Capítulo 2: Familias Exponenciales, Suficiencia e Información

Teoría Estadística Avanzada

Sigla DES124

Prof. Jaime Lincovil

2.1. Introducción

En estadística, la suficiencia y la información son conceptos fundamentales, sin importar el enfoque que se adopte—bayesiano, frecuentista u otro. La idea básica es que, para un modelo estadístico dado $\{P_\theta : \theta \in \Theta\}$, indexado por un espacio de parámetros (finito-dimensional) Θ, existen funciones de los datos observables $X = (X_1, \ldots, X_n)$ que contienen toda la información disponible en X sobre el parámetro desconocido θ. Tales funciones se denominan **estadísticos suficientes**, y la idea es que, en general, es suficiente, por ejemplo, para la estimación puntual, restringir la atención a funciones de estadísticos suficientes. La noción de "información" introducida anteriormente se pretende que sea informal; sin embargo, para mayor rigor, debemos formular una noción precisa de información en un problema estadístico. En particular, nos enfocaremos en la información de Fisher, atribuida a R. A. Fisher. El resultado clave es: la información de Fisher para θ en una función $T = T(X)$ de los datos observables X no es mayor que la información de Fisher para θ en X mismo, y ambas medidas de información son iguales si y solo si T es un estadístico suficiente.

La definición de suficiencia no es útil para encontrar un estadístico suficiente en un problema dado. Afortunadamente, el **teorema de factorización de Neyman-Fisher** facilita bastante esta tarea. La idea es que, con un poco de álgebra sobre la función de verosimilitud, se puede obtener un estadístico suficiente fácilmente. Sin embargo, los estadísticos suficientes no son únicos. Por lo tanto, existe interés en tratar de encontrar el "mejor" estadístico suficiente en un problema dado. Este mejor estadístico suficiente se denomina **mínimo**, y discutimos algunas técnicas para encontrar el estadístico suficiente mínimo. Sin embargo, puede ocurrir que incluso un estadístico suficiente mínimo T proporcione una reducción ineficiente de los datos X, ya sea porque la dimensión de T es mayor que la de θ, o porque hay información redundante en T. En tales casos, tiene sentido considerar la posibilidad de condicionar sobre un **estadístico auxiliar**, una especie de complemento de un estadístico suficiente que no contiene información sobre θ. Existen casos especiales en los que el estadístico suficiente mínimo T es **completo**. Esto significa que T no contiene información redundante sobre θ, por lo que condicionar sobre un estadístico auxiliar es innecesario (teorema de Basu).

Iniciamos este capítulo con una introducción a las familias exponenciales. Esta es una clase

amplia que contiene casi todas las distribuciones que se encuentran en un curso intermedio de estadística. De manera general, lo que hace que las familias exponenciales sean tan útiles es que sus funciones de verosimilitud correspondientes tienen propiedades convenientes que permiten aplicar muchas técnicas. Por ejemplo, las condiciones de regularidad necesarias para la normalidad asintótica de los estimadores de máxima verosimilitud o la desigualdad de Cramér-Rao se cumplen para las familias exponenciales (regulares). Un resultado clave es el **Teorema 1**, que es una aplicación interesante del teorema de convergencia dominada de Lebesgue.

De particular importancia aquí es que, esencialmente, solo las familias exponenciales (regulares) permiten una reducción adecuada de la dimensión mediante la suficiencia (**Teorema 4**). Los detalles sobre las familias exponenciales, incluyendo todo lo presentado aquí, se discuten en la monografía técnica de Brown (1986).

2.2. Familias exponenciales de distribuciones

Anteriormente discutimos el concepto general de una familia paramétrica de medidas de probabilidad, con cierto detalle sobre un caso especial con una estructura inducida por un grupo de transformaciones. En esta sección discutimos una clase muy importante de distribuciones que contiene muchos de los modelos estadísticos comunes, tales como la distribución binomial, de Poisson, normal, etc.

Definición 2.1

Una colección de medidas de probabilidad $\{P_\theta : \theta \in \Theta\}$ en $(\mathbb{X}, \mathcal{A})$, cada una dominada por una medida σ-finita μ, se llama **familia exponencial** si las derivadas de Radon-Nikodym $p_\theta(x) = \frac{dP_\theta}{d\mu}(x)$ satisfacen:

$$p_\theta(x) = h(x)e^{\langle \eta(\theta), T(x)\rangle - A(\theta)} \tag{1}$$

para algunas funciones h, A, η y T, donde

$$\eta(\theta) = (\eta_1(\theta), \ldots, \eta_d(\theta))^\top \quad \text{y} \quad T(x) = (T_1(x), \ldots, T_d(x))^\top.$$

Aquí, $\langle x, y\rangle$ denota el producto interno euclidiano usual entre vectores en $\mathbb{R}^d$, es decir,

$$\langle x, y\rangle = \sum_{i=1}^{d} x_i y_i.$$

Cuando es conveniente, podemos escribir $a(\theta) = e^{-A(\theta)}$ y reescribir la ecuación (1) como

$$p_\theta(x) = a(\theta)h(x)e^{\langle \eta(\theta), T(x)\rangle}.$$

Al considerar esperanzas con respecto a una distribución de la familia exponencial, ocasionalmente podemos absorber el término $h(x)$ en $d\mu(x)$; ver, por ejemplo, el Teorema 1.

Ejemplo 2.1

Supongamos que $X \sim \text{Poisson}(\theta)$. La distribución de Poisson está dominada por la medida de conteo en $\mathbb{X} = \{0, 1, \dots\}$, con densidad

$$p_\theta(x) = \frac{e^{-\theta}\theta^x}{x!} = \frac{1}{x!}e^{x\log\theta - \theta}, \quad x = 0, 1, \dots$$

El lado derecho de esta expresión tiene la forma de (1), por lo que la distribución de Poisson pertenece a la familia exponencial.

Ejemplo 2.2

Las siguientes distribuciones son miembros de la familia exponencial: $\mathcal{N}(\theta, 1)$, $\mathcal{N}(0, \theta^2)$, $\text{Exp}(\theta)$, $\text{Gamma}(\theta_1, \theta_2)$, $\text{Beta}(\theta_1, \theta_2)$, $\text{Bin}(n, \theta)$ y $\text{Geo}(\theta)$. Algunos ejemplos comunes de distribuciones que no pertenecen a una familia exponencial incluyen $\text{Cauchy}(\theta, 1)$ y $\text{Unif}(0, \theta)$.

Existen varias propiedades estadísticas interesantes de las familias exponenciales, en particular aquellas relacionadas con la existencia de estadísticos suficientes y, posteriormente, con la existencia de estimaciones insesgadas de varianza mínima. Las siguientes propiedades matemáticas de las familias exponenciales serán útiles para demostrar estos resultados estadísticos.

Proposición 2.1

Considere una familia exponencial con densidades respecto a μ:

$$p_\theta(x) = a(\theta)h(x)e^{\langle\theta, T(x)\rangle}. \tag{2}$$

El conjunto

$$\Theta = \left\{\theta : \int h(x)e^{\langle\theta,T(x)\rangle}d\mu(x) < \infty\right\}$$

es convexo.

En la Proposición 1, el correspondiente *espacio de parámetros naturales*. Hemos expresado esto en términos de la notación θ, pero la idea básica es partir de (1) y tomar $\eta(\theta)$ como el parámetro; es decir, simplemente hacer una reparametrización. El resultado establece que el espacio de parámetros naturales es un conjunto "bien comportado".

El siguiente teorema es útil para una serie de cálculos, en particular, para calcular momentos en familias exponenciales o la información de Fisher. La demostración es una aplicación interesante del teorema de convergencia dominada presentado anteriormente.

Teorema 2.1

Sea $X \in \mathbb{X} \subseteq \mathbb{R}^d$ con densidad $p_\theta(x) = a(\theta)e^{\langle\theta,x\rangle}$ con respecto a μ. Sea $\varphi : \mathbb{X} \to \mathbb{R}$ una función μ-medible y defina el conjunto

$$\Theta_\varphi = \left\{\theta : \int |\varphi(x)|e^{\langle\theta,x\rangle}d\mu(x) < \infty\right\}.$$

Entonces, para θ en el interior de Θ_φ, se define

$$m(\theta) := \int \varphi(x)e^{\langle\theta,x\rangle}d\mu(x)$$

y esta función es continua y tiene derivadas continuas de todo orden. Además, la derivada puede tomarse dentro del signo integral; es decir, para $i = 1, \ldots, d$, se cumple

$$\frac{\partial m(\theta)}{\partial \theta_i} = \int x_i \varphi(x) e^{\langle \theta, x \rangle} d\mu(x).$$

Demostración. Consideremos primero el caso de θ unidimensional; el caso general en dimensión d es similar. Sea $\theta \in \Theta_\varphi$ un valor fijo. Para un $\varepsilon > 0$ adecuado, definimos

$$d_n(x) = \frac{e^{\varepsilon x/n} - 1}{\varepsilon/n},$$

de modo que

$$\frac{m(\theta + \varepsilon/n) - m(\theta)}{\varepsilon/n} = \int \varphi(x) e^{\theta x} d_n(x) d\mu(x).$$

Es claro que $d_n(x) \to d(x) = x$ para cada x, donde $d(x)$ es la derivada de la función $z \mapsto e^{zx}$ en $z = 0$.
Definimos $f_n(x) = \varphi(x) e^{\theta x} d_n(x)$, por lo que $f_n(x) \to x\varphi(x) e^{\theta x}$ cuando $n \to \infty$ para todo x. Resta demostrar que la μ-integral de f_n converge a la μ-integral de f. Para ello, aplicamos el teorema de convergencia dominada.
Notamos las siguientes desigualdades para la función exponencial:

$$|e^z - 1| \leqslant |z| e^{|z|} \quad \text{y} \quad |z| \leqslant e^{|z|}, \quad \forall z.$$

