X}$:*

$$\frac{p_\theta(x)}{p_\theta(y)} \text{ no depende de } \theta \Leftrightarrow T(x) = T(y) \tag{1.13}$$

entones, $T(X)$ es suficiente minimal.

Antes de probar este teorema, veamos un ejemplo aplicado a la distribución de Poisson.

Ejemplo 1.1.5. *Recordemos que la distribución de Poisson (de parámetro θ) modela la cantidad de eventos en un intervalo de tiempo de la forma y consideremos las observaciones $x = (x_1, \ldots, x_n) \sim Poisson(\theta)$ con verosimilitud*

$$p_\theta(x) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \tag{1.14}$$

Notemos que la razón de verosimilitudes para dos observaciones $x, y \in \mathcal{X}$ toma la forma

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\theta^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i}}{\prod_{i=1}^n x_i! / \prod_{i=1}^n y_i!} = \tag{1.15}$$

lo cual no depende de θ únicamente si $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, consecuentemente, $T(x) = \sum_{i=1}^n x_i$ es un estadístico suficiente de acuerdo al Teorema 1.1.2.

Demostración de Teorema 1.1.2. Primero veremos que T es suficiente. Dada la partición inducida por el estadístico $T(X)$, para un valor $x \in \mathcal{X}$ consideremos $x_T \in \{x'; T(x') = T(x)\}$, entonces

$$p_\theta(x) = \underbrace{p_\theta(x)/p_\theta(x_T)}_{h(x) \text{ indep. } \theta} \underbrace{p_\theta(x_T)}_{q_\theta(T(x))} \tag{1.16}$$

donde la no dependencia de θ se tiene por el supuesto del Teorema.

Para probar que el estadístico es suficiente minimal, asumamos que existe otro estadístico $T'(X)$, consideremos dos valores en la misma clase de equivalencia, i.e., $x, y,$ t.q. $T'(x) = T'(y)$, y veamos que (mediante el CFNF) podemos escribir la razón de verosimilitudes de la forma

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{g'_\theta(T'(x))h'(x)}{g'_\theta(T'(y))h'(y)} = \frac{h'(x)}{h'(y)}, \quad \text{pues } T'(x) = T'(y) \tag{1.17}$$

consecuentemente, el enunciado nos permite aseverar que como $\frac{p_\theta(x)}{p_\theta(y)}$ no depende de θ, entonces $T(x) = T(y)$. Es decir, hemos mostrado que $T'(x) = T'(y)$ implica $T(x) = T(y)$, por lo que T es función de T'.

□

Como hemos discutido durante este capítulo, un objetivo principal de construir y estudiar estadísticos es su rol en el diseño y las propiedades de los estimadores. La noción de *completitud* es clave en esta tarea.

Definición 1.1.5 (Estadístico completo)**.** *Un estadístico $T(X)$ es completo si para toda función g, se tiene que*

$$\mathbb{E}\left(g(T)|\theta\right) = 0, \forall\theta \in \Theta \Rightarrow Pr(g(T) = 0) = 1 \tag{1.18}$$

El concepto de completitud dice relación con la construcción de estimadores usando estadísticos, lo cual puede ser ilustrado mediante el siguiente ejemplo

Ejemplo 1.1.6. *Consideremos dos estimadores, ϕ_1, ϕ_2 insesgados de θ distintos, es decir,*

$$\mathbb{E}\left(\phi_1\right) = \mathbb{E}\left(\phi_2\right) = \theta, \ \mathbb{P}_\theta\left(\phi_1 \neq \phi_2\right) > 0 \tag{1.19}$$

Definamos ahora $\phi = \phi_1 - \phi_2$, donde verificamos que $\mathbb{E}\left(\phi\right) = 0, \forall\theta$, es decir, ϕ es un estimador insesgado de cero. Sin embargo, del supuesto anterior tenemos que $\mathbb{P}_\theta\left(\phi_1 - \phi_2 = 0\right) > 0$, por lo que de acuerdo a la definición anterior, el estadístico ϕ no es completo.

