Estadística:
Teoría y Aplicaciones

Felipe Tobar

17 de octubre de 2019

Índice general

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

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

Definición 1.0.1 (Estadístico). *Sea $(S, \mathcal{A}, \mu)$ un espacio de probabilidad y $X \in \mathcal{X}$ una variable aleatoria con distribución paramétrica $\mathcal{P} = \{P_\theta \ \ t.q. \ \ \theta \in \Theta\}$. 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:

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

1.1. Suficiencia

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 uti-

lidad 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 1.1.1 (Estadístico Suficiente)**.** *Sea $(S, \mathcal{A}, \mu)$ un espacio de probabilidad y $X \in \mathcal{X}$ una variable aleatoria con distribución paramétrica $\mathcal{P} = \{P_\theta \ \ t.q. \ \ \theta \in \Theta\}$. 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 1.1.2 (Familia Dominada)**.** *Una familia de modelos paramétricos* $\mathcal{P} = \{P_\theta \ \ t.q. \ \ \theta \in \Theta\}$ *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* $\mathcal{P} = \{P_\theta \ \ t.q. \ \ \theta \in \Theta\}$ *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. □

1.2. Suficiencia minimal y completitud

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 1.2.1 (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 1.2.1 (Suficiencia minimal)**.** *Sea $\mathcal{P} = \{P_\theta \ \ t.q. \ \ \theta \in \Theta\}$ una familia dominada con densidades $\{p_\theta \ \ t.q. \ \ \theta \in \Theta\}$ 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 1.2.1. *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 1.2.1.

Demostración de Teorema 1.2.1. 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 1.2.2 (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 1.2.2. *Consideremos dos estimadores, ϕ_1, ϕ_2 insesgados de θ distintos, es decir,*

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

Definamos ahora $\phi = \phi_1 - \phi_2$, donde verificamos que $\mathbb{E}(\phi) = 0, \forall\theta$, es decir, ϕ es un estimador insesgado de cero. Sin embargo, del supuesto anterior tenemos que $\mathbb{P}(\phi_1 - \phi_2 = 0) > 0$, 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 1.2.3. *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*

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

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.

1.3. 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).

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

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 $\eta \in \{\eta | A(\eta) < \infty\}$

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

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

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

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

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

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*

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

2.1. Estimadores insesgados

Una clase muy importante de estimadores son los estimadores insesgados.

Definición 2.1.1 (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 2.1.1 (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 2.1.2 (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 2.1.3. *Consideremos una familia paramétrica $\mathcal{P} = \{P_\theta \ \ t.q. \ \ \theta \in \Theta\}$ 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

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

Esto quiere decir que el sesgo del estimador 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 estimador 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

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

cumple con

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

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

Clase 8: 29/8

2.2. Teorema de Rao-Blackwell

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 la 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:

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

Con esta métrica para comparar estimadores, el siguiente teorema establece que la información reportada por un estadístico suficiente (Definición 1.1.1), puede solo mejorar un estimador.

Teorema 2.2.1 (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 2.2.1. *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 2.2.1 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 2.2.1. 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 2.2.1. *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.

2.3. Varianza uniformemente mínima

En base al riesgo cuadrático definido en la ecuación (2.12), podemos ver que si un estimador es insesgado (Definición 2.1.1), su riesgo cuadrático es únicamente su varianza. Esto motiva la siguiente definición de optimalidad para estimadores insesgados.

Definición 2.3.1 (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 2.3.1. *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

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

pues $\mathbb{V}_\theta(x_i) = \mathbb{E}_\theta((\theta - x_i)) = \mathbb{E}_\theta(x_i^2) - \theta^2 = (0^2 \cdot (1-\theta) + 1^2 \cdot \theta) - \theta^2 = \theta(1-\theta)$. 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 2.3.1 (Teorema de Lehmann-Scheffé)**.** *Sea X una VA con distribución paramétrica $\mathcal{P} = \{P_\theta \ \ t.q. \ \ \theta \in \Theta\}$ 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, $\mathbb{E}_\theta(\phi_1(T) - \phi_1(T)) = 0$ 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 2.3.2 (EIVUM para Bernoulli)**.** *Recordemos que en el Ejemplo 1.2.3 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,*

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

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

Clase 9: 3/9

2.4. Máxima Verosimilitud

Hasta ahora, hemos estudiado distintas propiedades de estimadores (y estadísticos en general) y distintas relaciones entre ellos. Esto nos ha permitido evaluar si un estimador dado cumple con ciertas características como ser insesgado o tener varianza mínima, adicionalmente, hemos visto como *mejorar* un estimador crudo mediante el Teorema de Rao-Blackwell. Sin embargo, nuestro estudio siempre ha comenzado con un estimador disponible en vez de construir un estimador, lo cual en la práctica puede ser no trivial. En esta sección, veremos cómo construir estimadores usando directamente la densidad de probabilidad de la VA $X \in \mathcal{X}$, para cual recordemos la siguiente definición

Definición 2.4.1 (Función de verosimilitud)**.** *Sea* $X \in \mathcal{X}$ *una VA con distribución paramétrica* $\mathcal{P} = \{P_\theta \ \ t.q. \ \ \theta \in \Theta\}$*, donde* P_θ *tiene densidad* p_θ*. Dada la observación* $x \in \mathcal{X}$*, la verosimilitud del parámetro de* θ *está definida por*

$$\begin{aligned} L : \Theta &\to \mathbb{R}_+ \\ \theta &\mapsto L(\theta|x) = p_\theta(x). \end{aligned}$$

Adicionalmente, nos referiremos a $l(\theta|x) = \log L(\theta|x)$ *como la log-verosimilitud.*

La definición anterior simplemente establece que la verosimilitud es la distribución (conjunta) de los datos pero donde tomamos los datos como fijos y el parámetro como variable, lo cual tiene sentido en aplicaciones de modelos estadísticos donde los datos son fijos y conocidos pero el modelo (parámetro) no lo es. Una consecuencia importante de este concepto es que la verosimilitud no es una densidad de probabilidad (en θ) pues no integra 1 (con respecto a θ). Notemos que nos referimos a la **verosimilitud del parámetro** θ como la densidad de x dado θ (y no al revés).

La verosimilitud da las condiciones para determinar un estimador que recibe mucha atención en la literatura estadística:

Definición 2.4.2 (Estimador de máxima verosimilitud (MV))**.** *Sea una observación x y una función de verosimilitud $L(\theta)$, el estimador de máxima verosimilitud es*

$$\theta_{MV} = \arg\max_{\theta} L(\theta|x) \tag{2.25}$$

Claramente, el estimador de MV puede ser definido con respecto a la verosimilitud o a cualquier función no decreciente de ésta, como también pude no existir o no ser único. En particular, nos enfocaremos en encontrar θ_{MV} mediante la maximización de la log-verosimilitud, la cual es usualmente más fácil de optimizar en términos computacionales o analíticos. De hecho, muchas veces incluso ignoraremos constantes de la (log) verosimilitud, pues éstas no cambian el máximo de $L(\theta)$.

Ejemplo 2.4.1 (Máxima verosimilitud: Bernoulli)**.** *Sea $X_1, \dots X_n \sim Ber\,(\theta)$, la verosimilitud de θ está dada por*

$$L(\theta) = \prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i} \tag{2.26}$$

y su log-verosimilitud por $l(\theta) = (\sum_{i=1}^n x_i) \log\theta + (n - \sum_{i=1}^n x_i) \log(1-\theta)$. El estimador de MV puede ser encontrado resolviendo $\frac{\partial l(\theta)}{\partial \theta} = 0$:

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \theta} = 0 \Rightarrow (\sum_{i=1}^{n} x_i)\theta^{-1} &= (n - \sum_{i=1}^{n} x_i)(1-\theta)^{-1} \\
\Rightarrow \sum_{i=1}^{n} x_i(1-\theta) &= (n - \sum_{i=1}^{n} x_i)\theta \\
\Rightarrow \theta &= \sum_{i=1}^{n} x_i/n.
\end{aligned}$$

Notemos que este estimador de MV ¡es a su vez el EIVUM!

Ejercicio 2.4.1. *Graficar $l(\theta)$ en el Ejemplo 2.4.1.*

Ejercicio 2.4.2. *Encuentre el estimador de MV de $\theta = (\mu, \Sigma)$ para la VA $X \sim \mathcal{N}(\mu, \Sigma)$.*

Ejemplo 2.4.2. *Sea la VA $X \sim$ Uniforme (θ), es decir, $p(x) = \theta^{-1}\mathbb{1}_{0\leq x\leq\theta}$. Para calcular la verosimilitud, recordemos en primer lugar que la verosimilitud factoriza de acuerdo a*

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i) \tag{2.27}$$

y observemos que necesariamente $p_\theta(x_i) = 0$ si $x_i > \theta$. Consecuentemente, $L(\theta) > 0$ solo si θ es mayor que toda las observaciones, en particular, si $\theta \geq \text{máx}\{x_i\}_1^n$.

Además, si efectivamente tenemos $\theta \geq \text{máx}\{x_i\}_1^n$, entonces notemos que $p_\theta(x_i) = 1/\theta$, por lo que la verosimilitud está dada por

$$L(\theta) = \theta^{-n}, \quad \theta \geq \text{máx}\{x_i\}_1^n \tag{2.28}$$

y consecuentemente, el estimador de máxima verosimilitud es $\theta_{MV} = \text{máx}\{x_i\}_1^n$

FALTA: propiedades del estimador de MV: familia exponencial, consistent, equivariant, asymptotically Normal, asymptotically optimal.

2.5. Estimador de MV en la práctica: tres ejemplos

2.5.1. Regresión lineal y gaussiana

Una aplicación muy popular del estimador de MV es en los modelos de regresión lineal y gaussianos. Consideremos el caso donde se desea modelar la cantidad de pasajeros que mensualmente viajan en una aerolínea, para esto, sabemos de nuestros colaboradores en la división de análisis de datos de la aerolínea que ésta cantidad tiene una tendencia de crecimiento cuadrática en el tiempo y además una componente oscilatoria de frecuencia anual. Estos fenómenos pueden ser explicados por el aumento de la población, los costos decrecientes de la aerolínea y la estacionalidad anual de las actividades económicas.

Asumiendo que la naturaleza de la cantidad de pasajeros es estocástica, podemos usar los supuestos anteriores para modelar la densidad condicional de dicha

cantidad (con respecto al tiempo t) mediante una densidad normal parametrizada de acuerdo a

$$X \sim \mathcal{N}\left(\theta_0 + \theta_1 t^2 + \theta_2 \cos(2\pi t/12), \theta_3^2\right), \tag{2.29}$$

donde $\theta_0, \theta_1, \theta_2$ parametrizan la media y θ_3 la varianza.

Consecuentemente, si nuestras observaciones están dadas por $\{(t_i, x_i)\}_{i=1}^n$ podemos escribir la log-verosimilitud de θ como

$$\begin{aligned} l(\theta) &= \log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta_3^2}} \exp\left(-\frac{(x_i - \theta_0 - \theta_1 t^2 - \theta_2 \cos(2\pi t/12))^2}{2\theta_3^2}\right)\right) \\ &= \frac{n}{2}\log\left(2\pi\theta_3^2\right) - \frac{1}{2\theta_3^2}\sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i^2 - \theta_2 \cos(2\pi t_i/12))^2 \end{aligned} \tag{2.30}$$

con lo que vemos que θ_{MV} puede ser calculado explícitamente y es función de $\{(t_i, x_i)\}_{i=1}^n$ debido a que la ecuación (2.30) es cuadrática en $[\theta_0, \theta_1, \theta_2]$.

2.5.2. Regresión no lineal: clasificación

La razón por la cual θ_{MV} pudo ser calculado de forma explícita es porque el modelo Gaussiano con media parametrizada de forma lineal resulta en una log-verosimilitud cuadrática, donde el mínimo es único y explícito. Sin embargo, en muchas situaciones el modelo lineal y gaussiano no es el apropiado.

Un ejemplo es esto es problema de evaluación crediticia (*credit scoring*) donde en base a un conjunto de *características* que definen a un cliente, un ejecutivo bancario debe evaluar si otorgarle o no el crédito que el cliente solicita. Para tomar esta decisión, el ejecutivo puede revisar la base de datos del banco e identificar los clientes que en el pasado pagaron o no pagaron sus créditos para determinar el perfil del *pagador* y el del *no-pagador*. Finalmente, un nuevo cliente puede ser *clasificado* como pagador/no-pagador en base su similaridad con cada uno de estos grupos.

Formalmente, denotemos las características del cliente como $t \in \mathbb{R}^N$ y asumamos que el cliente paga su crédito con probabilidad $\sigma(t)$ y no lo paga con probabilidad $1 - \sigma(t)$, la función $\sigma(t)$ a definir. Esto es equivalente a construir la VA X

$$X|t \sim \text{Ber}\left(\sigma(t)\right) \tag{2.31}$$

donde $X = 1$ quiere decir que el cliente paga su crédito y $X = 0$ que no. Una elección usual para la función $\sigma(\cdot)$ es la función logística aplicada a una transformación lineal de t, es decir,

$$\Pr(X = 1|t) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 t)}}. \tag{2.32}$$

Notemos que este es un clasificador lineal, donde $\theta = [\theta_0, \theta_1]$ define un hiperplano en $\mathbb{R}^N$ en donde los clientes $t \in \{t|0 \leq \theta_0 + \theta_1 t\}$ pagan con probabilidad mayor o igual a $1/2$ y el resto con probabilidad menor o igual a $1/2$. Esto es conocido como **regresión logística**.