Con estas desigualdades, podemos escribir

$$|f_n(x)| \leqslant |\varphi(x)| e^{\theta x} \varepsilon^{-1} e^{2\varepsilon |x|} \leqslant |\varphi(x)| e^{\theta x} \varepsilon^{-1} (e^{2\varepsilon x} + e^{-2\varepsilon x}).$$

Si elegimos ε de modo que $\theta \pm 2\varepsilon$ pertenezca a Θ_φ, entonces la cota superior, digamos $g(x)$, es μ-integrable. Por lo tanto, el teorema de convergencia dominada establece que

$$\frac{dm(\theta)}{d\theta} = \lim_{n \to \infty} \int f_n(x) d\mu(x) = \int f(x) d\mu(x) = \int x\varphi(x) e^{\theta x} d\mu(x).$$

Es decir, la derivada de la integral es la integral de la derivada, como se quería demostrar. Para demostrar que se pueden tomar más derivadas, y que estas derivadas adicionales pueden evaluarse tomando la derivada dentro de la integral, basta repetir el argumento anterior. ■

Existen un par de supuestos ocultos que vale la pena mencionar. Primero, hemos supuesto implícitamente que el soporte $\{x : p_\theta(x) > 0\}$ de la distribución no depende de θ. Esto descarta casos como Unif$(0, \theta)$. Segundo, también hemos supuesto implícitamente que Θ tiene un interior no vacío. Esto asegura que es posible encontrar un intervalo/caja abierto, dependiendo de ε, centrado en el valor dado θ y que encaje dentro de Θ. Sin esto, el enunciado podría ser falso. Estos dos supuestos forman parte de aquellas çondiciones de regularidad"mencionadas en las definiciones clásicas de familias exponenciales.

El mismo resultado es válido para familias exponenciales generales, no solo para aquellas en forma natural o canónica. El mensaje clave es que, para funciones φ suficientemente bien comportadas, el valor esperado de $\varphi(X)$ es una función bien comportada del parámetro θ. Como aplicación del Teorema 1, tenemos lo siguiente.

Corolario 2.1

Supongamos que $X = (X_1, \ldots, X_d)$ tiene una densidad de familia exponencial de la forma

$$p_\theta(x) = a(\theta)e^{\langle \theta, x\rangle}.$$

Entonces, para $i, j = 1, \ldots, d$, se cumple que

$$\mathbb{E}_\theta(X_i) = -\frac{\partial}{\partial \theta_i}\log a(\theta) \quad \text{y} \quad \text{Cov}_\theta(X_i, X_j) = -\frac{\partial^2}{\partial \theta_i \partial \theta_j}\log a(\theta).$$

Los momentos de orden superior de X pueden determinarse de manera similar.

Demostración. Partimos de la identidad

$$\int a(\theta)e^{\langle x, \theta\rangle} d\mu(x) = 1, \quad \text{para todo } \theta.$$

Ahora diferenciamos ambos lados con respecto a θ tantas veces como sea necesario, intercambiamos la derivada con la integral y resolvemos para el momento adecuado. ■

Podemos reescribir este resultado en una forma quizás más familiar reconociendo que $-\log a(\theta) = A(\theta)$. En este caso, por ejemplo, se tiene que

$$\mathbb{E}_\theta(X_i) = \frac{\partial}{\partial \theta_i}A(\theta).$$

Una derivación alternativa de este resultado se puede obtener notando que las familias exponenciales admiten una función generadora de momentos, dada por

$$M_\theta(u) = e^{A(\theta+u)-A(\theta)} = \frac{a(\theta)}{a(\theta+u)}. \tag{3}$$

2.3. Estadísticos Suficientes

2.3.1. Definición y el Teorema de Factorización

Un estadístico es simplemente una función medible de los datos; es decir, si $T : \mathbb{X} \to \mathbb{T}$ es medible, entonces $T(X)$ es un *estadístico*. Sin embargo, no todos los estadísticos serán útiles para el problema de inferencia estadística. El objetivo de esta sección es comprender qué tipo de funciones T son relevantes. La definición involucra distribuciones condicionales generales.

Definición 2.2

Suponga que $X \sim P_\theta$. Entonces, el estadístico $T = T(X)$, que mapea $(\mathbb{X}, \mathcal{A})$ a $(\mathbb{T}, \mathcal{B})$, es suficiente para $\{P_\theta : \theta \in \Theta\}$ si la distribución condicional de X, dado $T = t$, es independiente de θ. Más precisamente, suponga que existe una función $K : \mathcal{A} \times \mathbb{T} \to [0, 1]$, independiente de θ, tal que $K(\cdot, t)$ es una medida de probabilidad en $(\mathbb{X}, \mathcal{A})$ para

cada t, y $K(A,\cdot)$ es una función medible para cada $A \in \mathcal{A}$, con

$$P_\theta(X \in A, T \in B) = \int_B K(A,t)\, dP_\theta^T(t), \quad \forall A \in \mathcal{A}, B \in \mathcal{B}.$$

Aquí, P_θ^T denota la distribución marginal de T. Entonces, T es un Estadístico Suficiente para θ.

La clave de la definición anterior es que la probabilidad condicional de X, dado $T = t$, caracterizada por el núcleo $K(\cdot, t)$, no depende de θ. Por ejemplo, si φ es alguna función integrable, entonces

$$\mathbb{E}_\theta[\varphi(X) \mid T = t] = \int \varphi(x) dK(x,t)$$

no depende de θ. Llevando este argumento un paso más allá, se encuentra que la suficiencia implica que conocer el valor de T es suficiente para generar nuevos datos X', con las mismas propiedades probabilísticas que X.

Las dificultades de teoría de la medida que surgen con las distribuciones condicionales en el caso continuo hacen que la identificación de Estadísticos Suficientes mediante la definición sea complicada en estos casos; existe un método más práctico, el cual discutiremos en breve. Sin embargo, para problemas discretos, donde la condición es muy sencilla, la definición es adecuada.

Ejemplo 2.3

Suponga que $X_1, \ldots, X_n$ son variables iid Ber(θ). Definimos

$$T(X) = \sum_{i=1}^n X_i.$$

Entonces,

$$P_\theta(X = x \mid T(X) = t) = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}.$$

Esto es independiente de θ, por lo que $T(X)$ es un Estadístico Suficiente para θ. Aquí, $K(\cdot, t)$ es simplemente una distribución uniforme sobre todos los n-tuplas de 0's y 1's que contienen exactamente t 1's.

Ejemplo 2.4

Supongamos que $X_1, \ldots, X_n$ son iid Pois(θ). Definimos $T(X) = \sum_{i=1}^n X_i$. Entonces,

$$P_\theta(X_1 = x_1 \mid T(X) = t) = \frac{e^{-\theta}\theta^{x_1}/x_1! \cdot e^{-(n-1)\theta}[(n-1)\theta]^{t-x_1}/(t-x_1)!}{e^{-n\theta}(n\theta)^t/t!}$$

$$= \binom{t}{x_1}(1/n)^{x_1}(1-1/n)^{t-x_1}.$$

Es decir, la distribución condicional de X_1, dado $T = t$, es Bin($t, 1/n$). Esto se cumple para todos los X_i, no solo para X_1. De hecho, la distribución condicional del vector

X, dado $T = t$, es una distribución multinomial con tamaño t y pesos $1/n$. Como esta distribución es independiente de θ, se concluye que $T(X)$ es suficiente para θ.

Ejemplo 2.5

Sea $X_1, \ldots, X_n$ una muestra iid de una distribución P_θ. Definimos los *estadísticos de orden* $(X_{(1)}, \ldots, X_{(n)})$, que corresponden a la lista ordenada de los datos observados. Dado el orden, existen $n!$ posibles valores de X, y todos ellos tienen la misma probabilidad. Como esta probabilidad es independiente de P_θ, se concluye que los estadísticos de Orden deben ser suficientes para θ. No obstante, es importante notar que la suficiencia puede fallar si no se cumple la suposición de independencia idénticamente distribuida (iid).

En el caso en que la familia $\{P_\theta : \theta \in \Theta\}$ consista en distribuciones dominadas por una medida σ-finita común μ, existe una herramienta conveniente para identificar un estadístico suficiente.

Teorema 2.2 Teorema de Factorización de Neyman-Fisher

Sea $\{P_\theta : \theta \in \Theta\}$ dominada por una medida σ-finita común μ, con densidades $p_\theta = dP_\theta/d\mu$. Entonces, $T(X)$ es un estadístico suficiente para θ si y solo si existen funciones no negativas h y g_θ tales que

$$p_\theta(x) = g_\theta[T(x)]h(x) \quad \text{para todo } \theta, \quad \mu\text{-almost all } x. \tag{4}$$

Demostración. Para una demostración detallada, prestando especial atención a las preocupaciones de teoría de la medida sobre las condiciones, ver Keener (2010), Sec. 6.4. Sin embargo, la idea básica es bastante simple. Como T es un estadístico, una función de X, podemos ver aproximadamente que (i) la densidad conjunta de (X, T) es $g_\theta[T(x)]h(x)$, y (ii) la densidad marginal de T es $g_\theta(t)$. Entonces, la densidad condicional es el cociente de estas dos, y se observa que la dependencia en θ desaparece. Por lo tanto, la distribución condicional de X, dado T, no depende de θ, por lo que T es un estadístico suficiente. ■

Este teorema nos permite identificar fácilmente estadísticos suficientes, simplemente a través de manipulaciones algebraicas de la densidad conjunta o la función de verosimilitud.

Ejemplo 2.6

Supongamos que $X = (X_1, \ldots, X_n)$ consiste en muestras iid de Unif$(0, \theta)$. Entonces, la distribución conjunta se puede escribir como

$$p_\theta(x) = \prod_{i=1}^{n} \theta^{-1} I_{(0,\theta)}(x_i) = \theta^{-n} I_{(0,\theta)}(\max x_i).$$

Dado que $p_\theta(x)$ depende de θ solo a través de $T(X) = \max X_i$, se sigue del Teorema 2 que $T(X) = \max X_i$ es un estadístico suficiente para θ.