Intuitivamente entonces, podemos entender la noción de completitud como lo siguiente: un estadístico es completo si la única forma de construir un estimador insesgado de cero a partir de él es aplicándole la función idénticamente nula. Veamos un ejemplo de la distribución Bernoulli, donde el estadístico $T(x) = \sum x_i$ es efectivamente completo.

Ejemplo 1.1.7. *Sea $x = (x_1, \ldots, x_n)$ observaciones de $X \sim Ber(\theta)$, recordemos que $T(x) = \sum x_i \sim Bin(n, \theta)$, por lo que la esperanza $g(T)$ está dada por*

$$\mathbb{E}_\theta(g(T)) = \sum_{t=0}^{n} g(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = (1-\theta)^n \sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t \quad (1.20)$$

es decir un polinomio de grado t en $r = \theta/(1-\theta) \in \mathbb{R}_+$, entonces, $\mathbb{E}_\theta(g(T)) = 0$ implica que necesariamente los pesos de este polinomio sean todos idénticamente nulos, es decir,$g(T) = 0$. Consecuentemente, $T(x) = \sum x_i \sim Bin(n, \theta)$ es un estadístico completo.

1.2. La familia exponencial

Hasta este punto, hemos considerado algunas distribuciones paramétricas, tales como Bernoulli, Gaussiana o Poisson, para ilustrar distintas propiedades y definiciones de los estadísticos. En esta sección, veremos que realmente todas estas distribuciones (y otras más) pueden escribirse de forma unificada. Para esto, consideremos la siguiente expresión llamada *log-normalizador* (la razón de este nombre será clarificada en breve).

$$A(\eta) = \log \int_{\mathcal{X}} \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x)\mathrm{d}x \quad (1.21)$$

donde definimos lo siguiente:

- $\eta = [\eta_1, \ldots, \eta_s]^\top$ es el parámetro natural
- $T = [T_1, \ldots, T_s]^\top$ es un estadístico
- $h(x)$ es una función no-negativa

Definamos la siguiente función de densidad de probabilidad parametrizada por $\eta \in \{\eta | A(\eta) < \infty\}$

$$p_\eta(x) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x) \quad (1.22)$$

donde el hecho que $p_\eta(x)$ integra uno puede claramente verificarse reemplazando la ecuación (1.21) en (1.22), con lo cual se puede ver que A definido en (1.21) es precisamente el logaritmo de la constante de normalización de la densidad definida en (1.22).

Clase 7: 27/8

Notemos que el estadístico T es en efecto une estadístico suficiente para ν en la familia exponencial. En efecto, notemos que

$$p_\eta(x) = \underbrace{\exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right)}_{g_\theta(T(x))} \underbrace{h(x)}_{h(x)} \tag{1.23}$$

consecuentemente, por el CFNF en el Teorema 1.1.1, tenemos que T es un estadístico suficiente para ν.

Muchas de las distribuciones que usalmente consideramos pertenecen a la familia exponencial, por ejemplo, la distribución normal, exponencial, gamma, chi-cuadrado, beta, Dirichlet, Bernoulli, categórica, Poisson, Wishart (inversa) y geométrica. Otras distribuciones solo pertenecen a la familia exponencial para una determinada elección de sus parámetros, como lo ilustra el siguiente ejemplo.

Ejemplo 1.2.1 (El modelo binomial pertenece a la familia exponencial)**.** *Recordemos la distribución binomial está dada por*

$$\begin{aligned} Bin\,(x|\theta, n) &= \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x \in \{0, 1, 2\ldots, n\} \\ &= \underbrace{\binom{n}{x}}_{h(x)} \exp\left(x \underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{\text{parámetro natural}} + \underbrace{n \log\left(1-\theta\right)}_{-A(\theta)}\right) \end{aligned}$$

consecuentemente, para que $h(x)$ sea únicamente una función de la variable aleatoria, entonces el número de intentos n tiene que se una cantidad conocida, ***no un parámetro****.*