Entonces, usando los registros bancarios $\{(x_i, t_i)\}_{i=1}^n$ ¿cuál es el $\theta = [\theta_0, \theta_1]$ de máxima verosimilitud? Para esto notemos que la log-verosimilitud puede ser escrita como

$$\begin{aligned} l(\theta) &= \log \prod_{i=1}^{n} p(x_i|t) \\ &= \sum_{i=1}^{n} x_i \log \sigma(t) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1 - \sigma(t)) \\ &= \sum_{i=1}^{n} x_i \log \frac{1}{1 + e^{-(\theta_0 + \theta_1 t)}} + \left(n - \sum_{i=1}^{n} x_i\right) \log(1 - \frac{1}{1 + e^{-(\theta_0 + \theta_1 t)}}) \end{aligned}$$

Esta expresión no tiene mínimo global y a pesar que podemos calcular su gradiente, no podemos resolver $\partial l(\theta)/\partial\theta = 0$ de forma analítica, por lo que debemos usar métodos de descenso de gradiente.

2.5.3. Variables latentes: *Expectation-Maximisation*

En ciertos escenarios es natural asumir que nuestros datos provienen de una mezcla de modelos, por ejemplo, consideremos la distribución de estaturas en una población, podemos naturalmente modelar esto como una mezcla de distribuciones marginales para las estaturas de hombres y mujeres por separado, es decir,

$$X \sim p\mathcal{N}(X|\mu_H, \Sigma_H) + (1 - p)\mathcal{N}(X|\mu_M, \Sigma_M) \tag{2.33}$$

donde la verosimilitud de los parámetros $\theta = [p, \mu_H, \sigma_H, , \mu_M, \sigma_M]$ dado un conjunto de observaciones $\{x_i\}_{i=1}^n$ es

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} (p\mathcal{N}(X|\mu_H, \Sigma_H) + (1-p)\mathcal{N}(X|\mu_M, \Sigma_M)) \\ &= \prod_{i=1}^{n} \left(p\frac{1}{\sqrt{2\pi\Sigma_H^{-1}}} \exp\left(\frac{-(x_i-\mu_H)^2}{2\Sigma_H^2}\right) + (1-p)\frac{1}{\sqrt{2\pi\Sigma_M^{-1}}} \exp\left(\frac{-(x_i-\mu_M)^2}{2\Sigma_M^2}\right)\right). \end{aligned}$$

Optimizar esta expresión con respecto a las 5 componentes de θ es difícil, en particular por la suma en la expresión, lo cual no permite simplificar la expresión mediante la aplicación de $\log(\cdot)$.

Una interpretación de la diferencia de este modelo con respecto a los anteriores es la introducción implícita de una *variable latente* que describe de qué gaussiana fue generada cada observación. Si conociésemos esta variable latente, el problema sería dramáticamente más sencillo. En efecto, asumamos que tenemos a nuestra disposición las observaciones $\{z_i\}_{i=1}^n$ de la VA $\{Z_i\}_{i=1}^n$ las cuales denota de qué modelo es generada cada observación, por ejemplo, $Z_i = 0$ (cf. $Z_i = 1$) denota que el individuo con estatura X_i es hombre (cf. mujer).

En este caso, asumamos por un segundo que estas variables latentes están disponibles y consideremos los **datos completos** $\{(x_i, z_i)\}_{i=1}^n$ para escribir la función de verosimilitud completa mediante

$$\begin{aligned} l(\theta|z_i, x_i) &= \prod_{i=1}^{n} \mathcal{N}(X|\mu_H, \Sigma_H)^{z_i} \mathcal{N}(X|\mu_M, \Sigma_M)^{(1-z_i)} \\ &= \sum_{i=1}^{n} \left(z_i \log \frac{1}{\sqrt{2\pi\Sigma_H^{-1}}} \exp\left(\frac{-(x_i-\mu_H)^2}{2\Sigma_H^2}\right) + (1-z_i) \log \frac{1}{\sqrt{2\pi\Sigma_M^{-1}}} \exp\left(\frac{-(x_i-\mu_M)^2}{2\Sigma_M^2}\right)\right). \end{aligned}$$

Esta función objetivo es mucho más fácil de optimizar, pero no es observable pues la VA Z es desconocida. Una forma de resolver esto es tomando la esperanza condicional de la expresión anterior (con respecto a Z) condicional a los datos y los parámetros *actuales*, para luego maximizar esta expresión c.r.a. θ y comenzar nuevamente. Específicamente, como la expresión anterior es lineal en z_i basta con

tomar su esperanza:

$$\begin{aligned}\mathbb{E}_\theta\left(Z_i|\theta_t, x_i\right) &= 1\cdot\mathbb{P}\left(Z_i=1|\theta_t,x_i\right)+0\cdot\mathbb{P}\left(Z_i=0|\theta_t,x_i\right)\\ &= \frac{\mathbb{P}\left(x_i|\theta_t,z_i=1\right)p(z_i=1)}{p(x_i|\theta)}\\ &= \frac{\mathbb{P}\left(x_i|\theta_t,z_i=1\right)p(z_i=1)}{p(x_i|z=1,\theta)p(z=1)+p(x_i|z=0,\theta)p(z=0)}\end{aligned}$$

Clase 10: 5/9

2.6. Propiedades del EMV

2.6.1. Consistencia

La primera propiedad que veremos del EMV es su consistencia. Que un estimador sea *consistente* quiere decir que éste tiende (de alguna forma) al parámetro real a medida vamos considerando más datos. Para verificar esto, primero definamos la divergencia de Kullback-Leibler entre las densidades f y g como

$$\mathrm{KL}\left(f\|g\right)=\int_{\mathcal{X}} f(x)\log\left(\frac{f(x)}{g(x)}\right)\mathrm{d}x \tag{2.34}$$

Observemos que la divergencia KL es siempre positiva $\forall f,g$ (desigualdad de Gibbs):

$$\begin{aligned}-\mathrm{KL}\left(f\|g\right) &= \int_{\mathcal{X}} f(x)\log\left(\frac{g(x)}{f(x)}\right)\mathrm{d}x\\ &\leq \log\left(\int_{\mathcal{X}} f(x)\frac{g(x)}{f(x)}\right)\mathrm{d}x, \qquad \text{(Jensen's)}\\ &= \log\left(\int_{\mathcal{X}} g(x)\right)\mathrm{d}x\\ &= \log 1 = 0\end{aligned}$$

y como log es estrictamente convexo, la igualdad solo se cumple si el argumento $\frac{g(x)}{f(x)}$ es constante, lo cual se tiene solo para $g(x)=f(x)$.

Otra propiedad clave de la divergencia KL es que puede ser infinita y es asimétrica. La intuición detrás de la KL es que es una medida de *error* de estimar una distribución con respecto a la otra.

Con la KL, definiremos que un modelo/parámetro es **identificable** si los valores para los parámetros $\theta \neq \theta'$ implican $\mathrm{KL}\left(p_\theta \| p\theta'\right) > 0$ lo que significa que distintos valores del parámetro dan origen a distintos modelos—asumiremos desde ahora que los modelos considerados son identificables.

Observemos que el estimador de MV puede ser obtenido de la maximización de

$$M_n(\theta') = n^{-1}(l_n(\theta') - l_n(\theta)) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta'}(x_i)}{p_\theta(x_i)}\right) \tag{2.35}$$

donde n es la cantidad de observaciones $x_1, \ldots, x_n$, θ es el parámetro real y $l_n(\cdot)$ es la log-verosimilitud. Esto es posible porque $l_n(\theta)$ es constante para θ'. Veamos que gracias a la ley de los grandes números, tenemos que

$$M_n(\theta') \to \mathbb{E}_\theta\left(\log\left(\frac{p_{\theta'}(x)}{p_\theta(x)}\right)\right) = -\mathbb{E}_\theta\left(\log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right)\right) = -\mathrm{KL}\left(p_\theta \| p_{\theta'}\right) \tag{2.36}$$

Consecuentemente, como el objetivo del estimador de MV tiende a la KL negativa, entonces maximizar la verosimilitud es equivalente a minimizar la KL-divergencia entre el modelo real y el modelo generado por el parámetro. Si el modelo generado tiende al modelo real, nuestro supuesto de *identificabilidad* implica que el estimador de MV tiene al parámetro real.

clase 11: 10/9

Otra propiedad muy utilizada en la práctica es el **Principio de equivarianza**, el cual establece que si θ_{MV} es el estimador de MV de θ, entonces, $g(\theta_{\mathrm{MV}})$ es el estimador de MV del parámetro transformado $g(\theta)$.

Ejemplo 2.6.1. *(Cálculo del EMV en Gaussiana: varianza versus precisión versus log-precisión versus cholesky - reparametrisation trick)*

2.6.2. Normalidad asintótica

Otra propiedad es la **normalidad asintótica del EMV**, esto significa que la distribución del estimador ML (como cantidad aleatoria) es normal a medida se

incluyen más observaciones. Para estudiar esta propiedad, primero definamos la función de puntaje o *score function* como **función aleatoria** definida por la derivada de la log-verosimilitud, es decir,

$$S_\theta(X) = \frac{\partial \log p_\theta(X)}{\partial \theta}. \tag{2.37}$$

Observemos que la esperanza de la función de puntaje cero. En efecto, derivando la igualdad fundamental $1 = \int_{\mathcal{X}} p_\theta(x)\mathrm{d}x$ con respecto a θ, obtenemos

$$0 = \int_{\mathcal{X}} \frac{\partial p_\theta(X)}{\partial \theta}\mathrm{d}x = \int_{\mathcal{X}} \frac{1}{p_\theta(X)}\frac{\partial p_\theta(X)}{\partial \theta} p_\theta(X)\mathrm{d}x = \int_{\mathcal{X}} \frac{\partial \log p_\theta(X)}{\partial \theta} p_\theta(X)\mathrm{d}x = \mathbb{E}_\theta\left(S_\theta(X)\right) \tag{2.38}$$

Sorprendente.

Ahora, derivemos una vez más con lo cual obtenemos:

$$\begin{aligned}
0 &= \int_{\mathcal{X}} \frac{\partial}{\partial \theta}\left(\frac{\partial \log p_\theta(X)}{\partial \theta} p_\theta(X)\right)\mathrm{d}x \\
&= \int_{\mathcal{X}} \left(\frac{\partial^2 \log p_\theta(X)}{\partial \theta^2} p_\theta(X) + \frac{\partial \log p_\theta(X)}{\partial \theta}\frac{\partial p_\theta(X)}{\partial \theta}\right)\mathrm{d}x \\
&= \mathbb{E}_\theta\left(\frac{\partial^2 \log p_\theta(X)}{\partial \theta^2}\right) + \mathbb{E}_\theta\left(\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)^2\right)
\end{aligned}$$

donde podemos observar que cada una de estos términos tiene la misma magnitud (uno es negativo y el otro es positivo). Esta magnitud es conocida como *información de Fisher*, denotada como

$$I(\theta) = \mathbb{E}_\theta\left(\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)^2\right) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log p_\theta(X)}{\partial \theta^2}\right). \tag{2.39}$$

Recordemos además que como la esperanza de la función de puntaje es nulo, entonces, su varianza puede ser expresada como

$$\mathbb{V}_\theta\left(S_\theta(X)\right) = \mathbb{E}_\theta\left(S_\theta(X)^2\right) - \cancel{\mathbb{E}_\theta\left(S_\theta(X)\right)^2} = \mathbb{E}_\theta\left(\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)^2\right) \tag{2.40}$$

consecuentemente, la información de Fisher también es la varianza de la función de pérdida. Ya contamos con tres expresiones para poder calcular esta cantidad.