Ejemplo 2.7

Supongamos que $X = (X_1, \ldots, X_n)$ consiste en muestras iid de $\mathcal{N}(\mu, \sigma^2)$. La densidad conjunta es

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\mu^2 \right\}.$$

Por lo tanto, por el Teorema 2, $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ es un estadístico suficiente para (μ, σ^2). Equivalentemente, $T'(X) = (\bar{X}, s^2(X))$ también es un estadístico suficiente.

2.3.2. Estadísticos Suficientes Mínimos

Es claro que los estadísticos suficientes no son únicos; de hecho, en el Ejemplo 7 se identificaron dos estadísticos suficientes, y los estadísticos de orden también lo son, como es usual. Más generalmente, si T es suficiente, entonces también lo es $\psi(T)$ para cualquier función biyectiva ψ. Dicho esto, es deseable encontrar el estadístico suficiente que sea el "más pequeño" en algún sentido. Este *estadístico suficiente mínimo* $T = T_{\min}$ es aquel para el cual, dado cualquier otro estadístico suficiente U, existe una función h tal que $T = h(U)$. Una técnica poderosa para encontrar estadísticos suficientes mínimos es descrita en el siguiente teorema.

Teorema 2.3

Supongamos que, para cada $\theta \in \Theta$, P_θ tiene una densidad $p_\theta(x) = g_\theta[T(x)]h(x)$ con respecto a μ. Si

$$p_\theta(x) = cp_\theta(y), \quad \text{para alguna } c = c(x, y),$$

implica que $T(x) = T(y)$, entonces T es un estadístico suficiente mínimo.

Demostración. Ver Keener (2010), página 47. ■

Ejemplo 2.8

Supongamos que $X = (X_1, \ldots, X_n)$ es una muestra iid con densidad común

$$p_\theta(x) = h(x)e^{\sum_{j=1}^{d} \eta_j(\theta)T_j(x) - A(\theta)}. \tag{5}$$

Entonces, $T(X) = [T_1(X), \ldots, T_d(X)]$, con $T_j(X) = \sum_{i=1}^n T_j(X_i)$, es suficiente. Para ver que T es un estadístico suficiente mínimo (bajo cierta condición), aplicamos el Teorema 2.3. Tomemos x e y tales que $p_\theta(x) = cp_\theta(y)$ para alguna función $c = c(x, y)$. Esto implica que

$$\langle \eta(\theta), T(x) \rangle = \langle \eta(\theta), T(y) \rangle + c'$$

para algún $c' = c'(x, y)$. Tomando dos puntos θ_0 y θ_1 en Θ y restando, obtenemos

$$\langle \eta(\theta_0) - \eta(\theta_1), T(x) \rangle = \langle \eta(\theta_0) - \eta(\theta_1), T(y) \rangle,$$

lo que implica

$$\langle \eta(\theta_0) - \eta(\theta_1), T(x) - T(y) \rangle = 0,$$

es decir, $T(x) - T(y)$ y $\eta(\theta_0) - \eta(\theta_1)$ son ortogonales. Como θ_0 y θ_1 son arbitrarios, esto implica que $T(x) - T(y)$ debe ser ortogonal al espacio lineal generado por

$$S = \{\eta(\theta_0) - \eta(\theta_1) : \theta_0, \theta_1 \in \Theta\}.$$

Si S genera todo $\mathbb{R}^d$ (como se explica a continuación), entonces esto implica $T(x) = T(y)$ y, por lo tanto, T es un estadístico suficiente mínimo por el Teorema 3.

El resultado de la condición en el ejemplo anterior—que el espacio S abarque todo el espacio—es suficiente para demostrar que el Estadístico Suficiente natural, $T(X)$, en la familia exponencial es un Estadístico Suficiente mínimo. Sin embargo, esta condición por sí sola no es completamente satisfactoria. Primero, no es algo obvio de verificar y, en segundo lugar, algunas propiedades deseables, como la normalidad asintótica de los estimadores de máxima verosimilitud, requieren aún más regularidad. Por esta razón, a menudo imponemos una condición más fuerte, una que implique, en particular, que el espacio S mencionado anteriormente abarque todo el espacio.

Para formalizar esto, primero definimos que una familia exponencial de la forma (5) tiene *rango completo* si $\eta(\Theta)$ tiene interior no vacío y $[T_1(x), \ldots, T_d(x)]$ no satisface una restricción lineal para μ-almost all x. Reconocerás estas como condiciones de regularidad adicionales impuestas clásicamente en las familias exponenciales.

Si $\eta(\Theta)$ tiene interior no vacío, entonces también lo tiene

$$\{\eta(\theta_0) - \eta(\theta_1) : \theta_0, \theta_1 \in \Theta\},$$

lo que implica que la envoltura del espacio S en el ejemplo anterior llena todo el espacio. De hecho, si Θ contiene un conjunto abierto y η es una función continua biyectiva, entonces $\eta(\Theta)$ también contiene un conjunto abierto.

Es un ejercicio útil considerar cómo una colección de vectores que contiene un conjunto abierto implicaría que su envoltura abarca todo el espacio. Para esto, consideremos un conjunto abierto en dos dimensiones. Tomemos un vector $v = (v_1, v_2)$ en este conjunto abierto. Que el conjunto sea abierto significa que existe un $\varepsilon > 0$ suficientemente pequeño tal que $\tilde{v} = (v_1 + \varepsilon, v_2 + \varepsilon)$ también pertenece al conjunto.

Se afirma que, siempre que $v_1 \neq v_2$, el par de vectores $(v, \tilde{v})$ es linealmente independiente. Desde el álgebra lineal, existe una prueba de independencia lineal basada en el determinante de la matriz formada por la superposición de estos vectores. En este caso, se tiene:

$$\det \begin{pmatrix} v_1 & v_1 + \varepsilon \\ v_2 & v_2 + \varepsilon \end{pmatrix} = v_1 v_2 + v_1 \varepsilon - v_1 v_2 - v_2 \varepsilon = \varepsilon(v_1 - v_2).$$

Por supuesto, si $v_1 \neq v_2$, entonces este determinante no puede ser cero, por lo que $(v, \tilde{v})$ son linealmente independientes. Finalmente, un par de vectores linealmente independientes en dos dimensiones constituye una base y, por lo tanto, su envoltura abarca todo el espacio.

La mayoría de las familias exponenciales que conocemos tienen rango completo, por ejemplo: normal, binomial, Poisson, gamma, etc. Un ejemplo clásico de una familia exponencial que no tiene rango completo es la $N(\theta, \theta^2)$, que es una de las llamadas *familias exponenciales curvadas* (Keener 2010, Cap. 5). El término *curvado* proviene del hecho de que el espacio del parámetro natural es una curva o, más generalmente, un conjunto cuya dimensión efectiva es menor que la dimensión real. En este caso, el parámetro natural $\eta(\theta)$ está dado por

$$\eta_1(\theta) = \frac{1}{\theta} \quad \text{y} \quad \eta_2(\theta) = -\frac{1}{2\theta^2}.$$

Dado que $\eta_2 = -\eta_1^2/2$, es claro que el espacio del parámetro natural $\eta(\Theta)$ tiene la forma de una parábola invertida. Como el subconjunto unidimensional del espacio bidimensional no puede contener un conjunto abierto, concluimos que esta familia exponencial curvada no puede tener rango completo. Sin embargo, el estadístico suficiente natural $T = (T_1, T_2) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ sigue siendo mínimo suficiente. Para ver esto, necesitamos verificar que el conjunto S definido en el Ejemplo 8 abarque $\mathbb{R}^2$. Tomemos dos pares de puntos (x_1, y_1) y (x_2, y_2) y consideremos los dos vectores de diferencias:

$$v_j = \left(x_j - y_j, \frac{1}{2}(y_j^2 - x_j^2)\right)^\top, \quad j = 1, 2.$$

Coloquemos los dos vectores en una matriz, es decir,

$$\begin{pmatrix} x_1 - y_1 & x_2 - y_2 \\ \frac{1}{2}(y_1^2 - x_1^2) & \frac{1}{2}(y_2^2 - x_2^2) \end{pmatrix}.$$

Si, por ejemplo, tomamos $x_1 = 1$ y $y_1 = -1$, entonces el determinante de la matriz es igual a $y_2^2 - x_2^2$. Por lo tanto, en el caso $x_1 = 1$ y $y_1 = -1$, mientras $x_2 \neq \pm y_2$, el determinante es distinto de cero, los vectores son linealmente independientes y su espacio generado abarca todo el espacio. Por lo tanto, el estadístico suficiente natural T anterior es mínimo suficiente, incluso si la familia exponencial no es de rango completo.

La reducción de dimensión a través de la suficiencia simplifica enormemente las cosas. Por lo tanto, es interesante preguntar en qué problema es posible una reducción tan sustancial de la dimensión. Resulta que, esencialmente, esto solo ocurre en el caso de la familia exponencial. Como última parte de terminología, digamos que una familia de distribuciones admite un *estadístico suficiente continuo de dimensión k* si la factorización (4) es válida para todo x (no solo para casi todo x) y si $U(x) = [U_1(x), \ldots, U_k(x)]$ es continuo en x. El siguiente teorema se encuentra en Lehmann y Casella (1998, Cap. 1.6).

Teorema 2.4 Caracterización de Suficiencia

Suponga que $X_1, \ldots, X_n$ son valores reales e iid de una distribución con densidad continua $f_\theta(x)$ con respecto a la medida de Lebesgue, soportada en un intervalo $\mathbb{X}$ que no depende de θ. Denote la densidad conjunta por

$$p_\theta(x) = f_\theta(x_1) \cdots f_\theta(x_n),$$

y suponga que existe un estadístico suficiente continuo de dimensión k. Entonces:

- Si $k = 1$, entonces (5) es válido para algunas funciones h, η_1 y A.
- Si $k > 1$ y si $f_\theta(x_i)$ tiene derivadas parciales continuas con respecto a x_i, entonces (5) es válido para algún $d \leqslant k$.

