# Chapter 3

# Estimación puntual

## 3.1 Introducción

En este capítulo introduciremos algunos conceptos de la teoría de estimación puntual. Los resultados que se desarrollarán, se aplican al problema de ajustar distribuciones de probabilidad a los datos. Muchas familias de distribuciones, como la normal, $N(\mu, \sigma^2)$, o la Poisson, $P(\lambda)$, dependen de un número finito de parámetros y salvo que éstos se conozcan de antemano, deben ser estimados para conocer aproximadamente la distribución de probabilidad.

Consideremos el siguiente problema de inferencia estadística paramétrica. Supongamos se ha observado un vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ de cuya distribución sólo se conoce que pertenece a una familia $\mathcal{F} = \{F(x_1, x_2, \ldots, x_n, \boldsymbol{\theta})$ donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta \subset I\!R^p\}$. Supongamos que interese conocer *aproximadamente* $q(\boldsymbol{\theta})$, donde $q(\boldsymbol{\theta})$ es una función de $\Theta$ en $I\!R$. La única información que se tiene sobre $\boldsymbol{\theta}$ es el vector $\mathbf{X}$, por lo tanto cualquier estimación que se haga de $\boldsymbol{\theta}$, deberá estar basada en $\mathbf{X}$. Un *estimador puntual* de $q(\boldsymbol{\theta})$ será cualquier estadístico $\delta(\mathbf{X})$ de $I\!R^n$ en $I\!R$.

Un buen estimador $\delta(\mathbf{X})$ deberá tener la propiedad de que cualquiera sea el valor de $\boldsymbol{\theta}$, que es desconocido, la diferencia $\delta(\mathbf{X}) - q(\boldsymbol{\theta})$ sea pequeña. En qué sentido esta diferencia es pequeña será especificado más adelante.

Así en el ejemplo 1 de 2.4 se tenía para el problema (a) necesidad de estimar $q_1(\mu, \sigma^2) = \mu$ y $q_2(\mu, \sigma^2) = \sigma^2$, para el problema (b) se requería estimar $q(\mu, \sigma^2) = 1000\ \mu$. En cambio el problema (c) no era de estimación, ya que lo que se buscaba no era aproximar $q(\mu, \sigma^2)$ que vale 0 ó 1 según $\mu < 200$ ó $\mu \geq 200$, sino decidir si $q(\mu, \sigma^2)$ era 0 ó 1.

También podemos considerar problemas de estimación puntual no paramétrica. En este caso sólo se conoce que el vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ tiene una distribución $F(x_1, x_2, \ldots, x_n)$ perteneciente a una familia $\mathcal{F}$, pero esta familia no puede indicarse con un número finito de parámetros, y quiere estimarse una función $q(F)$ que va de $\mathcal{F}$ en $\mathbb{R}$. El ejemplo 2 de 2.4 es un ejemplo de este tipo.

El ejemplo 3 de 2.4 es otro ejemplo de estimación puntual paramétrica.

Comenzaremos describiendo distintos métodos de estimación que intuitivamente parecen razonables, su justificación queda diferida para más adelante.

## 3.2 Método de los momentos

Sea $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria de una familia de distribuciones $F(x, \theta)$, donde $\theta \in \Theta \subset \mathbb{R}$, y supongamos que se quiera estimar $\theta$.

Sea $g$ una función de $\mathbb{R}$ en $\mathbb{R}$, luego el método de los momentos estima $\theta$, por el valor $\widehat{\theta} = \delta(\mathbf{X})$ que satisface la ecuación

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = E_{\widehat{\theta}}(g(X_1)), \tag{3.1}$$

donde $E_\theta(X)$ significa la esperanza de $X$ cuando $X$ tiene la distribución $F(x, \theta)$. La justificación heurística de este método se basa en el hecho que de acuerdo a la ley de los grandes números

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to E_\theta(g(X_1)) \quad \text{c.t.p.}$$

y por lo tanto, si $\theta$ puede expresarse como una función continua de $E_\theta(g(X_1))$, se puede esperar que cuando $n$ es grande el valor $\widehat{\theta}$ que satisface la ecuación (3.1) estará cerca de $\theta$.