Ejercicio 2.6.1 (Cálculo de la información de Fisher para Bernoulli)**.** *Consideremos* $X \sim Ber\,(\theta)$*, entonces,*

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log\left(\theta^X(1-\theta)^{1-X}\right)\right) \\ &= -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}X\log\theta + \frac{\partial^2}{\partial\theta^2}(1-X)\log\left(1-\theta\right)\right) \\ &= \mathbb{E}_\theta\left(X\theta^{-2} + (1-X)(1-\theta)^{-2}\right) \\ &= \theta^{-1} + (1-\theta)^{-1} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned} \tag{2.41}$$

Ejercicio 2.6.2 (Cálculo de la información de Fisher para Poisson)**.** *Consideremos* $X \sim Poisson\,(\theta)$*, entonces,*

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log\left(\frac{\theta^X e^{-\theta}}{X!}\right)\right)^2\right) \\ &= \mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}X\log\theta - \frac{\partial}{\partial\theta}\theta - \frac{\partial}{\partial\theta}\log(X!)\right)^2\right) \\ &= \mathbb{E}_\theta\left(\left(X\theta^{-1} - 1\right)^2\right) \\ &= \mathbb{E}_\theta\left(X^2\theta^{-2} - 2X\theta^{-1} + 1\right) \\ &= (\theta+\theta^2)\theta^{-2} - 2\theta\theta^{-1} + 1 \\ &= \theta^{-1} \end{aligned}$$

El cálculo anterior ha sido para la verosimilitud en base a una variable aleatoria, si consideramos la verosimilitud evaluada calculada para un conjunto de observaciones (IID), tenemos que

$$S_\theta(X_1,\ldots,X_n) = \frac{\partial\log\prod_{i=1}^n p_\theta(X_i)}{\partial\theta} = \sum_{i=1}^n \frac{\partial\log p_\theta(X_i)}{\partial\theta} = \sum_{i=1}^n S_\theta(X_i) \tag{2.42}$$

y de igual forma para la información de Fisher, tenemos,

$$I_n(\theta) = \mathbb{V}_\theta\left(\sum_{i=1}^{n} S_\theta(X_i)\right) = nI(\theta) \tag{2.43}$$

Veamos ahora una desigualdad interesante para la información de Fisher y su relación con estimadores. Consideremos un estimador insesgado, es decir,

$$\mathbb{E}_\theta\left(\hat{\theta}(X) - \theta\right) = \int_{\mathcal{X}} (\hat{\theta}(X) - \theta) p_\theta(X) \mathrm{d}x = 0. \tag{2.44}$$

Derivando esta expresión con respecto a θ, obtenemos

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \int_{\mathcal{X}} (\hat{\theta}(X) - \theta) p_\theta(X) \mathrm{d}x \\ &= -\int_{\mathcal{X}} p_\theta(X) \mathrm{d}x + \int_{\mathcal{X}} (\hat{\theta}(X) - \theta) \frac{\partial p_\theta(X)}{\partial\theta} \mathrm{d}x \\ &= -1 + \int_{\mathcal{X}} (\hat{\theta}(X) - \theta) \frac{\partial \log p_\theta(X)}{\partial\theta} p_\theta(X) \mathrm{d}x. \end{aligned}$$

Lo que implica que

$$\begin{aligned} 1 &= \left(\int_{\mathcal{X}} (\hat{\theta}(X) - \theta) \frac{\partial \log p_\theta(X)}{\partial\theta} p_\theta(X) \mathrm{d}x\right)^2 \\ &= \left(\int_{\mathcal{X}} (\hat{\theta}(X) - \theta) \sqrt{p_\theta(X)} \sqrt{p_\theta(X)} \frac{\partial \log p_\theta(X)}{\partial\theta} \mathrm{d}x\right)^2 \\ &\leq \int_{\mathcal{X}} (\hat{\theta}(X) - \theta)^2 p_\theta(X) \mathrm{d}x \int \left(\frac{\partial \log p_\theta(X)}{\partial\theta}\right)^2 p_\theta(X) \mathrm{d}x. \end{aligned}$$

Notemos que la primera integral es la varianza del estimador insesgado $\hat{\theta}$ y la segunda es la esperanza del cuadrado de la función de puntaje (o la información de Fisher). Con esto, podemos enunciar el siguiente resultado

Definición 2.6.1 (Cota de Cramer-Rao)**.** *Sea $X_1, \ldots, X_n \sim p_\theta$ y $nI(\theta)$ su información de Fisher. Entonces para todo estimador insesgado θ' tenemos*

$$\mathbb{V}_\theta\left(\theta'\right) \geq (nI(\theta))^{-1}, \quad \forall\theta \in \Theta \tag{2.45}$$

La cota de Cramer-Rao es un elemento fundamental en el estudio estadístico, pues establece que cualquier estimador tiene necesariamente una varianza que está por sobre el recíproco de la información de Fisher.

Ahora podemos finalmente volver al concepto de normalidad asintótica. Si tenemos una colección de VA $X_1, \ldots, X_n \sim p_\theta$ con θ el parámetro real, entonces, la secuencia de estimadores de MV, ${\theta_{\text{MV}}}^{(n)}$ cumple con

$$\sqrt{n}({\theta_{\text{MV}}}^{(n)} - \theta) \to \mathcal{N}(0, I(\theta)^{-1}) \tag{2.46}$$

lo cual intuitivamente corresponde a que, para n suficientemente grande, el estimador de MV está distribuido de forma normal en torno al parámetro real con varianza $(nI(\theta))^{-1}$. Lo cual implica también *eficiencia asintótica*.

Este resultado es fundamental en estadística aplicada: no importa cómo ha sido obtenido el estimador de MV, si n es suficientemente grande, entonces la distribución del estimador es normal y su varianza tiende a cero.

clase 12: 12/9

2.7. Intervalos de Confianza

En muchas aplicaciones, no es suficiente reportar una estimación puntual, es decir, un valor único para le parámetro a estimar, sino que debe identificarse una rango donde, con cierta probabilidad, el parámetro real está contenido. Esto motiva la siguiente definición:

Definición 2.7.1 (Intervalo de confianza)**.** *Un $(1-\alpha)$-intervalo de confianza para el parámetro θ, $\alpha \in [0,1]$, es el intervalo aleatorio $(A(X), B(X))$ tal que*

$$\mathbb{P}_\theta\left(A(X) \leq \theta \leq B(X)\right) = 1 - \alpha, \forall \theta \in \Theta. \tag{2.47}$$

Observación 2.7.1. *Es importante notara que la definición arriba no es un enunciado de probabilidad en θ, o en otras palabras, no describe una probabilidad sobre el parámetro θ; pues recordemos que éste es fijo. Por el contrario, lo que es aleatorio en la ecuación* (2.47) *es el intervalo, no el parámetros. Entonces, si bien es una sutileza, la definición anterior se debe entender como la probabilidad de que "el intervalo (aleatorio) contenga al parámetro (fijo)", y no como la probabilidad de que 'el parámetro esté en el intervalo'.*

Una consecuencia clave de este concepto es que si $I_{1-\alpha}$ es un $(1-\alpha)$-intervalo de confianza, entonces si fuese posible repetir una gran cantidad de veces el ejercicio de recolectar datos X y calcular este intervalo para cada una de estas observaciones, entonces el parámetro θ estaría contenido en le intervalo un $(1-\alpha)\,\%$.

Esto es muy diferente de asegurar que para un solo experimento, la probabilidad de que el parámetro θ esté contenido en $I_{1-\alpha}$ es $1-\alpha$, lo cual no es cierto. Los siguientes ejemplos tienen por objetivo ayudar a aclarar este concepto.

Ejemplo 2.7.1 (Intervalo de confianza para la media de la distribución normal)**.** *Consideremos la muestra* $X_1, \ldots, X_\nu \sim \mathcal{N}(\theta, 1)$. *Como* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\theta, 1/n)$ *tenemos que*

$$\sqrt{n}(\bar{X} - \theta) \sim \mathcal{N}(0, 1). \tag{2.48}$$

Esta cantidad se llama pivote *y es una función (de la VA y del parámetro) cuya distribución no depende del parámetro. Consecuentemente, podemos identificar directamente un intervalo de confianza para el pivote desde una tabla de valores para la distribución normal de media cero y varianza unitaria. Si* $\phi(x)$ *denota la distribución Normal, entonces podemos elegir* x_1 *y* x_2 *tal que* $\phi(x_2) - \phi(x_1) = 1 - \alpha$ *con lo que tenemos*

$$\mathbb{P}_\theta\left(x_1 \leq \sqrt{n}(\theta - \bar{X}) \leq x_2\right) = 1-\alpha \Leftrightarrow \mathbb{P}_\theta\left(\bar{X} + x_1/\sqrt{n} \leq \theta \leq \bar{X} + x_2/\sqrt{n}\right) = 1-\alpha, \tag{2.49}$$

es. decir, $(\bar{X} + x_1/\sqrt{n}, \bar{X} + x_2/\sqrt{n})$ *es el* $(1-\alpha)$*-intervalo de confianza de* θ. *Eligiendo* $\alpha = 0{,}05$ *tenemos* $x_2 = -x_1 = 1{,}96$ *con lo que el intervalo de confianza del 95 % para* θ *está dado por*

$$(\bar{X} - 1{,}96/\sqrt{n}, \bar{X} + 1{,}96/\sqrt{n}). \tag{2.50}$$

El procedimiento estándar para encontrar intervalos de confianza es como el ilustrado en el ejemplo anterior, en donde construimos una cantidad que tiene una distribución que no depende del parámetro desconocido (llamada pivote). Construir un intervalo de confianza para esta cantidad es directo desde las tablas de distribuciones, luego, podemos encontrar el intervalo de confianza para la cantidad deseada, e.g., el parámetro desconocido, mediante transformaciones de la expresión del pivote.

Observación 2.7.2. *El intervalo de confianza no es único. Por ejemplo, en el caso gaussiano podemos elegir en intervalo centrado en cero o desde* $-\infty$. *Esta elección dependerá de las aplicación en cuestión: una regla general es elegirlo de forma centrada para densidades que son simétricas, centrado en la moda para distribuciones unimodales, mientas que para densidades con soporte positivo podemos elegirlo desde cero.*

Hasta ahora hemos solo definido intervalos de confianza para cantidades escalares, en donde el concepto de intervalo tiene sentido. Para parámetros vectoriales, nos referiremos a conjuntos de confianza. Siguiendo la Definición 2.7.1, un $(1-\alpha)$*-conjunto de confianza* $S(X)$ *es tal que*

$$\mathbb{P}_\theta\left(\theta \in S(X)\right) = 1-\alpha, \forall \theta \in \Theta. \tag{2.51}$$

Ejercicio 2.7.1. *Considere $X_1, \ldots, X_{50} \sim \mathcal{N}(0, \sigma^2)$, calcule el intervalo de confianza del 99 %.*

Ejemplo 2.7.2 (Intervalo de confianza para Bernoulli)**.** *Consideremos $X_1, \ldots, X_n \sim Ber(\theta)$ y calculemos un intervalo de confianza para θ. Recordemos que el EMV es $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y debido a la normalidad asintótica del EMV, tenemos que para n grande, podemos asumir*

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right) \tag{2.52}$$

donde la varianza $\frac{\theta(1-\theta)}{n} = I_n(\theta)^{-1}$ es la inversa de la información de Fisher.

Podemos entonces considerar el pivote

$$\frac{\sqrt{n}(\hat{\theta}-\theta)}{\sqrt{\theta(1-\theta)}} \sim \mathcal{N}(0,1) \tag{2.53}$$

y calcular el $(1-\alpha)$-intervalo de confianza asumiendo los valores x_1 y x_2 mediante

$$\mathbb{P}_\theta\left(x_1 \leq \frac{\sqrt{n}(\theta-\hat{\theta})}{\sqrt{\theta(1-\theta)}} \leq x_2\right) = 1-\alpha \Leftrightarrow \mathbb{P}_\theta\left(\hat{\theta} + \frac{x_1\sqrt{\theta(1-\theta)}}{\sqrt{n}} \leq \theta \leq \hat{\theta} + \frac{x_2\sqrt{\theta(1-\theta)}}{\sqrt{n}}\right) = 1-\alpha.$$

Sin embargo, los bordes de este intervalo no son conocidos, pues dependen de θ. Una forma de aproximar el intervalo es reemplazar el parámetro por su EMV.

Ejercicio 2.7.2 (Encuesta de elecciones presidenciales)**.** *Considere una encuesta que ha consultado a 1000 votantes y su candidato ha recibido 200 votos. Use el resultado del ejemplo anterior para determinar el intervalo de confianza del 95 % de la cantidad de votos que su candidato obtendría en la elección presidencial.*

Finalmente, revisaremos el siguiente ejemplo, el cual pretende ejemplificar el concepto de que en solo un experimento, la determinación del $(1-\alpha)$-intervalo de confianza no quiere decir que la probabilidad de que el parámetro esté contenido en él es $(1-\alpha)$ %.

Ejemplo 2.7.3 (Intervalo de confianza para una distribución uniforme)**.** *Considere $X_1, X_2 \sim$ Uniforme $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ y observe que*

$$\begin{aligned}\mathbb{P}_\theta\left(\text{mín}(X_1, X_2) \leq \theta \leq \text{máx}(X_1, X_2)\right) &= \mathbb{P}_\theta\left(X_1 \leq \theta \leq X_2\right) + \mathbb{P}_\theta\left(X_2 \leq \theta \leq X_1\right) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \\ &= \frac{1}{2}\end{aligned}$$

corresponde al intervalo del 50 %.

Sin embargo, si observáramos $X_1 = x_1$ *y* $X_2 = x_2$ *tal que* $|x_1 - x_2| \geq \frac{1}{2}$ *entonces necesariamente está contenido en el intervalo* $(\text{mín}(X_1, X_2), \text{máx}(X_1, X_2))$ *con probabilidad 1. Esto ilustra la idea de que, en un experimento dado, la probabilidad de que* θ *esté en intervalo de confianza del* $(1 - \alpha)$ *% no es necesariamente* $(1 - \alpha)$ *%.*

Capítulo 3

Test de Hipótesis

Clase 13: 24/9

3.1. Teoría de decisiones

En términos generales, la teoría de decisiones estudia las acciones que puede tomar un agente en un escenario dado. En este contexto afloran de forma natural los conceptos de incertibumbre (de aspectos clave del escenario), funciones de pérdida y procedimientos de decisión. En estadística, podemos identificar al menos los siguientes problemas de decisión.