Falta: dar ejemplos de cómo las distribuciones conocidas (Bernoulli, Gaussian,Poisson, etc) se pueden generar desde la ecuación (1.22)

La familia exponencial va a ser ampliamente usada durante el curso, lo cual se debe a sus propiedades favorables para el análisis estadístico. Por ejemplo, el producto de dos distribuciones de la familia exponencial también pertenece a la familia exponencial. En efecto, consideremos dos VA X_1, X_2, con distribuciones en la familia exponencial respectivamente dadas por

$$p_1(x_1) = h_1(x_1) \exp\left(\theta_1 T_1(x_1) - A_1(\theta_1)\right) \tag{1.24}$$
$$p_2(x_2) = h_2(x_2) \exp\left(\theta_2 T_2(x_2) - A_2(\theta_2)\right) \tag{1.25}$$

si asumimos que estas VA son independientes, entonces densidad conjunta de $X = (X_1, X_2) \sim p$ está dada por

$$\begin{aligned} p(X) &= p_1(x_1) p_2(x_2) \\ &= \underbrace{h_1(x_1) h_2(x_2)}_{h(x)} \exp\left(\underbrace{[\theta_1, \theta_2]}_{\theta} \underbrace{\begin{bmatrix} T_1(x_1) \\ T_2(x_2) \end{bmatrix}}_{T(x)} - \underbrace{(A_1(\theta_1) + A_2(\theta_2))}_{A(\theta)} \right) \end{aligned} \tag{1.26}$$

con lo que eligiendo $\theta = [\theta_1, \theta_2]$ y $T = [T_1, T_2]$, vemos que X está dado por una distribución de la familia exponencial.

Otra propiedad de las familia exponencial es la relación entre los momentos de la distribucion y el lognormalizador A. Denotando

$$Q(\theta) = \exp\left(A(\theta)\right) = \int_{\mathcal{X}} \exp\left(\theta T(x)\right) h(x) \mathrm{d}x \tag{1.27}$$

Observemos que la derivada de $A(\theta)$ está dada por

$$\begin{aligned} \frac{dA(\theta)}{d\theta} &= Q^{-1}(\theta) \frac{dQ(\theta)}{d\theta} \\ &= \frac{\int_{\mathcal{X}} T(x) \exp\left(\theta T(x)\right) h(x) \mathrm{d}x}{\int_{\mathcal{X}} \exp\left(\theta T(x)\right) h(x) \mathrm{d}x} \\ &= \frac{\int_{\mathcal{X}} T(x) \exp\left(\theta T(x) - A(\theta)\right) h(x) \mathrm{d}x}{\int_{\mathcal{X}} \exp\left(\theta T(x) - A(\theta)\right) h(x) \mathrm{d}x} \qquad \cdot A(\theta)/A(\theta) \\ &= \mathbb{E}\left(T(x)\right) \end{aligned} \tag{1.28}$$

Capítulo 2

Estimadores

Consideremos una función del parámetro de una familia paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$, $g(\theta)$. Un estimador puntual de $g(\theta)$ es un estadístico, es decir, una función de la VA X, que toma valores en el mismo conjunto que $g(\Theta)$. En general denotaremos como $\hat{g}(X)$ el estimador de $g(\theta)$ aplicado a X

Ejemplo 2.0.1 (Estimador de la media Gaussiana)**.** *Consideremos $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu, \sigma^2)$. Un estimador de $g(\theta) = g(\mu, \sigma) = \mu$ es el estadístico*

$$\hat{g}(X) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.1}$$

Una clase muy importante de estimadores son los estimadores insesgados.

Definición 2.0.1 (Estimador insesgado)**.** *Sea $\hat{g}(x)$ un estimador de $g(\theta)$. Este estimador es insesgado si*

$$\mathbb{E}\left(\hat{g}(x)\right) = g(\theta) \tag{2.2}$$

y el sesgo de $\hat{g}$ se define como

$$b_{\hat{g}}(\theta) = \mathbb{E}\left(\hat{g}(x)\right) - g(\theta) \tag{2.3}$$

Los estimadores insesgados juegan un rol importante en el estudio y aplicación de la estadística, sin embargo, uno no siempre debe poner exclusiva atención a ellos. Los siguiente ejemplos ilustran el rol del estimador insesgado en dos familias paramétricas distintas.