Este teorema indica que, entre aquellas familias absolutamente continuas suaves con soporte fijo, esencialmente las únicas que admiten un estadístico suficiente continuo son las familias exponenciales. Nótese que el teorema no dice nada sobre aquellos problemas irregulares donde el soporte depende de θ; de hecho, la familia $\text{Unif}(0, \theta)$ admite un estadístico suficiente unidimensional para todo n.

2.3.3. Estadísticos Completos y Auxiliares

Hay Estadísticos Suficientes de diferentes dimensiones; por ejemplo, si $X_1, \ldots, X_n$ son iid $N(\theta, 1)$, entonces los estadísticos de orden y la media $\bar{X}$ son ambos suficientes. La capacidad de un Estadístico Suficiente para admitir una reducción significativa parece estar relacionada con la cantidad de información Auxiliar que contiene.

Un estadístico $U(X)$ se dice que es Auxiliar si su distribución es independiente de θ. Los estadísticos Auxiliares en sí mismos no contienen información sobre θ, pero incluso los Estadísticos Suficientes mínimos pueden contener información Auxiliar. Por ejemplo, en el Ejercicio 7, el Estadístico Suficiente mínimo no es completo.

El hecho de que los estadísticos Auxiliares no contengan información sobre θ no significa que no sean útiles. Una sugerencia común, aunque no universalmente aceptada o utilizada, es realizar análisis condicionales sobre los valores de los estadísticos Auxiliares. Condicionar en algo que no contiene información sobre θ no causa dificultades lógicas, y Fisher argumentó que condicionar en estadísticos Auxiliares es una forma ingeniosa de dar un significado más relevante a la inferencia del problema en cuestión.

Intuitivamente, esto restringe el espacio muestral a un "subconjunto relevante"—el conjunto donde $U(X) = u_{\text{obs}}$—acercando la inferencia condicionada a la observación de X. Esta idea se usa a menudo cuando, por ejemplo, una estimación de máxima verosimilitud no es mínima suficiente.

Un estadístico T es completo si

$$\mathbb{E}_\theta\{f(T)\} = 0 \quad \text{para todo } \theta \text{ implica } f = 0 \text{ casi en todas partes.}$$

En otras palabras, no hay funciones no constantes de T que sean auxiliares.

Alternativamente, un Estadístico Suficiente Completo es aquel que contiene exactamente toda la información sobre θ en X; es decir, no contiene información redundante sobre θ, ya que cada característica $f(T)$ de T tiene información sobre θ. Para ver cómo esto se relaciona con la definición formal, notemos que ninguna función no nula de T es auxiliar.

Los Estadísticos Suficientes Completos son especialmente efectivos para reducir los datos; de hecho, los Estadísticos Suficientes Completos son mínimos.

Teorema 2.5

Si T es completo y suficiente, entonces T también es mínimo suficiente.

Demostración. Sea T' un Estadístico Suficiente Mínimo. Por minimalidad, tenemos $T' = f(T)$ para alguna función f. Escribimos $g(T') = \mathbb{E}_\theta(T \mid T')$, que no depende de θ por suficiencia de T'. Además, por la expectativa iterada, $\mathbb{E}_\theta g(T') = \mathbb{E}_\theta(T)$. Por lo tanto,

$$\mathbb{E}_\theta\{T - g(T')\} = 0 \quad \text{para todo } \theta$$

y, dado que $T - g(T') = T - g(f(T))$ es una función de T, la completitud implica que $T = g(T')$ casi en todas partes. Como $T = g(T')$ y $T' = f(T)$, se concluye que T y T' son equivalentes salvo transformaciones uno a uno; por lo tanto, T también es mínimo suficiente. ■

Dado el poder de un Estadístico Suficiente Completo, es útil poder identificar casos en los que existe. No es sorprendente que las familias exponenciales admitan un Estadístico Suficiente Completo.

Teorema 2.6

Si X se distribuye como una familia exponencial d-dimensional de rango completo con densidad (5), entonces $[T_1(X), \ldots, T_d(X)]$ es completo.

Demostración. Esto es solo un esquema en un caso simple; para la demostración detallada en el caso general, ver Brown (1986, Teorema 2.12). Consideremos el caso unidimensional con

$$p_\theta(x) = e^{\theta x - A(\theta)}$$

y medida dominante μ. Entonces $T(x) = x$. Sea $f(x)$ una función integrable con $\mathbb{E}_\theta\{f(X)\} = 0$ para todo θ. Escribiendo la forma integral de la esperanza, se obtiene

$$\int f(x)e^{\theta x} d\mu(x) = 0 \quad \forall\theta.$$

La integral es esencialmente la transformada de Laplace de f. La transformada de Laplace de la función cero es constante e igual a cero y, dado que las transformadas de Laplace son únicas (μ-a.e.), se sigue que f debe ser la función cero (μ-a.e.). Por lo tanto, X es completo. ■

Ejemplo 2.9

El Teorema 6 muestra que $T(X) = \sum_{i=1}^n X_i$ es completo cuando $X_1, \ldots, X_n$ es una muestra iid de Ber(θ), Pois(θ) y N($\theta, 1$).

Ejemplo 2.10

Sea $X_1, \ldots, X_n$ una muestra iid de N(θ, θ^2). Se mostró anteriormente que $T = (T_1, T_2) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ es minimalmente suficiente. Sin embargo, no es completo. Para ver esto, consideremos la función $f(t_1, t_2) = t_1^2 - \frac{n+1}{2} t_2$. Entonces,

$$\begin{aligned}
\mathbb{E}_\theta f(T_1, T_2) &= \mathbb{E}_\theta(T_1^2) - \frac{n+1}{2}\mathbb{E}_\theta(T_2) \\
&= n\theta^2 + (n\theta)^2 - \frac{n+1}{2} \cdot 2n\theta^2 \\
&= 0 \quad \forall\theta.
\end{aligned}$$

Dado que esta función no es exactamente cero para $T = (T_1, T_2)$ pero tiene esperanza cero, el estadístico T no es completo. Esto no contradice el Teorema 6 porque esta familia exponencial curva no tiene rango completo, es decir, el espacio de parámetros naturales es una curva unidimensional en el plano bidimensional y, por lo tanto, no contiene un conjunto abierto.

Ejemplo 2.11

Sea $X_1, \ldots, X_n$ una muestra iid de Unif$(0, \theta)$. Se afirma que $T(X) = X_{(n)}$ es completo. Un cálculo directo muestra que la densidad de T es

$$p_\theta(t) = nt^{n-1}/\theta^n, \quad 0 < t < n.$$

Supongamos que $\mathbb{E}_\theta\{f(T)\} = 0$ para todo θ. Entonces, tenemos

$$\int_0^\theta t^{n-1} f^+(t)dt = \int_0^\theta t^{n-1} f^-(t)dt \quad \forall \theta > 0.$$

Dado que esto se cumple para los intervalos de integración $[0, \theta]$ para todo θ, también debe cumplirse para todos los intervalos $[a, b]$. El conjunto de todos los intervalos genera la σ-álgebra de Borel, por lo que, en efecto,

$$\int_A t^{n-1} f^+(t)dt = \int_A t^{n-1} f^-(t)dt \quad \text{para todos los conjuntos de Borel } A.$$

Por lo tanto, f debe ser cero casi en todas partes, y así T es completo.

De acuerdo con el teorema de Basu, no tiene sentido condicionar sobre estadísticos auxiliares (como se describió brevemente al comienzo de esta sección) en los casos en los que el estadístico suficiente es completo.

Teorema 2.7 Basu

Si T es un estadístico suficiente y completo para $\{P_\theta : \theta \in \Theta\}$, entonces cualquier estadístico auxiliar U es independiente de T.

Demostración. Dado que U es auxiliar, la probabilidad $p_A = P_\theta(U \in A)$ no depende de θ para cualquier conjunto A. Definimos la distribución condicional $\pi_A(t) = P_\theta(U \in A \mid T = t)$; por expectativa iterada,

$$\mathbb{E}_\theta\{\pi_A(T)\} = p_A \quad \text{para todo } A \text{ y para todo } \theta.$$

Por lo tanto, por completitud, $\pi_A(t) = p_A$ para casi todo t. Dado que la distribución condicional $\pi_A(t)$ de U, dado $T = t$, no depende de t, las dos variables deben ser independientes. ■

Ejemplo 2.12

El teorema de Basu se puede usar para demostrar que la media y la varianza de una muestra independiente de N(μ, σ^2) son independientes. Supongamos primero que σ^2 es conocido y es igual a 1. Sabemos que la media muestral $\bar{X}$ es un estadístico suficiente y completo para μ, y también que la varianza muestral $s^2(X) = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ es auxiliar. Por lo tanto, el teorema de Basu establece que $\bar{X}$ y $s^2(X)$ son independientes. Pero esto fue para $\sigma^2 = 1$, ¿cómo extenderlo al caso de σ^2 desconocido? La clave es que el caso general de σ^2 corresponde a una transformación de escala simple de los datos, lo que claramente no puede alterar la estructura de correlación entre $\bar{X}$ y $s^2(X)$. Por lo tanto, $\bar{X}$ y $s^2(X)$ son independientes para todos los valores de (μ, σ^2).

Ejemplo 2.13

Supongamos que $X_1, \ldots, X_n$ es una muestra iid de $\mathrm{N}(0,1)$, y sea $\bar{X}$ y M la media muestral y la mediana muestral, respectivamente. El objetivo es calcular la covarianza entre $\bar{X}$ y M.
Introducimos un parámetro de media ξ; al hacerlo, encontramos que $\bar{X}$ es un estadístico suficiente y completo, mientras que $\bar{X} - M$ es auxiliar. Luego, el teorema de Basu establece que $\bar{X}$ y $\bar{X} - M$ son independientes, y por lo tanto:

$$0 = C(\bar{X}, \bar{X} - M) = V(\bar{X}) - C(\bar{X}, M) \quad \Longrightarrow \quad C(\bar{X}, M) = n^{-1}.$$

Es común en los cursos de teoría estadística y en los libros de texto dar la impresión de que el teorema de Basu es solo un truco para realizar ciertos cálculos, como en los dos ejemplos anteriores. Sin embargo, la verdadera contribución del teorema de Basu es el punto mencionado anteriormente sobre la condición de los Estadísticos Auxiliares.