- **Estimación**: Decidir el valor apropiado para un parámetro desconocido usando datos X y un distribución condicional P_θ
- **Test**: Decidir la hipótesis correcta usando datos $X \sim P_\theta$

$$H_0 : P_\theta \in \mathcal{P}_0 \tag{3.1}$$
$$H_1 : P_\theta \notin \mathcal{P}_0 \tag{3.2}$$

- **Ranking**: Elaborar una lista ordenada de ítems, por ejemplo, productos evaluados por una muestra de la población, resultados de eventos deportivos o juegos online.
- **Predicción**: Estimar/decidir el valor de una variable dependiente en base a observaciones de observaciones pasadas.

Como se puede apreciar, la teoría de decisiones presenta un contexto general para abordar una gran cantidad de situaciones. A continuación se describen los elementos básicos de un problema de decisión, en donde, con fines ilustrativos, ponemos como ejemplo su contraparte en el problema de estimación.

- $\Theta = \{\theta\}$ es el espacio de estado, donde la cantidad θ es el *estado del mundo*. En el problema de estimación, donde convenientemente se ha usado la misma notación, θ es el parámetro del modelo
- $\mathcal{A} = \{a\}$ es el espacio de acciones, donde a es la acción a tomar por el estadístico. En estimación, podemos abusar de la notación y decir que la acción a es elegir el valor a para el parámetro θ.
- $L(\theta, a)$ es la función de pérdida asociada a tomar la decisión a cuando el estado es θ; nótese que $L : \Theta \times \mathcal{A} \to \mathbb{R}$. En el caso de estimación, usualmente consideramos la pérdida cuadrática:

$$L(\theta, a) = (\theta - a)^2. \tag{3.3}$$

Ejemplo 3.1.1. *(Inversión bajo incertidumbre) Consideremos los estados $\Theta = \{\theta_1, \theta_2\}$, donde θ_1 quiere decir mercado sano y θ_2 quiere decir mercado no sano. Se debe elegir una estrategia de inversión del siguiente conjunto $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ con la siguiente función de costo $L(\theta, a)$ Notemos que*

$L(\theta, a)$	a_1	a_2	a_3	a_4	a_5
θ_1	-4	-4	-1	2	4
θ_2	4	0	-1	-6	-4

- *a_1 es bueno cuando $\theta = \theta_1$*
- *a_2 es bueno cuando $\theta = \theta_2$*
- *a_3 es medianamente bueno (pérdida negativa) siempre*

entonces, ¿cómo elegimos la acción?

Además de los elementos básicos del problema de decisión (estado, acciones y pérdida), en el enfoque estadístico de la teoría de decisiones existen los siguientes elementos:

- $X \sim P_\theta$ es la variable aleatoria, la cual define la distribución condicional, el espacio muestral, la densidad, etc.
- $\delta(X)$ es el procedimiento de la decisión, es decir, el mapa que asocia una observación $X = x$ con la acción a:

$$\delta(\cdot) : \mathcal{X} \to \mathcal{A}. \tag{3.4}$$

- $\mathcal{D} = \{\delta : \mathcal{X} \to \mathcal{A}\}$ es el espacio de decisiones
- $R(\theta, \delta)$ es el riesgo asociado a δ y θ, el cual está definido como el valor esperado de la pérdida incurrida al tomar la acción $\delta(X)$ cuando el parámetro es θ. Es decir,

$$R(\theta, \delta) = \mathbb{E}_\theta \left(L(\theta, \delta(X)) \right). \tag{3.5}$$

Ejemplo 3.1.2. *Volviendo al contexto del problema de estimación, consideremos el caso en que usando una VA $X \sim \mathcal{N}(\theta, 1), \theta \in \mathbb{R}$ debemos encontrar el valor de θ. En el contexto de teoría estadística de decisiones, el espacio de posibles acciones es precisamente en espacio de parámetros, es decir,*

$$\mathcal{A} = \Theta = \mathbb{R}. \tag{3.6}$$

Elegimos además la pérdida cuadrática, $L(\theta, \hat{\theta}(X)) = (\theta - \hat{\theta}(X))^2$, asociada a elegir $\hat{\theta}(X)$ como el valor del parámetro θ. Consideremos que el espacio de acciones está dado por versiones escaladas de la observación X, es decir,

$$\mathcal{A} = \{cX, c \in [0, 1]\}. \tag{3.7}$$

Con esta forma del estimador, podemos calcular el riesgo asociado mediante

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \left((\hat{\theta}(X) - \theta)^2 \right) = \mathbb{V}_\theta \left(\hat{\theta}(X) \right) + \mathbb{E}_\theta \left(\hat{\theta}(X) - \theta \right)^2 = c^2 + (c-1)^2 \theta^2 \tag{3.8}$$

¿Qué valor de c sugiere elegir?

3.2. Intuición en un test de hipótesis

El objetivo del análisis estadístico es obtener conclusiones razonables mediante el uso de observaciones, como también aseveraciones precisas sobre la incertidumbre asociada a dichas conclusiones. De forma ilustrativa, consideremos el siguiente escenario hipotético.

En base a estudios preliminares, se sabe que los pesos de los recién nacidos (RN) en Santiago, Chile, distribuyen aproximadamente normal con promedio 3000gr y desviación estándar de 500gr. Creemos que los RNs en Osorno pesan, en promedio, más que los RNs en Santiago. Nos gustaría formalmente aceptar o rechazar esta hipótesis.

Intuitivamente, una forma de evaluar esta hipótesis es tomar una muestra de RNs en Osorno, calcular su peso promedio y verificar si éste es *significativamente mayor* que 3000gr. Asumamos que hemos tenido acceso al peso de 50 RNs nacidos en Osorno, los cuales exhiben un preso promedio de 3200gr. ¿Podemos entonces concluir directamente y decir que efectivamente los RNs de Osorno pesan más que los de Santiago? Si bien esta es una posibilidad, una postura más escéptica podría argumentar que el obtener una población de 50 RNs con peso promedio de 3200gr es perfectamente plausible de una población de RNs distribuidos de acuerdo a $\mathcal{N}(3000, 500^2)$. Entonces, ¿cómo justificamos la plausibilidad de este resultado?

Para esto distingamos entre las dos hipótesis:

- H_1: Los RNs en Osorno pesan en promedio más de 3000gr (esta es la hipótesis alternativa)
- H_0: Los RNs en Osorno pesan en promedio 3000gr (esta es la hipótesis nula)

Para decidir cuál es verdadera, trataremos de *falsificar* H_0. La forma de hacer esto es calcular la probabilidad de obtener el resultado observado bajo el supuesto que H_0 es cierta. En este caso, sabemos que una muestra

$$X = X_1, \ldots, X_{50} \sim \mathcal{N}(3000, 500^2), \tag{3.9}$$

tiene una media que está distribuida de acuerdo a la siguiente densidad

$$\bar{X} = \frac{1}{50} \sum_{i=1}^{50} X_i \sim \mathcal{N}(3000, 500^2/50). \tag{3.10}$$

Entonces, cuál es la probabilidad de que la muestra obtenida, $\bar{X} = 3200$, haya sido generada por la distribución anterior? Para calcular esto, construyamos el **pivote**

$$z = \frac{\bar{X} - 3000}{500/\sqrt{50}} \sim \mathcal{N}(0, 1) \tag{3.11}$$

con la cual podemos realizar el cálculo:

$$\mathbb{P}\left(\bar{X} \geq 3200\right) = \mathbb{P}\left(z = \frac{\bar{X} - 3000}{500/\sqrt{50}} \geq 2\sqrt{2}\right) = 0{,}0023388674905235884, \quad (3.12)$$

donde el valor de esta probabilidad puede ser calculado usando la función[1] `cdf` de SciPy mediante la siguiente instrucción.

```
from scipy.stats import norm
import numpy as np
print(1-norm.cdf(2*np.sqrt(2)))
```

Concluimos entonces que la probabilidad de que una muestra de 50 RNs exhiban un promedio de peso mayor o igual a 3200gr, bajo el supuesto que H_0 es cierta, es del orden del 0.23 %.

Nos referiremos a esta probabilidad como **p-valor**, el cual nos dice cuán verosímil es que obtener la observación dada bajo el supuesto que la hipótesis nula es cierta. Mientras más pequeño es el p-valor, entonces más fuerte es la evidencia en contra de H_0. Entonces nos encontramos ante dos posibles explicaciones:

- H_0 es falsa
- hemos obtenido un resultado que solo ocurre una de cada 500 veces.

Agregar comentario sobre doble negación y falsificación

En cuando al umbral en el cual rechazamos H_0, nos referiremos a significancia del test α al umbral para el p-valor en el cual se rechaza el test. En general, este umbral es del 1 % o del 5 %, sin embargo esto depende de la aplicación en cuestión. Por ejemplo, si estamos considerando la administración de una droga que puede tener consecuencias fatales, entonces necesariamente nuestro nivel de significancia sea muy bajo, lo que quiere decir que la hipótesis nula requiere mucha evidencia en su contra para ser rechazada. Notemos que hay dos tipos de errores: El error de Tipo I en el cual H_0 es rechazada a pesar de que es verdadera y el error de Tipo II, donde H_0 no es rechazada a pesar de que es falsa (lo cual diremos que tiene probabilidad β). Los tipos de errores se definen mediante la siguiente tabla:

	H_0 es cierto	H_0 no es cierto
se rechaza H_0	error Tipo I	no hay error
no se rechaza H_0	no hay error	error Tipo II

[1] Acrónimo de *cumulative denstiy function*.

Volviendo a nuestro ejemplo de los recién nacidos, el p-valor del test es del orden de 0.0023, lo cual, si consideramos una significancia del $\alpha = 0{,}01 = 1\,\%$, resulta en el rechazo de H_0. Decimos entonces que **hay suficiente evidencia para rechazar H_0 al 1 %**, o bien que **rechazamos la hipótesis nula H_0 al 1 %**. Por el contrario, en el caso que el p-valor fuese mayor que el nivel de significancia del test, entonces no rechazamos H_0 y simplemente decimos que **la evidencia para rechazar H_0 no es significativa al 1 %**

clase 14: 26/9

Test de Hipótesis

En resumen, un test de hipótesis consta de los siguientes pasos:

1. Proponer una hipótesis alternativa H_1
2. construir una hipótesis nula (básicamente lo contrario de H_0)
3. Recolectar datos
4. Calcular el pivote (un estadístico de prueba) usando los datos
5. Calcular el p-valor para el pivote
6. Comparar el p-valor con la significancia estadística.
7. Rechazar si corresponde

Sobre p-valor y región crítica. Otra forma de cuantificar la evidencia en contra de H_0 es mediante la identificación de una región crítica, es decir, un subconjunto de $\mathcal{X}$ en donde, de tomar valores la observación (o el estadístico) , su p-valor estaría por debajo del nivel de significancia y consecuentemente H_0 se rechazaría. En el ejemplo anterior, este puede ser calculado usando la función de SciPy `ppf`[2]. Considerando una significancia del 1 % podemos ejecutar

```
1 from scipy.stats import norm
2 print(norm.ppf(0.99))
```

lo cual nos da una región crítica $[2{,}326, \infty)$, la cual contiene a nuestro p-valor $2\sqrt{2} = 2{,}82$; concluimos de igual forma y rechazamos H_0 al 1 %.

[2] Acrónimo para *percent point function*.

Falta agregar gráfico ilustrando el uso de p-valor, pivote, significancia y región crítica.

Falta discusión sobre test simétricos y asimétricos: gráfico ilustrando el uso de p-valor, pivote, significancia y región crítica.

3.3. Rechazo, potencia y nivel

Formalmente, frente a dos hipótesis generales denotadas por

$$H_0 : \theta \in \Theta_0 \tag{3.13}$$
$$H_1 : \theta \in \Theta_1, \tag{3.14}$$

definiremos el problema del test de hipótesis como la búsqueda de una funció

$$\phi : \mathcal{X} \to \{0, 1\}, \tag{3.15}$$

donde:

- Si $\phi(x) = 0$ entonces aceptamos H_0 (no rechazamos H_0).
- Si $\phi(x) = 1$ entonces rechazamos H_0, lo cual implícitamente acepta H_1.

En teoría de decisiones, diríamos que ϕ es una regla de decisión.

A continuación, revisamos definiciones que serán de utilidad para analizar y construir tests.

Definición 3.3.1 (Región crítica de un test)**.** *La región crítica o región de rechazo de un test de hipótesis ϕ se define como*

$$R_\phi = \{x \in \mathcal{X} | \phi(x) = 1\} = \phi^{-1}(1). \tag{3.16}$$

Definición 3.3.2 (Función de probabilidad de rechazo)**.** *Para un test ϕ y cualquier parámetro $\theta \in \Theta$ podemos definir la probabilidad de rechazo mediante*

$$\alpha_\phi(\theta) = \mathbb{P}_\theta\left(\phi(x) = 1\right) = \mathbb{P}_\theta\left(x \in R_\phi\right), \forall \theta \in \Theta, \tag{3.17}$$

donde nos gustaría entonces que $\alpha \approx 0$ si $\theta \in \Theta_0$ es cierto y que $\alpha \approx 1$ si $\theta \in \Theta_1$. Luego, usaremos esta función para evaluar la calidad del test.