Ejemplo 2.0.2 (Estimador insesgado de la media Gaussiana)**.** *El estimador de $g(\theta) = \mu$ descrito en el Ejemplo 2.0.1 es insesgado, en efecto:*

$$\mathbb{E}\left(\hat{g}(x)\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \tag{2.4}$$

Ejemplo 2.0.3 (Estimador de la taza de la distribución exponencial[1])**.** *Consideremos $X \sim Exp(\theta)$, donde $Exp(x|\theta) = \theta\exp(-\theta x)$, y asumamos que existe un estimador insesgado $\hat{g}(X)$ de $g(\theta) = \theta$, entonces,*

$$\mathbb{E}\left(\hat{g}(X)\right) = \int_0^\infty \hat{g}(x)\theta\exp(-\theta x)dx = \theta, \forall\theta, \tag{2.5}$$

lo cual es equivalente a decir que $\int_0^\infty \hat{g}(x)\exp(-\theta x)dx = 1, \forall\theta$ o bien que (al derivar ambos lados de esta expresión c.r.a. θ) $\int_0^\infty x\hat{g}(x)\exp(-\theta x)dx = 0, \forall\theta$.

Esta última expresión es equivalente a que $\mathbb{E}\left(X\hat{g}(X)\right) = 0$, lo que a su vez y considerando que X es un estadístico suficiente y completo, implica que necesariamente la función $X\hat{g}(X) = 0$ c.s. $\forall\theta$, y también que $\hat{g}(X) = 0$ c.s. $\forall\theta$. Como esto contradice el hecho de que $\hat{g}(X)$ es insesgado, no es posible construir estimadores insesgados para θ en la distribución exponencial.

Veamos ahora un ejemplo de un estimador sesgado de la varianza y cómo se puede construir un estimador insesgado.

Ejemplo 2.0.4. *Consideremos una familia paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ y denotemos por μ y σ^2 su media y su varianza respectivamente. Usando las observaciones $x_1, x_2, \ldots, x_n$, calculemos la varianza del estimador de la media, dado por $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ mediante*

$$\mathbb{V}_\theta\left(\bar{x}\right) = \mathbb{V}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \underbrace{=}_{i.i.d.} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}_\theta\left(x_i\right) = \frac{\sigma^2}{n} \tag{2.6}$$

es decir, el estimador de la media usando n muestras, tiene una varianza σ^2/n.

Consideremos ahora el siguiente estimador para la varianza:

$$S_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.7}$$

[1] Schervish

y notemos que la esperanza de dicho estimador es

$$\begin{aligned}
\mathbb{E}_\theta\left(S_2\right) &= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu+\mu-\bar{x})^2\right)\\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2+2\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(\mu-\bar{x})+\frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{x})^2\right)\\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-2(\mu-\bar{x})^2+(\mu-\bar{x})^2\right)\\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-(\mu-\bar{x})^2\right)\\
&= \mathbb{V}_\theta\left(x_i\right)-\mathbb{V}_\theta\left(\bar{x}\right) \quad \textit{ver ecuación (2.6)}\\
&= \sigma^2+\sigma^2/n=\left(\frac{n+1}{n}\right)\sigma^2
\end{aligned} \tag{2.8}$$

Esto quiere decir que el sesgo del estimado en la ecuación (2.7) *es asintóticamente insesgado, es decir, que su sesgo tiende a cero cuando el número de muestas n tiende a infinito. Sin embargo, notemos que podemos corregir el estimado de la varianza multiplicando el estimador original, S_2 en la ecuación* (2.7) *por $n/(n+1)$, con lo que el estimador corregido denotado por*

$$S_2' = \frac{n}{n+1}S_2 = \frac{1}{n+1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{2.9}$$

cumple con

$$\mathbb{E}_\theta\left(S_2'\right) = \left(\frac{n}{n+1}\right)\mathbb{E}_\theta\left(S_2\right) \underbrace{=}_{ec.(2.8)} \left(\frac{n}{n+1}\right)\left(\frac{n+1}{n}\right)\sigma^2 = \sigma^2 \tag{2.10}$$

es decir, el estimador S_2' en la ecuación (2.9) *es insesgado.*

Clase 8: 29/8

Para tener una notación más limpia, desde ahora nos referiremos a estimadores $\phi = \hat{g}$ de θ en general para evitar la expresión más engorrosa estimador $\hat{g}(X)$ de $g(\theta)$.