2.4. Información de Fisher

2.4.1. Definición

Entendemos coloquialmente que un Estadístico Suficiente contiene toda la información en $X_1, \ldots, X_n$ sobre el parámetro de interés θ. El concepto de información de Fisher hará esto más preciso.

Definición 2.3

Suponga que θ es d-dimensional y que $p_\theta(x)$ es la densidad de X con respecto a μ. Entonces, se cumplen las siguientes *condiciones de regularidad de la Información de Fisher*:

1. $\frac{\partial p_\theta(x)}{\partial \theta_i}$ existe μ-c.t.p. para cada i.
2. $\int p_\theta(x) d\mu(x)$ puede diferenciarse dentro del signo integral.
3. El soporte de p_θ es el mismo para todo θ.

Definición 2.4 Función Score e Información de Fisher

Suponga que se cumplen las condiciones de regularidad de la Información de Fisher. El vector score se define como $\frac{\partial \log p_\theta(X)}{\partial \theta_i}$ para $i = 1, \ldots, d$. La Información de Fisher $I_X(\theta)$ es la matriz de covarianza del vector score; es decir,

$$I_X(\theta)_{ij} = C_\theta \left(\frac{\partial \log p_\theta(X)}{\partial \theta_i}, \frac{\partial \log p_\theta(X)}{\partial \theta_j} \right). \tag{6}$$

Se tiene que el valor esperado del score es cero. En este caso, si podemos diferenciar dos veces dentro del signo integral (como en familias exponenciales; cf. Teorema 1), entonces hay una fórmula alternativa para la Información de Fisher:

$$I_X(\theta)_{ij} = -\mathbb{E}_\theta \left\{ \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p_\theta(X) \right\}.$$

Si $X_1, \ldots, X_n$ son iid de una distribución que satisface las condiciones de regularidad de la Información de Fisher, entonces es fácil demostrar que $I_{X_1,\ldots,X_n}(\theta) = nI_{X_1}(\theta)$. Es decir, la información se acumula a medida que se reciben más datos, lo cual tiene sentido si se pretende medir la información en un conjunto de datos. Si los datos no son iid, entonces la información sigue aumentando, pero a una tasa no mayor que en el caso iid.

Cabe mencionar que las condiciones de regularidad de la Información de Fisher no son necesarias aquí. En particular, requerir que $\theta \mapsto p_\theta(x)$ sea diferenciable para todo x es demasiado restrictivo. La Información de Fisher puede definirse bajo la condición mucho menos estricta de *diferenciabilidad en media cuadrática*.

2.4.2. Suficiencia e información

El siguiente resultado ayuda a interpretar que los estadísticos suficientes contienen toda la información relevante sobre θ.

Teorema 2.8

Supongamos que se cumplen las condiciones de regularidad de la Información de Fisher. Supongamos que θ es d-dimensional y que P_θ está dominada por μ. Si $T = g(X)$ es un estadístico, entonces $I_X(\theta) - I_T(\theta)$ es semidefinida positiva. La matriz es nula si y solo si T es suficiente.

Demostración. Como T es una función de X, la distribución conjunta está determinada por la distribución marginal de X. En particular,

$$p_\theta^{X|T}(x \mid t) = \begin{cases} p_\theta^X(x)/p_\theta^T(t) & \text{si } T(x) = t \\ 0 & \text{en otro caso.} \end{cases}$$

Aquí utilizamos p_θ para todas las densidades, y los superíndices indican la distribución. Por lo tanto,

$$p_\theta^X(x) = p_\theta^{X,T}(x,t) = p_\theta^T(t) p_\theta^{X|T}(x \mid t), \quad \text{si } T(x) = t.$$

Tomando logaritmos, obtenemos

$$\frac{\partial \log p_\theta^X(X)}{\partial\theta_i} = \frac{\partial \log p_\theta^T(T)}{\partial\theta_i} + \frac{\partial \log p_\theta^{X|T}(X \mid T)}{\partial\theta_i}, \quad \forall\theta. \tag{7}$$

Mostraremos que los dos términos en el lado derecho son incorrelacionados y que el último término es cero si y solo si T es suficiente. Usando la expectativa iterada,

$$\begin{aligned} C_\theta\left(\frac{\partial \log p_\theta^T(T)}{\partial\theta_i}, \frac{\partial \log p_\theta^{X|T}(X|T)}{\partial\theta_i} \right) &= \mathbb{E}_\theta \left\{ \frac{\partial \log p_\theta^T(T)}{\partial\theta_i} \frac{\partial \log p_\theta^{X|T}(X|T)}{\partial\theta_i} \right\} \\ &= \mathbb{E}_\theta \left\{ \frac{\partial \log p_\theta^T(T)}{\partial\theta_i} \mathbb{E}_\theta \left(\frac{\partial \log p_\theta^{X|T}(X|T)}{\partial\theta_i} \Big| T \right) \right\}. \end{aligned}$$

Afirmamos que la expectativa condicional interna es cero con probabilidad P_θ^T igual a 1. Para mostrar esto, primero notemos que

$$1 = \int p_\theta^{X|T}(x \mid t)\, d\mu(x) \implies 0 = \frac{\partial}{\partial \theta_i} \int p_\theta^{X|T}(x \mid t)\, d\mu(x),$$

para todo t fuera de un conjunto nulo bajo P_θ^T. Si podemos intercambiar la última derivada y la expectativa condicional, hemos terminado. Dado que

$$\int p_\theta^{X|T}(x \mid t)\, d\mu(x) = \frac{1}{p_\theta^T(t)} \int_{\{x:T(x)=t\}} p_\theta^X(x)\, d\mu(x),$$

tomamos la derivada con respecto a θ_i y simplificamos.

$$\frac{1}{p_\theta^T(t)} \frac{\partial}{\partial \theta_i} \int_{\{x:T(x)=t\}} p_\theta^X(x)\, d\mu(x) - \frac{\partial}{\partial \theta_i} \log p_\theta^T(t). \tag{8}$$

La restricción en el rango de integración no nos impide intercambiar la derivada y la integral de p_θ^X (como en FI). Así, obtenemos

$$\begin{aligned}\frac{\partial}{\partial \theta_i} \int_{\{x:T(x)=t\}} p_\theta^X(x)\, d\mu(x) &= \int_{\{x:T(x)=t\}} \left[\frac{\partial}{\partial \theta_i} \log p_\theta^T(t) + \frac{\partial}{\partial \theta_i} \log p_\theta^{X|T}(x|t)\right] p_\theta^X(x)\, d\mu(x) \\ &= p_\theta^T(t) \frac{\partial}{\partial \theta_i} \log p_\theta^T(t) + p_\theta^T(t) \int \frac{\partial}{\partial \theta_i} \log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\, d\mu(x).\end{aligned}$$

Este último cálculo muestra que (8) se simplifica a

$$\int \frac{\partial}{\partial \theta_i} \log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\, d\mu(x).$$

Por lo tanto,

$$\frac{\partial}{\partial \theta_i} \int p_\theta^{X|T}(x \mid t)\, d\mu(x) = \int \frac{\partial}{\partial \theta_i} \log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\, d\mu(x),$$

y dado que podemos intercambiar la derivada y la integral con respecto a la distribución condicional, obtenemos que la esperanza condicional de la función de score condicional es cero (con P_θ^T-probabilidad 1). Esto, a su vez, muestra que los dos términos en el lado derecho de (7) son incorrelacionados. Luego, la matriz de covarianza de la suma en el lado derecho de (7) es la suma de las respectivas matrices de covarianza. Por lo tanto,

$$I_X(\theta) = I_T(\theta) + \mathbb{E}_\theta\{I_{X|T}(\theta)\},$$

y es claro que $I_X(\theta) - I_T(\theta)$ es semidefinida positiva. La matriz $\mathbb{E}_\theta\{I_{X|T}(\theta)\}$ es cero si y solo si el score condicional $\frac{\partial}{\partial \theta_i} \log p_{X|T,\theta}(X|T)$ es constante (debe ser cero, ¿verdad?) en θ o, en otras palabras, T es suficiente. ■

Esto formaliza la afirmación hecha en la introducción de que los estadísticos suficientes T preservan toda la información sobre θ en los datos X. Es decir, en el caso unidimensional, se tiene $I_T(X) \leqslant I_X$ con igualdad si y solo si T es suficiente.

2.4.3. Desigualdad de Cramer–Rao

Hemos visto que la información de Fisher proporciona una justificación para la afirmación de que los estadísticos suficientes contienen toda la información relevante en una muestra. Sin embargo, la información de Fisher juega un papel aún más profundo en la inferencia estadística; en particular, está involucrada en muchos resultados de optimalidad que proporcionan una referencia para la comparación entre estimadores, pruebas, etc. A continuación, se presenta un resultado familiar pero importante, que establece que, bajo ciertas condiciones, la varianza de un estimador no puede ser menor que un límite que involucra la información de Fisher.

Teorema 2.9 Cramer–Rao

Para simplificar, tomemos θ como un escalar y supongamos que p_θ satisface las condiciones de regularidad FI. Sea $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta$ y sea $T = T(X_1, \ldots, X_n)$ un estadístico real con $\mathbb{E}_\theta(T) = g(\theta)$. Entonces,

$$V_\theta(T) \geqslant \{g'(\theta)\}^2 \{nI(\theta)\}^{-1}.$$

Demostración. La covarianza entre T y la función de score es $g'(\theta)$. Dado que la varianza del score es $nI(\theta)$, la desigualdad de Cauchy–Schwarz nos da

$$g'(\theta)^2 \leqslant V_\theta(T)\{nI(\theta)\}.$$

Despejando $V_\theta(T)$, se obtiene el resultado deseado. ■

Una aplicación de la desigualdad de Cramer–Rao se encuentra en el diseño experimental. En estos problemas, se tiene control sobre ciertas entradas y el objetivo es seleccionar dichas entradas de manera que el estimador tenga, por ejemplo, la menor varianza posible. En estos casos, la estrategia consiste en elegir esas entradas de tal forma que la información de Fisher se maximice, lo cual tiene una cierta intuición basada en Cramer–Rao, es decir, la varianza es pequeña si la información de Fisher es grande.