Definición 3.3.3 (Potencia de un test)**.** *En el caso que H_1 sea cierta, es decir, $\theta \in \Theta_1$, podemos definir la potencia del test como la probabilidad rechazar H_0 cuando H_1 es efectivamente cierta ($\theta \in \Theta_1$). Es decir,*

$$\pi_\phi(\theta) = \mathbb{P}\left(\textit{rechazar } H_0 | H_1 \textit{ es cierta}\right) = \mathbb{P}_{\theta_1}\left(\phi(x) = 1\right) \tag{3.18}$$

Nos gustaría entonces minimizar $\alpha(\theta)$ cuando H_0 y maximizar $\alpha(\theta)$ cuando H_1, lo cual es equivalente a minimizar la probabilidad de cometer errores de Tipo I y II respectivamente.

Ejemplo 3.3.1 (Un test absurdo)**.** *Existen tests absurdos, por ejemplo $\phi(x) = 0, \forall x \in \mathcal{X}$. Este test tiene $\alpha(\theta) = 0$ cuando H_0 (lo cual es bueno), por también tiene potencia nula, es decir, incluso si H_1, no rechaza a H_0.*

En general, consideramos más importante prevenir un error de tipo I que uno de tipo II.

Definición 3.3.4 (Nivel de un test)**.** *Decimos que un test es de nivel $\alpha \in [0, 1]$ si*

$$\alpha_\phi(\theta) \leq \alpha, \forall \theta \in \Theta_0, \tag{3.19}$$

equivalentemente, $\sup_{\theta \in \Theta_0} \alpha_\phi(\theta) \leq \alpha$. Además, denotamos por T_α la clase de todos los tests de nivel α.

Dentro de esta clase, la cual nos restringe únicamente a los test que tienen probabilidad de rechazo acotada superiormente por α para $\theta \in \Theta_0$ (probabilidad de cometer error tipo I), podemos buscar el test de mayor potencia (probabilidad de rechazar H_0 cuando H_1 es cierta). Caracterizamos este test mediante:

Definición 3.3.5 (Test uniformemente más potente, UMP)**.** *Diremos que $\phi^\star$ es un test UMP (de nivel α) si*

$$\pi_{\phi^\star}(\theta) \geq \pi_\phi(\theta), \forall \theta \in \Theta_1. \tag{3.20}$$

Falta definición/discusión sobre tests simples y compuestos.

3.4. Test de Neyman-Pearson

Consideremos el siguiente problema de test de dos hipótesis simples.

$$H_0 : \theta \in \Theta_0 = \{\theta_0\} \quad \text{v.s.} \quad H_1 : \theta \in \Theta_1 = \{\theta_1\}, \tag{3.21}$$

donde por una notación más simple escribiremos simplemente

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta_1, \tag{3.22}$$

y asumiremos que $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\} = \{P_{\theta_0}, P_{\theta_1}\}$ con densidades respectivamente dadas Por $p_0(x) = p_{\theta_0}(x)$ y $p_1(x) = p_{\theta_1}(x)$.

Denotamos además la región crítica (donde se rechaza H_0) mediante

$$R^* = \{x \in \mathcal{X} | p_1(x) \geq k p_0(x)\}, \tag{3.23}$$

donde $k \in \mathbb{R}_+$ es una constante a determinar.

Podemos entonces definir el test ϕ^* como el test que tiene el conjunto R^* como región de rechazo, es decir,

$$\phi^*(x) = 1 \Leftrightarrow x \in R^*. \tag{3.24}$$

Finalmente, determinaremos la constante k de tal manera de que

$$\alpha_{\phi^\star}(\theta_0) = \mathbb{P}_{\theta_0}(x \in R^*) = \alpha, \ \alpha \in [0,1], \tag{3.25}$$

donde, por definición, $\phi^\star \in T_\alpha$. Consecuentemente, de acuerdo al siguiente lema, $\phi^\star$ es el test UMP en T_α.

Lema 3.4.1 (Neyman-Pearson)**.** *Consideremos un test de hipótesis de la forma*

$$H_0 : \theta = \theta_0 \quad \textit{v.s.} \quad H_1 : \theta = \theta_1 \tag{3.26}$$

con probabilidad de rechazo dada por

$$\alpha_\phi(\theta) = \mathbb{P}(p_1 \geq k p_0). \tag{3.27}$$

Entonces, el test definido arriba que rechaza H_0 si $p_1 \geq k p_0$ con $\alpha_\phi(\theta) = \alpha, \theta \in \Theta_0$ es el UMP.

Demostración. Consideremos un test $\phi \in T_\alpha$ con región crítica dada por R. Recordemos que la probabilidad de los datos estén en la región R es

$$\mathbb{P}_\theta(R) = \int_R p_\theta(x)\mathrm{d}x. \tag{3.28}$$

Luego, podemos escribir

$$\mathbb{P}_\theta(R) = \mathbb{P}_\theta(R \cap R^*) + \mathbb{P}_\theta(R \cap \bar{R}^*) \tag{3.29}$$
$$\mathbb{P}_\theta(R^*) = \mathbb{P}_\theta(R^* \cap R) + \mathbb{P}_\theta(R^* \cap \bar{R}), \tag{3.30}$$

restando y evaluando para $\theta = \theta_1$, tenemos

$$\begin{aligned}
\mathbb{P}_{\theta_1}(R^*) - \mathbb{P}_{\theta_1}(R) &= \mathbb{P}_{\theta_1}(R^* \cap \bar{R}) - \mathbb{P}_{\theta_1}(R \cap \bar{R}^*) \\
&= \int_{R^* \cap \bar{R}} p_{\theta_1}(x)\mathrm{d}x - \int_{R \cap \bar{R}^*} p_{\theta_1}(x)\mathrm{d}x \\
&\geq k \int_{R^* \cap \bar{R}} p_{\theta_0}(x)\mathrm{d}x - k \int_{R \cap \bar{R}^*} p_{\theta_0}(x)\mathrm{d}x \quad [\text{pues } p_1 \geq k p_0 \text{ en } R^*] \\
&= k\left(\mathbb{P}_{\theta_0}(R^* \cap \bar{R}) - \mathbb{P}_{\theta_0}(R \cap \bar{R}^*)\right) \\
&= k\left(\underbrace{\mathbb{P}_{\theta_0}(R^*)}_{=\alpha} - \underbrace{\mathbb{P}_{\theta_0}(R)}_{\leq \alpha}\right) \quad [\text{primera igualdad de este desarrollo}] \\
&\geq 0
\end{aligned} \tag{3.31}$$

Hemos probado que $\mathbb{P}_{\theta_1}(R^*) \geq \mathbb{P}_{\theta_1}(R)$, es decir, si $\theta = \theta_1$ entonces R^* tiene mayor probabilidad que cualquier otra región, es es decir, el test que tiene a R^* por región es el test UMP.

□

clase 16: 3/10

Ejemplo 3.4.1. *Sea $X_1, \ldots, X_n$ iid Ber(θ), $\theta \in \{\theta_0, \theta_1\}$:*

$$H_0 : \theta = \theta_0 \quad v.s. \quad H_1 : \theta = \theta_1. \tag{3.32}$$

Asumamos que $\theta_1 > \theta_0$ y expresemos las densidades de cada hipótesis como

$$p_i(x) = \theta_i^{\sum x_j}(1-\theta_i)^{n-\sum x_j},\ i = 0, 1. \tag{3.33}$$

Para rechazar H_0 según el test de Neyman-Pearson, es decir, $x \in R^*$ de acuerdo a la ecuación (3.23)*, el test requiere:*

$$\frac{p_1(x)}{p_0(x)} = \frac{\theta_1^{\sum x_j}(1-\theta_1)^{n-\sum x_j}}{\theta_0^{\sum x_j}(1-\theta_0)^{n-\sum x_j}} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_j} \geq k, \tag{3.34}$$

donde, usando el hecho de que $\theta_1 \geq \theta_0$, podemos justificar que la expresión anterior es monótona en $\sum x_j$, consecuentemente, $\sum x_j$ debe ser lo suficientemente grande para rechazar H_0.

Ahora, para calcular el valor de k dado un α, tenemos que resolver $\mathbb{P}_{\theta_0}(x \in R^*) = \alpha$. Primero notemos que la ecuación (3.34) *es equivalente a*

$$\begin{aligned}\left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_j} \geq k &\Leftrightarrow \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_j} \geq k\left(\frac{1-\theta_0}{1-\theta_1}\right)^n \\ &\Leftrightarrow \sum x_j \log\left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right) \geq n\log\left(k\left(\frac{1-\theta_0}{1-\theta_1}\right)\right) \\ &\Leftrightarrow \sum x_j \geq \frac{n\log\left(k\left(\frac{1-\theta_0}{1-\theta_1}\right)\right)}{\log\left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)} = k'\end{aligned} \tag{3.35}$$

como $\sum x_j$ es binomial, podemos resolver directamente para k' (y consecuentemente para k).

3.5. Test de Wald

Este test nos permite evaluar si un parámetro θ toma no un valor θ_0 dado. Consideremos un parámetro escalar y $\hat{\theta}$ un estimador asintóticamente normal, es decir,

$$\frac{\hat{\theta} - \theta_0}{\text{ee}} \sim \mathcal{N}(0,1), \tag{3.36}$$

cuando el número de observaciones tiende a infinito y $\text{ee} = \sqrt{\mathbb{V}\left(\hat{\theta}\right)}$ es conocido como el *error estándar* y puede ser calculado muestralmente o con respecto a θ_0. Entonces, el test de Wald de tamaño α para las hipótesis

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta \neq \theta_0 \tag{3.37}$$

indica rechazar H_0 cuando el pivote $W = \frac{\hat{\theta}-\theta_0}{\text{ee}}$ cumple con

$$|W| \geq z_{\alpha/2}, \tag{3.38}$$

donde $z_{\alpha/2} = \Phi(1-\alpha/2)$, es decir, $\mathbb{P}\left(Z \geq z_{\alpha/2}\right) = \alpha/2,\ Z \sim \mathcal{N}(0,1)$.

Observación 3.5.1. *Notemos que, asintóticamente, el tamaño del test de Wald de tamaño α, es α. En efecto,*

$$\mathbb{P}_{\theta_0}\left(|W| \geq z_{\alpha/2}\right) = \mathbb{P}_{\theta_0}\left(\left|\frac{\hat{\theta}-\theta_0}{ee}\right| \geq z_{\alpha/2}\right) \to= \mathbb{P}_{\theta_0}\left(|Z| \geq z_{\alpha/2}\right) = \alpha \tag{3.39}$$

donde, hemos usado que $Z \sim \mathcal{N}(0,1)$.

Ejemplo 3.5.1. *Consideremos dos conjuntos de observaciones $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_n$, con medias respectivas μ_1 y μ_2. Se requiere evaluar las hipótesis*

$$H_0 : \mu_x = \mu_y \quad v.s. \quad H_1 : \mu_x \neq \mu_y, \tag{3.40}$$

lo cual está dentro del alcance del test de Wald denotando $\delta = \mu_x - \mu_y$ e identificando las hipótesis

$$H_0 : \delta = 0 \quad v.s. \quad H_1 : \delta \neq 0. \tag{3.41}$$

Utilicemos el estimador no-paramétrico 'plug in' de δ dado por $\hat{\delta} = \bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{m}\sum_{i=1}^{m} y_i$, Además, la varianza de este estimador está dada por $v = \frac{1}{n}s_x^2 + \frac{1}{m}s_y^2$, con lo que la condición sobre el estadístico de Wald indica que si

$$W = \frac{\hat{\delta}-0}{ee} = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{n}s_x^2 + \frac{1}{m}s_y^2}} \geq z_{\alpha/2}, \tag{3.42}$$

rechazamos H_0.

Observación 3.5.2. *Notemos que el test de Wald de tamaño α rechaza $H_0 : \theta = \theta_0$ (v.s. $H_1 : \theta \neq \theta_0$) si y solo si*

$$\theta_0 \notin (\hat{\theta} - ee z_{\alpha/2}, \hat{\theta} + ee z_{\alpha/2}), \tag{3.43}$$

es decir, realizar el test de Wald es equivalente a calcular el α intervalo de confianza para el parámetro θ_0.

3.6. Test de razón de verosimilitud

Consideremos un caso más general que los anteriores, donde al menos una de las hipótesis es compuesta, es decir, especifican que el parámetro pertenece a un conjunto en vez de tomar un valor puntual. Es decir,

$$H_0 : \theta \in \Theta_0 \quad \text{v.s.} \quad H_1 : \theta \notin \Theta_0. \tag{3.44}$$

El test de razón de verosimilitud (TRV) indica que se debe rechazar H_0 si

$$\lambda(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \leq C, \tag{3.45}$$

donde $\hat{\theta}$ es el EMV y $\hat{\theta}_0$ es el EMV restringido a $\{\theta \in \Theta_0\}$. Notemos que la región de rechazo está dada por

$$R^* = \{x \in \mathcal{X} | \lambda(x) \leq C\}. \tag{3.46}$$

Observación 3.6.1. *Para el caso de hipótesis simples, es decir, $\Theta = \{\theta_0, \theta_1\}$ y $\Theta_0 = \{\theta_0\}$, entonces el TRV coincide con el test de Neyman-Pearson.*

Al igual que en el TNP, fijamos C en función del un nivel deseado α.