Es natural evaluar la bondad de distintos estimadores (sesgados o insesgados), una forma de hacer esto es definir una función de *pérdida* o *costo* que compara el

valor reportado por el estimador y el valor real del parámetro. En general, elegimos la función de pérdida cuadrática para un estimador ϕ y un parámetro θ definida por

$$L_2(\phi, \theta)^2 = (\phi - \theta)^2. \tag{2.11}$$

Luego, como el estimador es una VA, también lo es función de pérdida, por lo que podemos calcular la esperanza de la función de pérdida, lo cual conocemos como *riesgo*. El riesgo asociado a la pérdida cuadrática en la ecuación anterior está dado por:

$$\begin{aligned} R(\theta, \hat{g}) &= \mathbb{E}\left((\theta - \phi)^2\right) \\ &= \mathbb{E}\left((\theta - \bar{g} + \bar{g} - \phi)^2\right); \quad \text{denotando } \bar{\phi} = \mathbb{E}(\phi) \\ &= \mathbb{E}\left((\theta - \bar{\phi})^2 + 2(\theta - \bar{\phi})\cancel{(\bar{\phi} - \phi)} + (\bar{\phi} - \phi)^2\right) \\ &= \underbrace{(\theta - \bar{\phi})^2}_{=b_\phi^2 \text{ (sesgo}^2\text{)}} + \underbrace{\mathbb{E}\left((\bar{\phi} - \phi)^2\right)}_{=V_\phi \text{ (varianza)}}. \end{aligned} \tag{2.12}$$

Con esta métrica para comparar estimadores, el siguiente teorema establece que la información reportada por un estadístico suficiente (Definición 1.1.2), puede solo mejorar un estimador.

Teorema 2.0.1 (Teorema de Rao-Blackwell)**.** *Sea $\phi = \phi(X)$ un estimador de θ tal que $\mathbb{E}_\theta(\phi) < \infty, \forall\theta$. Asumamos que existe T estadístico suficiente para θ y sea $\phi^\star = \mathbb{E}_\theta(\phi|T)$. Entonces,*

$$\mathbb{E}_\theta\left((\phi^\star - \theta)^2\right) \leq \mathbb{E}_\theta\left((\phi - \theta)^2\right), \forall\theta, \tag{2.13}$$

donde la desigualdad es estricta salvo en el caso donde ϕ es función de T.

En otras palabras, el Teo. de Rao-Blackwell establece que un estimador puede ser *mejorado* si es reemplazado por su esperanza condicional dado un estadístico suficiente. El proceso de mejorar un estimador poco eficiente de esta forma es conocido como *Rao-Blackwellización* y veremos un ejemplo a continuación.

Ejemplo 2.0.5. *Consideremos $X = (X_1, \ldots, X_n) \sim$ Poisson (θ) y estimemos el parámetro θ. Para esto, consideremos el estimador básico $\phi = X_1$ y Rao-Blackwellicémoslo usando el estimador suficiente $T = \sum_{i=1}^n X_i$, es decir,*

$$\phi^* = \mathbb{E}_\theta\left(X_1 \middle| \sum_i X_i = t\right). \tag{2.14}$$