2.4.4. Otras medidas de información

¿Es la información de Fisher la única medida de información? Técnicamente, la respuesta es NO, existen otras medidas, pero no reciben tanta atención. La razón es que la información de Fisher es la elección "correcta" siempre que el modelo satisfaga las condiciones de regularidad de FI. Dado que la mayoría de los modelos (por ejemplo, las familias exponenciales regulares) las satisfacen, no hay mucha razón para buscar más allá de la información de Fisher. Sin embargo, existen modelos que no satisfacen las condiciones de regularidad, como la $\text{Unif}(0, \theta)$. En tales casos, la información de Fisher no está definida, por lo que obviamente no puede usarse. La pregunta es si existe algún otro tipo de información, una que se reduzca a la información de Fisher cuando esta existe, pero que sea más versátil en el sentido de que pueda definirse cuando la información de Fisher no puede.
Para extender la información de Fisher, es útil comprender de dónde proviene. Recordemos la divergencia de Kullback–Leibler del Capítulo 1, que (aproximadamente) mide la distancia entre dos modelos. Consideremos aquí dos modelos con funciones de densidad p_θ y $p_{\theta+\varepsilon}$, respecto a la misma medida dominante μ, donde el segundo representa un pequeño cambio en el parámetro. Kullback(1997) muestra que la divergencia de Kullback–Leibler de $p_{\theta+\varepsilon}$ desde p_θ es aproximadamente cuadrática en ε, en particular,

$$K(p_\theta, p_{\theta+\varepsilon}) \approx \varepsilon^\top I(\theta)\varepsilon,$$

cuando $\varepsilon \to 0$, donde $I(\theta)$ es la matriz de información de Fisher mencionada anteriormente. Entonces, la idea clave para generalizar la matriz de información de Fisher es reconocer que, fuera de los casos regulares, la divergencia de Kullback–Leibler o, mejor aún, la divergencia de Hellinger, definida como

$$h(\theta, \theta') = \int \left(p_\theta^{1/2} - p_{\theta'}^{1/2}\right)^2 d\mu,$$

no es cuadrática en ε. Sin embargo, esta misma expansión puede realizarse y el coeficiente define una adecuada "información de Hellinger." Por ejemplo, consideremos el caso de la distribución $\text{Unif}(0, \theta)$. La divergencia de Hellinger es

$$h(\theta+\varepsilon, \theta) = \int \left[\frac{1}{\sqrt{\theta+\varepsilon}} I_{(0,\theta+\varepsilon)}(x) - \frac{1}{\sqrt{\theta}} I_{(0,\theta)}(x)\right]^2 d\mu(x) = \cdots = \frac{\varepsilon}{\theta} + o(\varepsilon). \tag{9}$$

Esto tiene una aproximación lineal en lugar de cuadrática, resultado de la no regularidad de la distribución $\text{Unif}(0, \theta)$. Sin embargo, una "información de Hellinger" para $\text{Unif}(0, \theta)$ puede definirse como θ^{-1}. También existen versiones de la cota de Cramer–Rao para la información de Hellinger, pero no se presentarán aquí.

2.5. Condicionamiento

Aquí discutimos algunos ejemplos interesantes en los cuales el enfoque frecuentista clásico da respuestas extrañas. Estos ejemplos se utilizarán para motivar el condicionamiento en la inferencia.

Ejemplo 2.14

Suponga que X_1 y X_2 son iid con distribución P_θ que satisface

$$P_\theta(X = \theta - 1) = P_\theta(X = \theta + 1) = 0{,}5, \quad \theta \in \mathbb{R}.$$

El objetivo es construir un intervalo de confianza para el desconocido θ. Considere

$$C = \begin{cases} \{\bar{X}\} & \text{si } X_1 \neq X_2, \\ \{X_1 - 1\} & \text{si } X_1 = X_2. \end{cases}$$

Para ser claros, en cualquier caso, C es un conjunto unitario. Se puede demostrar que C tiene un nivel de confianza del 75 %. Pero analicemos este procedimiento con más cuidado. A partir de la estructura del problema, si $X_1 \neq X_2$, entonces una observación es $\theta - 1$ y la otra es $\theta + 1$. En este caso, $\bar{X}$ es exactamente igual a θ, por lo que, *dado* que $X_1 \neq X_2$, C es garantizado como correcto. Por otro lado, si $X_1 = X_2$, entonces C es $\{\theta\}$ con probabilidad 0.5 y $\{\theta - 2\}$ con probabilidad 0.5, por lo que, *dado* que $X_1 = X_2$, C es correcto con probabilidad 0.5. Juntando esto, C tiene confianza del 100 % cuando $X_1 \neq X_2$ y del 50 % cuando $X_1 = X_2$. En promedio, la confianza es del 75 %, pero dado que, para un problema específico, sabemos en qué caso nos encontramos, ¿no tendría sentido reportar la *confianza condicional* de 100 % o 50 %?

Ejemplo 2.15

Suponga que los datos X pueden tomar valores en $\{1, 2, 3\}$ y que $\theta \in \{0, 1\}$. La distribución de probabilidad de X para cada θ está descrita en la siguiente tabla.

x	1	2	3
$p_0(x)$	0,0050	0,0050	0,99
$p_1(x)$	0,0051	0,9849	0,01

La prueba más poderosa de nivel $\alpha = 0{,}01$ de $H_0 : \theta = 0$ contra $H_1 : \theta = 1$ se basa en la razón de verosimilitud $p_0(x)/p_1(x)$ para un valor observado de $X = x$. Se puede demostrar que esta prueba tiene una potencia de 0.99, lo que sugiere que hay una gran confianza en la decisión basada en el valor observado de x. Pero, ¿es esto cierto? Si se observa $X = 1$, entonces la razón de verosimilitud es $0{,}005/0{,}0051 \approx 1$. En general, una razón de verosimilitud cercana a 1 no da una fuerte preferencia ni por H_0 ni por H_1, por lo que medir nuestra certeza sobre la decisión del procedimiento usando la medida "global" de potencia podría ser engañoso.

Ejemplo 2.16

Considere el siguiente experimento: lanzar una moneda justa y, si la moneda cae en cara, entonces tomar $X \sim \mathrm{N}(\theta, 1)$; de lo contrario, tomar $X \sim \mathrm{N}(\theta, 99)$. Suponga que el resultado del lanzamiento de la moneda es *conocido*. El objetivo es usar X para estimar θ. ¿Qué distribución deberíamos usar para construir un intervalo de confianza, por ejemplo? La varianza marginal de X es $(1 + 99)/2 = 50$. Sin embargo, esto parece una mala representación del error real en X como estimador de θ, ya que en realidad sabemos si X fue muestreado de $\mathrm{N}(\theta, 1)$ o de $\mathrm{N}(\theta, 99)$. Entonces la pregunta es, ¿por qué no usar la varianza "condicional", dado el resultado del lanzamiento de la moneda? Esto es algo intuitivamente natural de hacer, pero esto *no* es lo que el frecuentismo sugiere hacer.

Ejemplo 2.17

Sea (X_{1i}, X_{2i}), $i = 1, \ldots, n$ una muestra bivariada iid de una distribución con densidad $p_\theta(x_1, x_2) = e^{-\theta x_1 - x_2/\theta}$, donde x_1, x_2 y θ son todos positivos. Se puede demostrar que el estadístico suficiente mínimo es $T = (T_1, T_2)$, donde $T_j = \sum_{i=1}^{n} X_{ji}$, $j = 1, 2$. Note que el estadístico suficiente mínimo es bidimensional, mientras que el parámetro es unidimensional. Para estimar θ, una elección razonable es $\hat{\theta} = \{T_2/T_1\}^{1/2}$, el estimador de máxima verosimilitud. Sin embargo, este no es un estadístico suficiente mínimo, por lo que debemos elegir si debemos condicionar o no. Un estadístico auxiliar para condicionar es $A = \{T_1 T_2\}^{1/2}$. Como se discute en Ghosh et al. (2010), la información de Fisher incondicional en T y en $\hat{\theta}$, respectivamente, son

$$I_T(\theta) = \frac{2n}{\theta^2} \quad \text{y} \quad I_{\hat{\theta}}(\theta) = \frac{2n}{\theta^2}\frac{2n}{2n+1};$$

por supuesto, como se esperaba, $I_T(\theta) > I_{\hat{\theta}}(\theta)$. Sin embargo, la información de Fisher condicional es

$$I_{\hat{\theta}|A}(\theta) = I_T(\theta)\frac{K_1(2A)}{K_0(2A)}, \tag{10}$$

donde K_0 y K_1 son funciones de Bessel. Una gráfica de la razón—llamada $r(A)$—en el lado derecho de la ecuación anterior, como función de $A = a$, se muestra en la Figura

2.1. Cuando A es grande, $r(A)$ está cerca de 1, por lo que $I_{\hat{\theta}|A}(\theta) \approx I_T(\theta)$. Sin embargo, si A no es grande, entonces la información condicional puede ser mucho mayor y, dado que una mayor información es "mejor", podemos ver que en este caso hay una ventaja en condicionar.

El propósito de los ejemplos anteriores es destacar las desventajas del frecuentismo puro. Al menos en algunos ejemplos, hay una razón clara para considerar el condicionamiento sobre algo: a veces es evidente sobre qué condicionar (Ejemplo 16) y otras veces no lo es (Ejemplo 17). La inferencia condicional se da cuando las distribuciones muestrales se basan en distribuciones condicionales de estimadores dados los valores observados de un estadístico auxiliar; por ejemplo, en el Ejemplo 14, $|X_1 - X_2|$ es un estadístico auxiliar.