Observación 3.6.2. *Notemos que podemos escribir la expresión en la ecuación* (3.45) *como*

$$\lambda(x_1, \ldots, x_n) = \mathbb{1}_{\hat{\theta} \in \Theta_0} + \mathbb{1}_{\hat{\theta} \in \Theta_1} \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_1} L(\theta)} \tag{3.47}$$

donde el segundo término (de activarse) es estrictamente menor que 1, con lo que el TRV puede enunciarse en función del estadístico

$$\tilde{\lambda}(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_1} L(\theta)} \leq \tilde{k}. \tag{3.48}$$

Ejemplo 3.6.1 (TRV Bernoulli)**.** *Sea $X_1, \ldots, X_n \sim Ber(\theta)$ iid, se quiere resolver*

$$H_0 : \theta \leq \theta_0 \quad v.s. \quad H_1 : \theta > \theta_0, \tag{3.49}$$

donde θ_0 es conocido y sabemos que $p_\theta(x) = \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}$. En la notación de la definición anterior del TRV, podemos identificar

$$\Theta_0 = [0, \theta_0] \quad \& \quad \Theta_1 = (\theta_0, 1] \tag{3.50}$$

calculamos el EMV (restricto e irrestricto) mediante

$$\hat{\theta} = \hat{x} \quad \textit{irrestricto} \tag{3.51}$$

$$\hat{\theta}_0 = \hat{x} \quad \textit{si } \hat{x} \in \Theta_0,\ \theta_0 \textit{ si no.} \tag{3.52}$$

podemos escribir esta última expresión como $\hat{\theta}_0 = \hat{x}\mathbb{1}_{\hat{x}\in\Theta_0} + \theta_0\mathbb{1}_{\hat{x}\notin\Theta_0}$*, entonces*

$$\lambda(x) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \tag{3.53}$$

$$= \frac{L(\hat{x})}{L(\hat{x})}\mathbb{1}_{\hat{x}\in\Theta_0} + \frac{L(\theta_0)}{L(\hat{x})}\mathbb{1}_{\hat{x}\notin\Theta_0} \tag{3.54}$$

$$= \mathbb{1}_{\hat{x}\in\Theta_0} + \mathbb{1}_{\hat{x}\notin\Theta_0}\left(\frac{\theta_0}{\hat{x}}\right)^{n\bar{x}}\left(\frac{1-\theta_0}{1-\hat{x}}\right)^{n(1-\bar{x})} \tag{3.55}$$

graficar $\lambda(x)$ *(en función de* $\bar{x}$*)*

Donde ahora rechazaremos si $\lambda(x) \leq C$*, pero, ¿cómo elegimos* C*?*
Recordemos que queremos que el test sea de nivel α*, es decir,*

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\lambda(x) \leq C\right) = \alpha \tag{3.56}$$

donde recordemos que $\lambda(x)$ *es una función decreciente de* $\bar{x}$*, por lo que la condición* $\lambda(x) \leq C$ *puede expresarse como* $\bar{x} \geq C'$*, para algún* C'*. Esta expresión dependerá de* C'*, que es función de* C*, de* θ_0 *y de* α*; despejamos para* C*.*

clase 17: 8/10

3.7. Test χ^2

Este test es usado para verificar si datos multinomiales siguen una distribución dada. Es decir, denotando $p_0 = (p_{01}, \ldots, p_{0k}) \in [0,1]^k$ y $n \in \mathbb{N}$, estamos interesados en verificar si la distribución p de una variable multinomial sigue

$$H_0 : p = p_0 \quad \text{v.s.} \quad H_1 : p \neq p_0. \tag{3.57}$$

Para esto consideramos $X_1, \ldots, X_n \sim \text{Bin}(X; p, n)$ y construyamos el siguiente estadístico

$$T = \sum_{i=1}^{k} \frac{(X_i - np_{0j})^2}{np_{0j}} = \sum_{i=1}^{k} \frac{(X_i - \mathbb{E}(X_j))^2}{\mathbb{E}(X_j)}, \tag{3.58}$$

donde las esperanzas $\mathbb{E}(X_j) = np_{0j}$ son tomadas con respecto a p_0. Este estadístico representa una medida de discrepancia entre frecuencias observadas (X_i) y esperadas (np_{0j}).

Bajo H_0, la distribución asintótica (es decir, cuando el número de observaciones tiene a infinito) es chi cuadrado con $k-1$ grados de libertad, lo cual denotamos por χ^2_{k-1}. Consecuentemente, el test rechaza la hipótesis nula si $T \geq \chi^2_{k-1,\alpha}$, donde $\mathbb{P}\left(T \geq \chi^2_{k-1,\alpha}\right) = \alpha$.

Ejemplo 3.7.1 (Perritos vagos)**.** *Consideremos un criadero de perros recogidos de la calle, donde podemos identificar cada individuo según tamaño (grande/chico) y color (blanco/negro). Consecuentemente, existen las categorías*

{grande-negro, grande-blanco, chico-negro, chico-blanco}.

En base a un estudio preliminar realizado por el GrUpo cAnino Universitario (GUAU), se sabe que estas clases deben distribuir, respectivamente, de acuerdo a

$$\{9/16, 3/16, 3/16, 1/16\}.$$

El último censo del criadero, el cual solo pudo recolectar información de 556 ejemplares, arrojó los siguiente valores observados: $X = (315, 101, 108, 32)$. *Esto permite calcular el estadístico del test* χ^2 *mediante la ecuación* (3.58) *obteniendo*

$$T_{\chi^2} = 0{,}47. \tag{3.59}$$

Para $\alpha = 0{,}05$, *tenemos* $\chi^2_{3,\alpha} = 7{,}815 > 0{,}47$, *por lo que no existe evidencia suficiente para rechazar* H_0. *Es decir, el censo no contradice el estudio realizado por el GUAU.*

Si bien a primera vista, el test χ^2 parece un tanto simple, pues es *solo* definido para VA multinomiales, es un test del tipo *bondad de ajuste* que puede ser extendido a casos no paramétricos.

definir qué es un problema estadístico no paramétrico

Asumamos $X_1, \ldots, X_n \sim F$ iid, donde F es una distribución desconocida sobre $\mathbb{R}$ (tanto en forma como en parámetros). Planteamos además el siguiente problema de test de hipótesis

$$H_0 : F = F_0 \quad \text{v.s.} \quad H_1 : F \neq F_0, \tag{3.60}$$

donde F_0 sí es una distribución conocida. Podemos aplicar el test χ^2 mediante la discretización de este problema: definimos los intervalos

$$I_i = (a_{i-1}, a_i], \quad -a_0, a_k = \infty, \quad \{a_i\}_{i=1}^{k-1} \subset \mathbb{R}, \tag{3.61}$$

y definimos las VAs

$$N_i = \sum_{j=1}^{n} \mathbb{1}_{x_j \in I_i}, \tag{3.62}$$

las cuales tienen ley multinomial con parámetro $\theta_i = F(a_i) - F(a_{i-1}), \quad i = 1, \ldots, k$. Finalmente, bajo H_0 resulta $\theta_i = F_0(a_i) - F_0(a_{i-1}), i = 1, \ldots, k$, lo cual nos permite aplicar el test χ^2.

3.8. Test de Kolmogorov-Smirnov

Ahora consideramos otro enfoque, basado en una estrategia muy distinta, al problema de test de hipótesis anterior para distribuciones no paramétricas:

$$H_0 : F = F_0 \quad \text{v.s.} \quad H_1 : F \neq F_0. \tag{3.63}$$

En vez de discretizar, podemos construir la distribución empírica dada por

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_j \leq x}, \tag{3.64}$$

la cual realmente es una distribución (discontinua).

Sabemos que, debido a la ley de los grandes números,

$$F_n(x) \to \mathbb{E}\left(\mathbb{1}_{X \leq x}\right) = \mathbb{P}\left(X \leq x\right) = F(x), \tag{3.65}$$

además, por el teorema de Glivenko-Cantelli, tenemos

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{c.s.} \tag{3.66}$$

Lo anterior nos permite definir el estadístico $D_n = \sup_x |F_n(x) - F_0(x)|$ y la región crítica

$$R = \{x | D_n \geq k_\alpha\}, \tag{3.67}$$

donde k_α se elige imponiendo $\mathbb{P}_{\theta_0}\left(D_n \geq k_\alpha\right) = \alpha$.

Observación 3.8.1. *El test de Kolmogorov-Smirnoff sirve tanto para verificar si una VA sigue una distribución dada o si bien dos distribuciones siguen la misma distribución (desconocida).*

3.9. Test de Wilcoxon

Este es otro testo no paramétrico para verificar si dos VAs siguen la misma distribución. Consideremos las observaciones

$$X_1, \dots, X_n \sim F, \quad Y_1, \dots, Y_n \sim G, \tag{3.68}$$

donde F y G son dos distribuciones, de las cuales solo sabemos que son continuas.

Consideremos el siguiente problema de test de hipótesis:

$$H_0 : F = G \quad \text{v.s.} \quad H_1 : F \neq G. \tag{3.69}$$

El test de Wilcoxon se enfoca en este escenario pero solo es sensible a diferentes *localizaciones*, es decir, si G es una versión desplazada de F.

Antes de ver el test de Wilcoxon, notemos que si nos interesase detectar estas desviaciones, entonces podríamos considerar un test que rechace H_0 si $|\bar{X} - \bar{Y}| \geq K$. Esto es exactamente lo que hace el TRV en el problema

$$H_0 : \mu = \eta \quad \text{v.s.} \quad H_1 : \mu \neq \eta, \tag{3.70}$$

cuando $X \sim \mathcal{N}(\mu, \sigma^2)$, $Y \sim \mathcal{N}(\eta, \sigma^2)$.

Sin embargo, en el caso general (cuando no sabemos nada de F) obtener la ley de $|\bar{X} - \bar{Y}|$ bajo H_0 no es trivial, lo cual es necesario para $\mathbb{P}_{\theta_0}(|\bar{X} - \bar{Y}| \geq K) = \alpha$. En esta situación, el test de Wilcoxon propone considerar la siguiente observación conjunta

$$(z_1, \dots, z_{m+n}) = (x_1, \dots, x_n, y_1, \dots, y_m), \tag{3.71}$$

para luego considerar la secuencia ordenada de valores z_i dados por

$$\min_i\{z_i\} = z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n+m)} = \max_i\{z_i\}. \tag{3.72}$$

Ahora podemos definir el concepto de *rango* como la posición en el orden anterior, es decir, donde el rango de $z_{(1)}$ es 1, el rango de $z_{(2)}$ es dos y así sucesivamente.

dibujo de bolas negras y blancas.

Denotando el rango de x_i como R_i, podemos construir el estadístico

$$W = \sum_{i=1}^{n} R_i, \tag{3.73}$$

esta cantidad debe intuitivamente interpretarse como el promedio de los rangos (es decir de la posiciones) que toman las observaciones de la variable X, por rechazamos H_0 si W es muy pequeño o muy grande, es decir, si las muestras de X no quedan *mezcladas* con las de Y.

Esto es posible por que la distribución de W bajo H_0 puede ser calculada y de hecho no depende de F, esto es porque (bajo H_0) los elementos de z_i son iid, con lo que todas las posible permutaciones del los valores z_i tienen la misma probabilidad dada por $\binom{n+m}{n}$.

Observación 3.9.1 (¿Cómo obtenemos la región crítica R para este test?)**.** *Podemos proceder de forma iterativa: Asumimos H_0, consideramos $R = \emptyset$ y agregamos las configuraciones de bolitas que tienen el menor y mayor valor de W, luego seguimos con las siguientes configuraciones hasta acumular una probabilidad $\mathbb{P}_{\theta_0}(W \in \mathbb{R}) = \alpha$.*

Capítulo 4

Inferencia bayesiana

El enfoque que hemos revisado hasta ahora es el clásico o *frecuentista*, este punto de vista asume lo siguiente:

- El concepto de probabilidad está relacionado con frecuencias límites, es decir, la probabilidad de un evento es la razón de veces que este ocurre versus las veces que no ocurre (usualmente referido como *casos favorables dividido por casos totales*). En este sentido, la probabilidad es una propiedad del mundo real.
- Los parámetros son constantes (fijos) y desconocidos, es decir, no existe *aleatoriedad* relacionada a los parámetros, por ende no podemos construir enunciados probabilísticos con respecto a ellos
- El procedimiento estadístico debe comportarse bien en el largo plazo, un ejemplo de esto es que un $(1-\alpha)$-intervalo de confianza debe capturar (asintóticamente) el parámetro una fracción $1-\alpha$ de las veces luego de infinitos experimentos.