Para calcular esta esperanza condicional, observemos primero que

$$\sum_{j=1}^{n} \mathbb{E}_\theta \left(X_j \middle| \sum_{i=1}^{n} X_i = t \right) = \mathbb{E}_\theta \left(\sum_{j=1}^{n} X_j \middle| \sum_{i=1}^{n} X_i = t \right) = t \tag{2.15}$$

y que como $X_1, \ldots, X_n$ son iid, entonces todos los términos dentro de la suma del lado izquierdo de la ecuación anterior son iguales. Consecuentemente, recuperamos el estimador

$$\phi^* = \frac{t}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.16}$$

Para demostrar el Teorema 2.0.1 consideremos dos variable aleatorias $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, y recordemos dos propiedades básicas. En primer lugar la ley de esperanzas totales, la cual establece que

$$\begin{aligned}
\mathbb{E}_Y \mathbb{E}_{X|Y}(X|Y) &= \int_{\mathcal{Y}} \int_{\mathcal{X}} x \mathrm{d}P(x|y) \mathrm{d}P(y) && \text{def. esperanza} \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} \mathrm{d}P(x|y) \mathrm{d}P(y) && \text{linealidad} \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} \mathrm{d}P(x,y) && \text{def. esperanza condicional} \\
&= \int_{\mathcal{X}} x \mathrm{d}P(x) = \mathbb{E}_X(X) && \text{def. esperanza}
\end{aligned} \tag{2.17}$$

y la desigualdad de Jensen, la cual para el caso particular del costo cuadrático, puede verificarse que

$$0 \leq \mathbb{V}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2 \Rightarrow \mathbb{E}\left(X^2\right) \geq \mathbb{E}(X)^2 . \tag{2.18}$$

Falta: dibujo con la intuición detrás de Jensen en el caso general

Entonces, utilizando las expresiones en (2.17) y (2.18), podemos demostrar el teorema anterior.

Demostración de Teorema 2.0.1. La varianza del estimador $\phi^\star$ está dada por

$$\begin{aligned}
\mathbb{E}_\theta\left((\phi^\star - \theta)^2\right) &= \mathbb{E}_\theta\left((\mathbb{E}_\theta(\phi|T) - \theta)^2\right) && \text{def.} \\
&= \mathbb{E}_\theta\left((\mathbb{E}_\theta(\phi - \theta|T))^2\right) && \text{linealidad} \\
&\leq \mathbb{E}_\theta\left(\mathbb{E}_\theta\left((\phi - \theta)^2|T\right)\right) && \text{Jensen} \\
&= \mathbb{E}_\theta\left((\phi - \theta)^2\right) && \text{ley esperanzas totales}
\end{aligned}$$

Donde las esperanzas exteriores son con respecto a T y las interiores con respecto a X (o equivalentemente a ϕ). Observemos además que la desigualdad anterior viene de la expresión en la ecuación (2.18), por lo que la igualdad es obtenida si $\mathbb{V}(\phi - \theta|T) = 0$, es decir, la VA $\phi - \theta$ tiene que ser constante para cada valor de T, es decir, ϕ es función de T. Intuitivamente podemos entender esto como que si el estadístico ya fue considerado en el estimador, entonces conocer el valor del estadístico no reporta información adicional. □

Observación 2.0.1. *Notemos que si el estimador ϕ es insesgado, su Rao-Blackwellización ϕ^* también lo es, en efecto*

$$\mathbb{E}_\theta(\phi^*) = \mathbb{E}_\theta(\mathbb{E}_\theta(\phi|T)) = \mathbb{E}_\theta(\phi) = \theta, \tag{2.19}$$

donde la segunda igualdad está dada por la ley de esperanzas totales y la tercera por el supuesto de que ϕ es insesgado.

En base al riesgo cuadrático definido en la ecuación (2.12), podemos ver que si un estimador es insesgado (Definición 2.0.1), su riesgo cuadrático es únicamente su varianza. Esto motiva la siguiente definición de optimalidad para estimadores insesgados.