Cuando el estimador es un Estadístico Suficiente completo, el teorema de Basu establece que no hay necesidad de condicionar. Pero en problemas donde el estimador no es un Estadístico Suficiente completo (Ejemplo 17), hay necesidad de condicionar. Existen amplias discusiones en la literatura sobre la inferencia condicional, por ejemplo, Fraser (2004) y Ghosh et al. (2010); Berger (2014) proporciona una discusión más reciente. A pesar de los beneficios de la inferencia condicional, este enfoque no ha permeado realmente la estadística aplicada. Esto se debe a algunas dificultades técnicas adicionales, tanto en la identificación de un estadístico auxiliar adecuado como en la implementación efectiva del condicionamiento. Una revisión aplicada de la inferencia condicional y temas relacionados se encuentra en Brazzale et al. (2007).

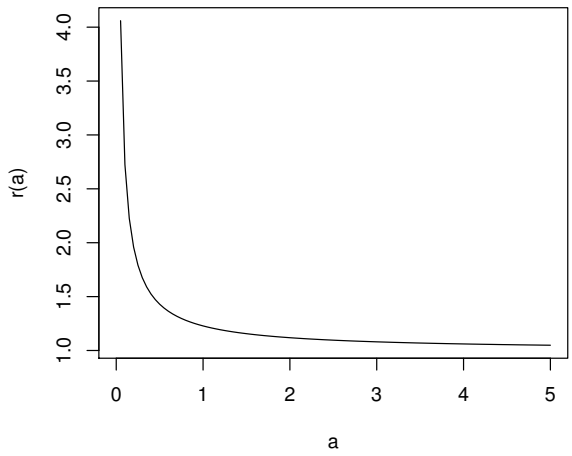

Figura 1: Gráfica de la razón $r(a)$ en el lado derecho de (10) como una función de $A = a$, el estadístico auxiliar.

2.6. Discusión

2.6.1. Modelos lineales generalizados

Una aplicación importante de las distribuciones de la familia exponencial son los llamados *modelos lineales generalizados* (GLM por sus siglas en inglés). Estos modelos generalizan los modelos lineales usuales (por ejemplo, regresión y análisis de varianza) que se presentan en

los cursos introductorios de metodología estadística. Este tema típicamente no aparece en un curso como este—no hay mención de los GLM en Keener (2010)—y creo que la razón de su omisión es que los detalles de la teoría pueden entenderse en el contexto más simple de la familia exponencial, como se discutió en la Sección 2, dejando los detalles específicos de modelado y computación para otros cursos/textos.

Sin embargo, creo que es importante que los estudiantes tengan al menos una exposición mínima a esta aplicación de los modelos de la familia exponencial en este curso teórico, aunque solo sea para que sepan de la existencia de estos temas y puedan leer más por su cuenta o tomar cursos más especializados. Aquí doy una breve explicación de los GLM con algunos ejemplos.

Consideremos un problema con dos variables: Y es llamada la *variable de respuesta* y X la *variable predictora* o *covariable*, donde X es, por ejemplo, d-dimensional. El modelo lineal usual establece que, dado $X = x$, la media de la variable respuesta Y es una función lineal de x, es decir, $\mathbb{E}(Y \mid X = x) = x^\top \beta$, donde β es un parámetro d-dimensional de coeficientes. Si tenemos muestras independientes, es decir, $\{(X_i, Y_i) : i = 1, \ldots, n\}$, entonces el modelo establece que, dado $X_i = x_i$, los Y_i son independientes con media $\mu_i = x_i^\top \beta$, $i = 1, \ldots, n$. Un punto clave es que β es el mismo para cada i; además, este es uno de los modelos independientes pero no iid más comunes que los estudiantes verán. El método de mínimos cuadrados puede utilizarse para estimar β basándose en las observaciones, y esta solución tiene muchas propiedades deseables que no profundizaremos aquí.

Una cuestión a considerar es la siguiente: ¿es este tipo de modelo lineal siempre adecuado? Es decir, ¿debería la media de la distribución de la variable de respuesta (condicionada a la variable predictora X) expresarse como una función lineal del predictor? Como ejemplo, consideremos el caso donde Y sigue una distribución de Poisson o Bernoulli. En ambos casos, la media de la distribución tiene una restricción—en $(0, \infty)$ en un caso y en $(0, 1)$ en el otro—por lo que una función lineal, que no tiene restricciones y puede tomar valores en $(-\infty, \infty)$, podría no ser adecuada. Un modelo lineal generalizado (GLM) puede abordar esto sin extenderse demasiado fuera del marco de un modelo lineal.

Supongamos que las variables de respuesta $Y_1, \ldots, Y_n$ son independientes con densidades

$$p_{\theta_i}(y_i) = h(y_i) e^{\eta(\theta_i) y_i - A(\theta_i)}, \quad i = 1, \ldots, n,$$

que tiene la forma de la familia exponencial descrita en la Sección 2, pero con un parámetro diferente θ_i para cada punto de datos. Supongamos que existe cierta estructura común, en particular, que la media $\mu_i = \mathbb{E}_\theta(Y_i)$ satisface la condición $g(\mu_i) = x_i^\top \beta$ para alguna función uno-a-uno y suave g, llamada *función de enlace*. Cuando la función de enlace es tal que $\eta(\theta_i) = x_i^\top \beta$, se denomina *enlace canónico*. El resultado de esta construcción es una forma general de introducir un modelo lineal efectivo que conecta la variable de respuesta Y_i con la variable predictora X_i, pero evitando las limitaciones de un modelo lineal real.

Como ejemplo rápido, consideremos el caso donde $Y_i \sim \text{Poisson}(\theta_i)$, con $i = 1, \ldots, n$ independientes. Es fácil comprobar que la distribución de Poisson pertenece a la familia exponencial y que $\eta(\theta_i) = \log \theta_i$. Dado que θ_i es también la media de Y_i, si queremos construir un GLM de Poisson con enlace canónico, entonces $g(u) = \log u$, por lo que

$$\theta_i = e^{x_i^\top \beta} \quad \Longleftrightarrow \quad \log \theta_i = x_i^\top \beta.$$

Esta última fórmula explica por qué este GLM de Poisson suele llamarse *modelo log-lineal*.

2.6.2. Un poco más sobre la condicionalidad

En los cursos y libros de teoría estadística, la suficiencia se trata como un aspecto críticamente importante de la inferencia estadística. Aquí quiero argumentar que no hay nada realmente especial sobre los estadísticos suficientes, siempre que se realice una condicionalidad apropiada. El mensaje aquí es que la condicionalidad es un concepto más fundamental que la suficiencia.

Voy a argumentar este punto usando un ejemplo simple. Sea X_1, X_2 una muestra iid de $N(\theta, 1)$. Un estimador razonable de θ es $\bar{X} = (X_1 + X_2)/2$, un estadístico suficiente, cuya distribución muestral es $N(\theta, 1/2)$. Por otro lado, considere el estimador $\hat{\theta} = X_1$, el cual no es un estadístico suficiente. Consideraciones clásicas sugieren que la inferencia basada en X_1 es peor que aquella basada en $\bar{X}$. Sin embargo, considere la distribución muestral condicional de X_1, dado $X_2 - X_1$. Es fácil verificar que

$$X_1 \mid (X_2 - X_1) \sim N\left(\theta + \frac{X_2 - X_1}{2}, 1/2\right),$$

y, por ejemplo, los intervalos de confianza basados en esta distribución condicional son los mismos que aquellos basados en la distribución muestral marginal del estadístico suficiente $\bar{X}$.

Por lo tanto, en este problema, se podría argumentar que no hay nada realmente especial sobre el estadístico suficiente $\bar{X}$, ya que uno puede obtener esencialmente la misma distribución muestral usando otro estadístico no suficiente, siempre que se realice una condicionalidad adecuada. El resultado aquí es más general (ver Ejercicio 20), aunque la continuidad parece ser importante.

La suficiencia, cuando proporciona algo significativo, puede ser conveniente, ya que la condicionalidad no es necesaria, lo que ahorra algo de esfuerzo. Sin embargo, hay casos en los que la suficiencia no proporciona ninguna mejora. Por ejemplo, en el problema de localización de Student-t con grados de libertad conocidos, los datos completos constituyen el estadístico suficiente mínimo. Sin embargo, se puede obtener fácilmente un estimador razonable (equivariante a la localización), como la media muestral, y condicionar en el invariante máximo, un estadístico auxiliar. El punto es que la condicionalidad funciona cuando la suficiencia no lo hace, y, aun cuando la suficiencia funcione, la condicionalidad puede ser igual de buena. Por lo tanto, argumentaría que la condicionalidad es más fundamental que la suficiencia.

2.7. Ejercicios

1. La desigualdad de Hölder es una generalización de la desigualdad de Cauchy-Schwartz.

Sea $1 \leqslant p, q \leqslant \infty$ números tales que $\frac{1}{p} + \frac{1}{q} = 1$. Sean f y g funciones tales que f^p y g^q son integrables respecto a μ. Entonces,

$$\int |fg|\, d\mu \leqslant \left(\int |f|^p \, d\mu \right)^{1/p} \left(\int |g|^q \, d\mu \right)^{1/q}.$$

La desigualdad de Cauchy-Schwartz corresponde al caso $p = q = 2$.

Utilice la desigualdad de Hölder para probar la Proposición 1.

2. Suponga que X sigue una distribución de familia exponencial con densidad

$$p_\theta(x) = h(x) e^{\eta(\theta) T(x) - A(\theta)}.$$

Derive las fórmulas de la media y la varianza:

$$\mathbb{E}_\theta[T(X)] = \frac{A'(\theta)}{\eta'(\theta)}, \quad V_\theta[T(X)] = \frac{A''(\theta)}{[\eta'(\theta)]^2} - \frac{\eta''(\theta) A'(\theta)}{[\eta'(\theta)]^3}.$$

3. Demuestre la ecuación (3), una fórmula para la función generadora de momentos de la familia exponencial.

4. Una variable aleatoria discreta con función de masa de probabilidad

$$p_\theta(x) = a(x)\theta^x / C(\theta), \quad x \in \{0, 1, \dots\}; \quad a(\theta) \geqslant 0, \quad \theta > 0$$

sigue una *distribución de serie de potencias*.

   a) Demuestre que la distribución de serie de potencias es una familia exponencial.

   b) Demuestre que las distribuciones binomial y de Poisson son casos especiales de distribuciones de serie de potencias.