En este capítulo consideraremos un enfoque alternativo al análisis estadístico, el cual llamaremos **enfoque bayesiano**, y que se caracteriza por lo siguiente:

- La probabilidad es subjetiva y denota un grado de *creencia*, es decir, la aleatoriedad de un evento no solo es intrínseca de éste sino también de nuestra observación
- Lo anterior permite considerar aleatoriedad en los parámetros, pues el hecho de que éstos sean fijos no quiere decir que los conozcamos.

- Podemos considerar los parámetros como VAs y, consecuentemente, calcular su distribución de probabilidad. Inferencias puntuales o la incidencia de este parámetro en otras VAs está completamente determinada por su distribución.

Existen ventajas y desventajas para ambos enfoques, lo cual hace que ambos sean considerados en distintas aplicaciones. Si bien el enfoque bayesiano es muy antiguo, la estadística clásica ha privilegiado un punto de vista frecuentista, mientras que disciplinas como minería de datos y aprendizaje de máquinas se inclinan por el enfoque bayesiano. De todas formas actualmente ambos métodos se consideran en base a sus propios méritos, entonces, dejando materias filosóficas de lado nos dedicaremos a estudiar como se hace inferencia bayesiana.

clase 18: 10/10

4.1. Distribuciones a priori y a posteriori

Consideraremos que tanto el parámetro θ y la cantidad de interés X son VAs. En este sentido, definiremos la distribución del parámetro como $p(\theta)$ y la distribución de X condicional al parámetro como $p(X|\theta)$. Observemos que ya no usamos la notación $p_\theta(X)$, pues damos énfasis en que θ es una VA. Nos referiremos a $p(\theta)$ como la distribución **a priori** sobre θ, o simplemente el prior sobre θ. Formalmente, re-definamos el concepto de *familia paramétrica*.

Definición 4.1.1. *Sea $(S, \mathcal{A}, \mu)$ un espacio de probabilidad y sean $(\mathcal{X}, \mathcal{B})$ y (Ω, τ) espacios borelianos. Sean $X : S \to \mathcal{X}$ y $\Theta : S \to \Omega$ funciones medibles, entonces, Θ se llama parámetro y Ω se llama espacio de parámetros. La distribución condicional X dado Θ se llama familia paramétrica de distribuciones de X y está denotada por*

$$\mathcal{P} = \{P_\theta \; t.q. \; \forall A \in \mathcal{B}, P_\theta(A) = \mathbb{P}\left(X \in A | \Theta = \theta\right), \theta \in \Omega\} \tag{4.1}$$

.

Observemos que la definición de familia paramétrica en el contexto bayesiano solo dice relación con las distribuciones condicionales $\mathbb{P}\left(X|\theta\right)$ que denota la distribución de X una vez que $\Theta = \theta$ ha sido observado. Sin embargo, el contexto

también permite identificar la distribución *a priori* de Θ, μ_Θ, la cual es una medida en (Ω, τ) inducida por Θ desde μ.

Asumiremos que P_θ como medida en $(\mathcal{X}, \mathcal{B})$ tiene densidad (con respecto a alguna medida ν) dada por

$$p_\theta(x|\theta) = \frac{dP_\theta}{d\nu}(x). \tag{4.2}$$

Asumiremos que $p_\theta(x|\theta)$ es medible con respecto a la sigma-álgebra producto, lo cual permitirá integrarla tanto en $x \in \mathcal{X}$ como en $\theta \in \Omega$. Esto nos permite verificar que la densidad condicional de X dado θ con respecto a ν cumple con

$$\mathbb{P}\left(X \in A | \Theta = \theta\right) = \int_A p(x|\theta) \mathrm{d}\nu(x), \ A \in \mathcal{B}. \tag{4.3}$$

Adicionalmente, podemos notar que la distribución **marginal** de X puede ser calculada integrando la distribución condicional con en todo el espacio de parámetros: Para $A \in \mathcal{B}$ tenemos

$$\begin{aligned} \mu_X(A) &= \int_\Omega \left(\int_A p(x|\theta) \mathrm{d}\nu(x) \right) \mathrm{d}\mu_\Theta(\theta) \\ &= \int_A \int_\Omega p(x|\theta) \mathrm{d}\mu_\Theta(\theta) \mathrm{d}\nu(x) \end{aligned} \tag{4.4}$$

lo cual implica directamente que $\mu_X(A)$ es absolutamente contínua c.r.a. ν también con densidad

$$p(x) = \int_\Omega p(x|\theta) \mathrm{d}\mu_\Theta(\theta) \tag{4.5}$$

la cual conocemos como densidad marginal de X.

La distribución condicional de Θ dado $X = x$, denotada por $\mu_{\Theta|X}(\theta|x)$, es conocida como la distribución a posteriori o simplemente *posterior*. El siguiente teorema es el elemento fundamental de la inferencia bayesiana, pues permite calcular la distribución posterior.

Teorema 4.1.1 (Bayes)**.** *Asumamos que X sigue una familia paramétrica $\{P_\theta\}$ como en la Definición 4.1.1 y asumamos que $P_\theta \ll \nu, \forall \theta \in \Omega$, para alguna medida ν en $(\mathcal{X}, \mathcal{B})$. Denotemos*

- *$p(x|\theta)$: la densidad (con respecto a ν) de X dado $\Theta = \theta$*
- *μ_Θ: la distribución a priori de Θ*

- $\mu_{\Theta|X}(\cdot|x)$: *la distribución condicional de* Θ *dado* $X = x$.

Entonces, $\mu_{\Theta|x} \ll \mu_\Theta$ *c.s. con respecto a la ley marginal de* X, μ_X, *y su derivada de Radon-Nikodym es*

$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta|x) = \frac{p(x|\theta)}{\int_\Omega p(x|t)d\mu_\Theta(t)} \tag{4.6}$$

demostración está pendiente

Con el resultado anterior, podemos verificar que para $T \in \Omega$ la posterior cumple

$$\mathbb{P}(\Theta \in T) = \int_C \frac{p(x|\theta)}{\int_\Omega p(x|t)\mathrm{d}\mu_\Theta(t)} d\mu_\Theta(\theta) \tag{4.7}$$

$$= \frac{1}{\int_\Omega p(x|t)\mathrm{d}\mu_\Theta(t)} \int_C p(x|\theta)d\mu_\Theta(\theta) \tag{4.8}$$

Observación 4.1.1 (Prior absolutamente continuas)**.** *Si bien no es necesario, asumamos que la distribución a priori* μ_X *es a.c. con respecto a la medida de Lebesgue,* λ, *y su densidad con respecto a dicha medida es*

$$p(\theta) = \frac{d\mu_\Theta}{d\lambda}(\theta). \tag{4.9}$$

Entonces, como el teorema de Bayes establece que $\mu_{\Theta|x} \ll \mu_\Theta$ *y nosotros hemos considerado* $\mu_\Theta \ll \lambda$, *entonces tenemos que* $\mu_{\Theta|x} \ll \lambda$ *con densidad respecto a la medida de Lebesgue*

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}, \tag{4.10}$$

donde recordemos que $p(x) = \int_\Omega p(x|\theta)p(\theta)d\theta$.

Ejemplo 4.1.1 (Posterior Gaussiana: varianza conocida)**.** *Consideremos la familia paramétrica* $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \sigma^2 \text{ conocido.}\}$ *y el prior sobre* μ *dado por* $\mathcal{N}(\mu; \mu_0, \sigma_0^2)$. *Dada una muestra* $X_1, \ldots, X_n$ *la posterior está dada por*

$$p(\theta|x) = \mathcal{N}\left(\theta; \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i^n x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right) \tag{4.11}$$

Ejemplo 4.1.2 (Posterior Bernoulli - prior uniforme). *Consideremos las VAs $X_1, \ldots, X_n \sim$ Ber (θ) con un prior uniforme sobre θ. La posterior está dada por*

$$p(\theta) = \frac{\theta^{\sum X_i}(1-\theta)^{(n-\sum X_i)} \cdot 1}{p(x)} \tag{4.12}$$

lo cual podemos identificar como una distribución beta de parámetros $\alpha = \sum X_i + 1$ y $\beta = n - \sum X_i + 1$, es decir

$$p(\theta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{4.13}$$

Observación 4.1.2. *Observemos que no hemos calculado explícitamente la constante de normalización, esto es porque hemos calculado una versión proporcional de la posterior y luego hemos impuesto que su integral en todo Ω deber ser 1. Este procedimiento permite no solo una forma alternativa de calcular esta constante sino que muchas veces la única.*

Clase 19: 15/10

Ejemplo 4.1.3 (Posterior Bernoulli - prior Beta). *Consideremos las VAs $X_1, \ldots, X_n \sim$ Ber (θ) pero ahora con un prior Beta dado por*

$$p(\theta) = \frac{\Gamma(\alpha_0)\Gamma(\beta_0)}{\Gamma(\alpha_0+\beta_0)}\theta^{\alpha_0-1}(1-\theta_0)^{\beta_0-1}. \tag{4.14}$$

Es este caso, la posterior está dada por

$$p(\theta) = \frac{\Gamma(\alpha_0)\Gamma(\beta_0)}{\Gamma(\alpha_0+\beta_0)}\frac{\theta^{\sum X_i}(1-\theta)^{(n-\sum X_i)}\theta^{\alpha_0-1}(1-\theta_0)^{\beta_0-1}}{p(x)} \tag{4.15}$$

$$= \frac{\Gamma(\alpha_0)\Gamma(\beta_0)}{p(x)\Gamma(\alpha_0+\beta_0)}\theta^{\sum X_i+\alpha_0-1}(1-\theta)^{(n-\sum X_i+\beta_0-1)} \tag{4.16}$$

$$= Beta\left(\sum X_i + \alpha_0, n - \sum X_i + \beta_0\right), \tag{4.17}$$

lo cual podemos identificar como una distribución beta de parámetros $\alpha = \sum X_i + 1$ y $\beta = n - \sum X_i + 1$, es decir

$$p(\theta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{4.18}$$

Consecuentemente, pasar del prior al posterior es simplemente actualizar los parámetros de la forma

$$\alpha_0 \to \sum X_i + \alpha_0 \tag{4.19}$$
$$\beta_0 \to n - \sum X_i + \beta_0 \tag{4.20}$$

Notemos que, para ciertas elecciones de distribuciones a priori la forma de la densidad posterior cambia y otras veces se mantiene (con respecto a la forma del prior). La siguiente definición caracteriza este fenómeno.

Definición 4.1.2 (Prior conjugado)**.** *Sea un modelo con densidad $p(x|\theta)$ y un prior sobre θ con densidad $p(\theta)$. Decimos que $p(\theta)$ es conjugado con la verosimiltud $p(x|\theta)$ si la posterior*

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{4.21}$$

pertenece a la misma familia que el prior $p(\theta)$.

El hecho que dos distribuciones (i.e, prior y posterior) pertenezcan a la *misma familia*, quiere decir que ambas tienen la misma forma funcional, e.g., $f_\lambda(\theta)$ pero con distintos valores para el *parámetro* λ, el cual es un *hiperparámetro* del modelo.

Ejemplo 4.1.4 (Distribución Multinomial)**.** *Consideremos un VA multinomial $X \sim Mult\,(n, \theta)$ donde θ pertenece al simplex*

$$\{\theta \in [0,1]^k : \theta_1 + \cdots + \theta_k = 1\}. \tag{4.22}$$

La distribución multinomial genera vectores $X \in \mathbb{N}^k$ cuya $i-$ésima componente modela la cantidad de veces que ocurre el evento i dentro de k eventos en n intentos. Por ejemplo, si lanzamos un dado balanceado 100 veces, el vector que contiene el conteo de veces que obtenemos cada cara puede modelarse como

$$\theta_{dado} \sim Mult\left(100, \left[\tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}\right]\right). \tag{4.23}$$

Denotando $X = [x_1, \ldots, x_n]$, observemos que una muestra multinomial $X \sim Mult\,(n, \theta)$ cumple con

$$\{x_i\}_{i=1}^k \subset \{0, 1, \ldots, n\}, \quad \sum_{i=1}^k x_i = n. \tag{4.24}$$

Finalmente, la distribución Multinomial está dada por

$$Mult\,(X; n, \theta) = \frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k}, \tag{4.25}$$

y es la generalización de

- *Bernoulli cuando* $k = 2$ *y* $n = 1$*; pues* $Ber\,(X; \theta) = \theta^x (1-\theta)^{1-x}$
- *Categórica: cuando* $n = 1$*; pues* $Cat\,(X; \theta) = \theta_1^{x_1} \cdots \theta_k^{x_k}$
- *Binomial: cuando* $k = 2$*; pues* $Bin\,(X; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

Observemos que el parámetro θ en la distribución multinomial (y las otras tres) es directamente la distribución discreta. Es decir, el construir un prior $p(\theta)$ implica definir una distribución sobre distribuciones discretas.

Definición 4.1.3 (Distribución de Dirichlet)**.** *Consideremos la distribución de Dirichlet*

$$\theta \sim Dir\,(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}, \tag{4.26}$$

donde $\alpha = (\alpha_1, \ldots, \alpha_k)$ *es el parámetro de concentración y la constante de normalización está dada por* $B(\alpha) = \prod_{i=1}^{k} \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^{k} \alpha_i)$*. El soporte de esta distribución es el simplex presentado en la ecuación* (4.22)*.*

En el caso $k = 3$, la distribución de Dirichlet puede ser graficada en el simplex de 2 dimensiones. La Figura 4.1 presenta tres gráficos para distintos valores del parámetro de concentración.