Definición 2.0.2 (Estimador insesgado de varianza uniformemente mínima)**.** *El estimador ϕ de θ es un estimador insesgado de varianza uniformemente mínima (EIVUM) si es insesgado y si $\forall \phi' : \mathcal{X} \to \Theta$ estimador insesgado se tiene*

$$\mathbb{V}_\theta(\phi) \leq \mathbb{V}_\theta(\phi'), \forall \theta \in \Theta. \tag{2.20}$$

Ejemplo 2.0.6. *Consideremos $x = (x_1, \ldots, x_n) \sim Ber(\theta)$ y los siguientes estimadores de θ*

- $\phi_1(x) = x_1$
- $\phi_2(x) = \frac{1}{2}(x_1 + x_2)$
- $\phi_3(x) = \frac{1}{n}\sum_{i=1}^n x_i$

Observemos que todos estos estimadores son insesgados, pues como $\forall i, \mathbb{E}_\theta(x_i) = \theta$, entonces

$$\mathbb{E}_\theta(\phi_1(x)) = \mathbb{E}_\theta(\phi_2(x)) = \mathbb{E}_\theta(\phi_3(x)) = \theta \tag{2.21}$$

Veamos ahora que la varianza de $\phi_3(x)$ está dada por

$$\mathbb{V}_\theta\left(\phi_3(x)\right) = \mathbb{V}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}_\theta\left(x_i\right) = \frac{\theta(1-\theta)}{n} \tag{2.22}$$

pues $\mathbb{V}_\theta\left(x_i\right) = \mathbb{E}_\theta\left((\theta - x_i)\right) = \mathbb{E}_\theta\left(x_i^2\right) - \theta^2 = (0^2\cdot(1-\theta) + 1^2\cdot\theta) - \theta^2 = \theta(1-\theta)$. Consecuentemente, la varianza de los estimadores considerados decae como la inversa del número de muestras.

Con las definiciones anteriores, podemos mencionar el siguiente teorema, el cual conecta la noción de estadístico completo con la de EIVUM.

Teorema 2.0.2 (Teorema de Lehmann-Scheffé)**.** *Sea X una VA con distribución paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ y T un estadístico suficiente y completo para θ. Si el estimador $\phi = \phi(T)$ de θ es insesgado, entonces ϕ es el único EIVUM.*

Demostración. Veamos en primer lugar que es posible construir un estimador en función del estadístico $\phi(T)$ que tiene menor o igual varianza que un estimador arbitrario $\phi'(X)$. En efecto, el Teorema de Rao-Blackwell estable ce que el estimador

$$\phi(T) = \mathbb{E}_\theta\left(\phi'(X)|T\right), \tag{2.23}$$

tiene efectivamente menos varianza que $\phi'(X)$.

Ahora veamos que solo existe un único estimador insesgado que es función de T, en efecto, si existiesen dos estimadores insesgados de θ, $\phi_1(T), \phi_2(T)$, entonces, $\mathbb{E}_\theta\left(\phi_1(T) - \phi_1(T)\right) = 0$ y como T es completo, entonces, $\phi_1(T) = \phi_2(T)$ c.s.-P_θ.

Hemos probado que (i) para un estimador arbitrario, se puede construir un estimador que es función de T el cual tiene menor o igual varianza que el estimador original y, (ii) el estimador insesgado $\phi(T)$ es único. Consecuentemente, $\phi(T)$ es el único EIVUM. □

El Teorema de Lehmann-Scheffé da una receta para encontrar el EIVUM: simplemente es necesario encontrar un estadístico completo y construir un estimador insesgado en base a éste, esto garantiza que el estimador construido es el **único** EIVUM.

Ejemplo 2.0.7 (EIVUM para Bernoulli)**.** *Recordemos que en el Ejemplo 1.1.7 vimos que el estadístico $T = \sum_{i=1}^{n} X_i$ es completo para $X \sim Ber\,(\theta)$. Como el estimador de θ dado por $\phi(T) = T/n$ es insesgado,*

$$\mathbb{E}_\theta\left(\phi(T)\right) = \mathbb{E}_\theta\left(T/n\right) = \sum_{i=1}^{n}\mathbb{E}_\theta\left(X_i\right)/n = \theta, \tag{2.24}$$

entonces $\phi(T) = T/n$ *es el EIVUM para* θ *en Ber* (θ).