5. *a*) Demuestre la identidad de Stein. Para $X \sim N(\mu, \sigma^2)$, sea φ una función diferenciable con $\mathbb{E}_\theta|\varphi'(X)| < \infty$. Entonces,

$$\mathbb{E}[\varphi(X)(X - \mu)] = \sigma^2 \mathbb{E}[\varphi'(X)].$$

   [**Pista:** Sin pérdida de generalidad, suponga $\mu = 0$ y $\sigma = 1$. Use integración por partes. Necesitará mostrar que $\varphi(x) e^{-x^2/2} \to 0$ cuando $x \to \pm\infty$. También existe un enfoque que usa el teorema de Fubini.]

   b) Sea $X \sim N(\mu, \sigma^2)$. Use la identidad de Stein para encontrar los primeros cuatro momentos, $\mathbb{E}(X^k)$, con $k = 1, 2, 3, 4$.
   [**Pista:** Para $\mathbb{E}(X^k)$, use $\varphi(x) = x^{k-1}$.]

6. Demuestre que una función uno a uno de un estadístico suficiente minimal también es un estadístico suficiente minimal.

7. Suponga que $X_1, \dots, X_n$ son iid $N(\theta, \theta^2)$.

   a) Muestre que $N(\theta, \theta^2)$ tiene la forma de una familia exponencial.

   b) Encuentre el estadístico suficiente minimal para θ.

   c) Muestre que su estadístico suficiente minimal no es completo.

8. La familia Inversa Gaussiana, denotada por $\text{IG}(\lambda, \mu)$, tiene la función de densidad

$$(\lambda/2\pi)^{1/2} \exp\{(\lambda\mu)^{1/2}\} x^{-3/2} \exp\{-(\lambda x^{-1} + \mu x)/2\}, \quad x > 0; \quad \lambda, \mu > 0.$$

   a) Mostrar que $\text{IG}(\lambda, \mu)$ es una familia exponencial.

   b) Mostrar que $\text{IG}(\lambda, \mu)$ es invariante respecto al grupo de transformaciones de escala, es decir, $\mathcal{G} = \{g_c(x) = cx : c > 0\}$.

   c) Sean

$$T_1(X) = n^{-1} \sum_{i=1}^{n} X_i, \quad T_2(X) = \sum_{i=1}^{n} (1/X_i - 1/T_1(X)).$$

   Mostrar que (T_1, T_2) es completo y suficiente.

   d) Mostrar que $T_1 \sim \text{IG}(n\lambda, n\mu)$.

9. Suponga que los pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ son una muestra iid de una distribución normal bivariada, donde

$$\mathbb{E}(X_1) = \mathbb{E}(Y_1) = 0, \quad \mathbb{V}(X_1) = \mathbb{V}(Y_1) = 1, \quad \text{y} \quad \mathbb{E}(X_1 Y_1) = \theta.$$

   Aquí, $\theta \in (-1, 1)$ es la correlación entre X y Y.

   a) Encuentre un estadístico suficiente minimal (bidimensional) para θ.

   b) Demuestre que el estadístico suficiente minimal no es completo.

   c) Sea $Z_1 = \sum_{i=1}^{n} X_i^2$ y $Z_2 = \sum_{i=1}^{n} Y_i^2$. Demuestre que tanto Z_1 como Z_2 son auxiliares, pero que (Z_1, Z_2) no lo es.

10. Este ejercicio describe un enfoque alternativo para encontrar estadísticos suficientes minimales. Está relacionado con el dado en el Teorema 3.

   a) Demuestre el siguiente teorema:

   Considere una familia finita de distribuciones con densidades $p_0, p_1, \ldots, p_K$, todas con el mismo soporte. Entonces

$$T(X) = \left(\frac{p_1(X)}{p_0(X)}, \frac{p_2(X)}{p_0(X)}, \ldots, \frac{p_K(X)}{p_0(X)} \right)$$

   es suficiente minimal.

   b) Demuestre el siguiente teorema: Sea $\mathbb{P}$ una familia paramétrica de distribuciones con soporte común, y sea $\mathbb{P}_0$ un subconjunto de $\mathbb{P}$. Si T es suficiente minimal para $\mathbb{P}_0$ y suficiente para $\mathbb{P}$, entonces es suficiente minimal para $\mathbb{P}$.

   c) Use los dos resultados anteriores para demostrar que, para la familia $\text{Pois}(\theta)$, el estadístico

$$T = \sum_{i=1}^{n} X_i$$

   es suficiente minimal.

   Pista: Elija un subconjunto de dos elementos $\mathbb{P}_0 = \{p_0 = \text{Pois}(\theta_0), p_1 = \text{Pois}(\theta_1)\}$ de $\mathbb{P} = \{\text{Pois}(\theta) : \theta > 0\}$.

11. *a)* Considere una familia de localización con densidades $p_\theta(x) = p(x-\theta)$, donde $\theta \in \mathbb{R}$. Para $X \sim p_\theta$, demuestre que la información de Fisher para θ es

$$I_X(\theta) = \int_{-\infty}^{\infty} \frac{[p'(x)]^2}{p(x)} dx,$$

la cual es independiente de θ.

b) Considere una familia de escala con $p_\theta(x) = p(x/\theta)/\theta$, con $\theta > 0$. Para $X \sim p_\theta$, demuestre que la información de Fisher para θ es

$$I_X(\theta) = \frac{1}{\theta^2} \int \left[\frac{xp'(x)}{p(x)} + 1\right]^2 p(x)\, dx.$$

12. Para cada caso, encuentre la información de Fisher basada en una sola observación X.

(a) Ber(θ).

(b) Pois(θ).

(c) Cau($\theta, 1$).

(d) N($0, \theta$), donde $\theta > 0$ denota la varianza.

13. Para $X_1, \ldots, X_n$ iid, demuestre que $I_{X_1,\ldots,X_n}(\theta) = nI_{X_1}(\theta)$.

14. Sea p_θ una densidad que satisface las condiciones de regularidad FI, y sea $T = T(X_1, \ldots, X_n)$ con $\mathbb{E}_\theta(T) = g(\theta)$. Demuestre que $C_\theta(T, U_\theta) = g'(\theta)$, donde $U_\theta = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log p_\theta(X_i)$ es la función de puntuación.

15. Suponga que la información de Fisher en X sobre θ es $I_X(\theta)$, donde θ es un escalar. Sea ξ una reparametrización suave y uno a uno de θ, y escriba $\tilde{I}_X(\xi)$ para la información de Fisher en X sobre ξ. Demuestre que $\tilde{I}_X(\xi) = \left(\frac{d\theta}{d\xi}\right)^2 I_X(\theta)$. Generalice al caso de vectores θ y ξ.

16. Sea $X \sim \mathcal{N}_n(\theta, \Sigma)$ una única muestra normal n-dimensional; aquí, la matriz de covarianza Σ es conocida, pero el vector θ es desconocido.

(a) Encuentre la matriz de información de Fisher $I_X(\theta)$.

(b) Suponga que $\theta = D\xi$, donde D es una matriz $n \times p$ de rango p, con $p < n$, y ξ es un vector desconocido $p \times 1$. Aquí, D es la *matriz de diseño*. Utilice el resultado en el Ejercicio 15 para encontrar la información de Fisher $\tilde{I}_X(\xi)$ en X sobre ξ.

(La matriz de información en la parte (b) depende de la matriz de diseño D, y la teoría de diseños óptimos busca elegir D para hacer $\tilde{I}_X(\xi)$ lo "más grande posible". Por supuesto, la información de Fisher aquí es una matriz, por lo que se debe definir qué significa que una matriz sea grande, pero la intuición es perfectamente clara.)

17. Sea $\{p_\theta : \theta \in \Theta\}$ una clase de μ-densidades que satisfacen las condiciones de regularidad de la información de Fisher (FI). Mediante la permutación de la diferenciación e integración, derive una aproximación de Taylor de dos términos para la función $\eta \mapsto K(p_\theta, p_\eta)$, para η cercano a θ, donde K es la divergencia de Kullback–Leibler.

18. Let $Y_i \sim \text{Ber}(\theta_i)$, $i = 1, \ldots, n$, independent.

(a) Show that the Bernoulli model is an exponential family with $\eta(\theta) = \log \frac{\theta}{1-\theta}$.

(b) Find the canonical link and write down the formula for θ_i in terms of a predictor variable x_i and a parameter β like in Section 6.1.

(c) Look up the "logistic distribution" (e.g., on **wikipedia**) to see why they call this Bernoulli GLM with canonical link *logistic regression*.

19. Sean X_1, X_2 iid Unif$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

(a) Demuestre que $A = (X_2 - X_1)/2$ es un estadístico auxiliar.

(b) Encuentre la distribución de $\bar{X}$, dado $A = a$.

(c) Compare $\mathbb{V}(\bar{X})$ y $\mathbb{V}(\bar{X} \mid A = a)$.

(Ver el Ejemplo 2.2 en Fraser (2004) para una ilustración diferente de este ejemplo: allí se muestra que los intervalos de confianza incondicionales "óptimos"para θ son inútiles, mientras que el intervalo de confianza condicional es muy razonable.)

20. Sean X_1, X_2 iid exponenciales con media θ.

(a) Encuentre la distribución de $\bar{X} = (X_1 + X_2)/2$.

(b) Encuentre la distribución de X_1, dado X_2/X_1.

(c) Compare los intervalos de confianza obtenidos de las distribuciones en (a) y (b).