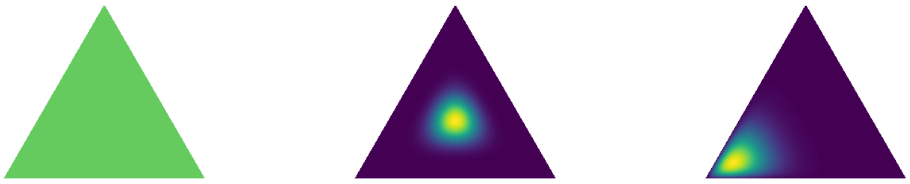

Figura 4.1: Distribuciones Dirichlet para $k = 3$ con parámetros de concentración α (desde izquierda a derecha) dado por $[1, 1, 1]$, $[10, 10, 10]$ y $[10, 2, 2]$.

.

Veamos a continuación que la distribución de Dirichlet es conjugada al modelo Multinomial, y consecuentemente para Bernoulli, Categórica y Binomial. En efecto, si $\theta \sim \text{Dir}(\theta;\alpha)$ y $X \sim \text{Mult}(X;n,\theta)$, entonces

$$\begin{aligned} p(\theta|x) &= \frac{\text{Mult}(x;n,\theta)\,\text{Dir}(\theta;\alpha)}{p(x)} \\ &= \frac{n!}{x_1!\cdots x_k!p(x)B(\alpha)}\prod_{i=1}^{k}\theta_i^{x_i+\alpha_i-1} \\ &= \frac{1}{B(\alpha')}\prod_{i=1}^{k}\theta_i^{\alpha'_i-1} \end{aligned} \tag{4.27}$$

donde $\alpha' = (\alpha'_1,\ldots,\alpha'_k) = (\alpha'_1 + x_1,\ldots,\alpha'_k + x_k)$ es el nuevo parámetro de concentración.

Ejemplo 4.1.5. *Consideremos $\alpha = [1,2,3,4,5]$ y generemos una muestra de $\theta \sim Dir(\theta|\alpha)$. El siguiente código genera, grafica e imprime esta muestra.*

```
import numpy as np
alpha = np.array([1,2,3,4,5])
theta = np.random.dirichlet(alpha)
plt.bar(np.arange(5)+1, theta);
print(f'theta = {theta}')
```

En nuestro caso, obtuvimos los parámetros $\theta = [0{,}034, 0{,}171, 0{,}286, 0{,}185, 0{,}324]$.

Ahora, usaremos un prior Dirichlet sobre θ con $\alpha_p = [1,1,1,1,1]$ calcularemos la posterior de acuerdo a la ecuación 4.27. La Figura 4.2 muestra 100 muestras de la distribución posterior para una cantidad de observaciones igual a

Los modelos y sus prior conjugados que hemos visto hasta ahora han pertenecido a la familia exponencial. En general, notemos que en el caso general si el modelo sigue una densidad de la familia exponencial, es decir,

$$p(x|\theta) = h_m(x)\exp\left(\theta^\top T(x) - A_m(\theta)\right), \tag{4.28}$$

tiene un prior conjugado dado por

$$p(\theta|\lambda) = h_p(\theta)\exp\left(\lambda_1^\top\theta + \lambda_2^\top(-A_m(\theta)) - A_p(\lambda)\right), \tag{4.29}$$

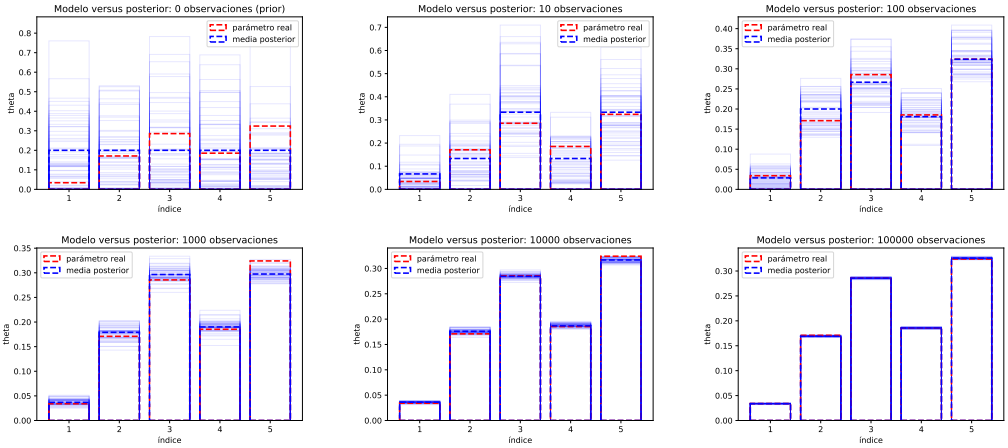

Figura 4.2: Concentración de la distribución posterior en torno al parámetro real para $X \sim \text{Mult}(\theta)$ y $\theta \sim \text{Dir}(\alpha)$. Se considera desde 0 hasta 10^5 observaciones y cada gráfico (desde izquierda-arriba hasta derecha-abajo) muestra el parámetro real (linea roja quebrada), la media de la posterior (línea azul quebrada) y 50 muestras de la posterior (azul claro). Observe como la distribución a priori (linea roja quebrada en la primera figura) pierde importancia a medida que el número de observaciones aumenta.

donde el parámetro natural $\lambda = [\lambda_1, \lambda_2]$ tiene dimensión $\dim(\theta) + 1$ y el estadístico suficiente es $[\theta, -A_m(\theta)]$. El resto de los términos como h_p y A_p dependen de la forma particular de modelo.

Calculando la distribución posterior tenemos que

$$
\begin{aligned}
p(\theta|x_{1:n}, \lambda) &\propto \prod_{i=1}^{n} p(x_i|\theta, \lambda) p(\theta|\lambda) \qquad (4.30)\\
&= \left(\prod_{i=1}^{n} h_m(x_i)\right) h_p(\theta) \exp\left(\theta^\top \sum_{i=1}^{n} T(x_i) - nA_m(\theta) + \lambda_1^\top \theta + \lambda_2^\top(-A_m(\theta)) - A_p(\lambda)\right)\\
&\propto h_p(\theta) \exp\left(\theta^\top \sum_{i=1}^{n} T(x_i) - nA_m(\theta) + \lambda_1^\top \theta + \lambda_2^\top(-A_m(\theta)) - A_p(\lambda)\right)\\
&\propto h_p(\theta) \exp\left(\left(\sum_{i=1}^{n} T(x_i) + \lambda_1\right)^\top \theta + (n + \lambda_2)(-A_m(\theta))\right)
\end{aligned}
$$

4.2. Elección del prior

Recordemos que la distribución a priori sobre el parámetro pierde importancia con una cantidad suficiente de datos, sin embargo, cuando se cuenta con pocos datos la elección del prior es fundamental. Esta distribución a priori se elige no solo en base a las propiedades de conjugación sino que también en función del conocimiento experto del modelo. De esta forma, es posible incorporar información adicional que puede no ser revelada inmediatamente por los datos pero también sesgos.

Cuando no tenemos información experta sobre el proceso en cuestión, podemos elegir una distribución a priori *no informativa*, es decir, una que no aporta información (pues no la tenemos). Es natural pensar que un prior no informativo es uniforme sobre el espacio de parámetros, es decir,

$$p(\theta) \propto 1. \tag{4.31}$$

Observemos que este prior, como densidad de probabilidad, está bien definido solo cuando la dimensión del espacio de parámetros es finita. Cuando el espacio de parámetros no es finito, por ejemplo cuando $\theta \in \Omega = \mathbb{R}$, entonces no podemos definir esta distribución como una uniforme. Sin embargo, notemos que si elegimos $p(\theta)$ una distribución que no necesariamente integra uno, el teorema de Bayes

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int_\Omega p(x|\theta)p(\theta)\mathrm{d}\theta} \tag{4.32}$$

aún entrega distribuciones posteriores apropiadas.

Definición 4.2.1 (Prior impropia)**.** *Una distribución a priori impropia es una distribución que no necesariamente integra 1, pero de todas formas puede ser utilizada como distribución a priori en el contexto de inferencia bayesiana.*

Observación 4.2.1. *Veamos que un prior impropio puede incluso tener integral infinita, en el caso de la distribución normal $X \sim \mathcal{N}(X; \mu, 1)$, $\mu \in \mathbb{R}$, podemos elegir $p(\mu) \propto 1$ y escribir*

$$p(\mu|x) \propto p(x|\mu) \cdot 1 = \mathcal{N}(x; \mu, 1) = \mathcal{N}(\mu; x, 1) \tag{4.33}$$

Considera priors uniformes impropias como priori no informativas parece tener sentido, pues intuitivamente no estamos dando preferencia (mayor probabilidad a priori) a ningún valor del parámetro por sobre otro. Sin embargo, este procedimiento sufre de una desventaja conceptual.

Consideremos $X \sim p(x|\theta)$, $\theta \in [a, b]$, en donde elegimos el prior *no informativo* uniforme dado por

$$p(\theta) = \text{Uniforme}\,(a, b) = \frac{1}{b-a}. \tag{4.34}$$

Consideremos ahora un modelo *reparametrizado* $\eta = e^{\theta} \in [c, d]$, donde el modelo es expresado como $X \sim q(x|\eta) = p(x|\theta)$. El prior uniforme para el nuevo parámetro es

$$p(\eta) = \text{Uniforme}\,(c, d) = \frac{1}{d-c}. \tag{4.35}$$

Observemos que la elección uniforme del parámetro θ en el intervalo $[a, b]$ es equivalente a elegir η según

$$\tilde{p}(\eta) = p(\theta) \left| \frac{d\theta}{d\eta} \right| = \frac{1}{b-a} \left| \frac{d \log \eta}{d\eta} \right| = \frac{1}{\eta(b-a)} \tag{4.36}$$

es decir, la distribución sobre η inducida por $p(\theta)$. Esta distribución por supuesto no es equivalente a elegir η uniformemente en el intervalo $[c, d]$.

Una forma de construir un prior que es invariante ante reparametrizaciones es mediante la metodología propuesta por Harold Jeffreys (1946), el que sugiere elegir un prior proporcional a la raíz cuadrada del determinante de la información de Fisher, es decir,

$$p(\theta) \propto (\det I(\theta))^{1/2} \tag{4.37}$$

donde recordemos que la información de Fisher está dada por

$$I(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \log p(X|\theta) \right) = \mathbb{E}_\theta \left(\left(\frac{\partial}{\partial \theta} \log p(X|\theta) \right)^2 \right) \tag{4.38}$$

y si $X_1, \ldots, X_n$ son iid, entonces $I(\theta) = nI_1(\theta)$ y el prior de Jeffreys puede ser expresado como

$$p(\theta) \propto I_1(\theta)^{1/2}. \tag{4.39}$$

Observemos que si $\int_\Omega \sqrt{I(\theta)} \mathrm{d}\theta$ es finito, entonces la constante de proporcionalidad es precisamente esta cantidad. Sin embargo, si esta cantidad es infinita el prior de Jeffreys aún es un prior válida pero impropio, siempre y cuando las posteriores respectivas sí sean propias.

Veamos ahora que el prior de Jeffreys es invariante bajo reparametrizaciones. Consideremos los modelos relacionados mediante reparametrización dados por

$$X \sim p(x|\theta),\ \theta \in \Omega \quad \& \quad X \sim q(x|\eta),\ \eta \in \Gamma \tag{4.40}$$

donde $\eta = h(\theta)$. Las informaciones de Fisher para ambos modelos, denotadas respectivamente $I_p(\theta)$ e $I_q(\theta)$, están relacionadas mediante

$$I_p(\theta) = \int_{\mathcal{X}} \left(\frac{\partial}{\partial \theta} \log p(x|\theta) \right)^2 p(x|\theta) \mathrm{d}x \tag{4.41}$$

$$= \int_{\mathcal{X}} \left(\frac{\partial}{\partial \theta} \log q(x|h(\theta)) \right)^2 q(x|h(\theta)) \mathrm{d}x \tag{4.42}$$

$$= \int_{\mathcal{X}} \left(\frac{\partial}{\partial \eta} \log q(x|\eta) h'(\theta) \right)^2 q(x|h(\theta)) \mathrm{d}x \tag{4.43}$$

$$= \left(h'(\theta) \right)^2 I_q(\eta). \tag{4.44}$$

Observemos ahora que el prior en θ, $p(\theta)$, inducido por el prior de Jeffreys en η, $p_J(\eta)$, es efectivamente el prior de Jeffreys en θ, $p_J(\theta)$. En efecto, debido al cambio de variable tenemos

$$p(\theta) = p_J(\eta) \left| \frac{d\eta}{d\theta} \right| = \sqrt{I_q(\eta)} \left| h'(\theta) \right| = \sqrt{I_p(\theta)} = p_J(\theta) \tag{4.45}$$