En general, se toman como funciones $g$ las funciones generadoras de momentos, ya que se supone que los parámetros de la distribución se relacionan con los momentos a través de alguna función continua.

**Ejemplo 1:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución de la cual sólo se conoce que está en la familia $N(\mu, 1)$. Usando el método

de los momentos y usando $g(x) = x$ se obtiene

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E_{\widehat{\mu}}(X_1) = \widehat{\mu}\ .$$

Luego $\widehat{\mu} = (1/n)\sum_{i=1}^{n} X_i$ es el estimador de $\mu$ resultante.

**Ejemplo 2:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $N(0, \sigma^2)$. Usando el método de los momentos con $g(x) = x^2$ se obtiene

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E_{\widehat{\theta}}(X_1^2) = \widehat{\sigma}^2\ .$$

Luego $\widehat{\sigma}^2 = \delta(X_1, \ldots, X_n) = (1/n)\sum_{i=1}^{n} X_i^2$ es el estimador de $\sigma^2$ resultante.

**Ejemplo 3:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $P(\lambda)$, usando la función $g_1(x) = x$ se obtiene como estimador de $\lambda$

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E_{\widehat{\lambda}}(X_i) = \widehat{\lambda}\ .$$

Luego el estimador de los momentos resultantes usando la función $g_1$ resulta

$$\widehat{\lambda}_1 = \delta_1(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i\ .$$

También podemos usar la función $g_2(x) = x^2$. Recordando que

$$E_\lambda(X_1^2) = \mathrm{Var}_\lambda(X_1) + (E_\lambda(X_1))^2 = \lambda + \lambda^2\ ,$$

obtenemos

$$\frac{1}{n}\sum_{i=1}^{n} X_1^2 = E_{\widehat{\lambda}}(X_1^2) = \widehat{\lambda} + \widehat{\lambda}^2\ ,$$

y resolviendo esta ecuación de segundo grado el valor resulta

$$\widehat{\lambda} = -\frac{1}{2} \pm \sqrt{\frac{1}{4} + \sum_{i=1}^{n} \frac{X_i^2}{n}}\ .$$

Como el parámetro $\lambda$ es positivo, la solución que interesa es la positiva. Luego el estimador correspondiente a $g_2$ vendrá dado por

$$\widehat{\lambda}_2 = \delta_2(X_1, X_2, \ldots, X_n) = -\frac{1}{2} + \sqrt{\frac{1}{4} + \sum_{i=1}^{n} \frac{X_i^2}{n}}$$

Luego observamos que eligiendo distintas funciones $g$, obtenemos diferentes estimadores. Todavía no estamos en condiciones de comparar uno con otro, por lo que dejamos este punto sin resolver hasta más adelante.

**Generalización cuando hay varios parámetros:** Supongamos que se tiene una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución perteneciente a la familia $\mathcal{F} = \{F(x, \theta_1, \theta_2, \ldots, \theta_p)$ con $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p) \in \Theta \subset I\!R^p\}$.

Para estimar $\theta_1, \theta_2, \ldots, \theta_p$ por el método de los momentos se procede como sigue: Se consideran $k$ funciones $g_1, g_2, \ldots, g_p$ de $I\!R$ en $I\!R$ y se resuelve el siguiente sistema

$$\frac{1}{n}\sum_{i=1}^{n} g_j(X_i) = E_{\widehat{\boldsymbol{\theta}}}(g_j(X_1)) \qquad j = 1, 2, \ldots, p\ .$$

**Ejemplo 4:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$. Consideremos $g_1(x) = x$ y $g_2(x) = x^2$. Como se tiene

$$E_{\mu,\sigma^2}(g_1(X_1)) = \mu \quad \text{y} \quad E_{\mu,\sigma^2}(g_2(X_1)) = \sigma^2 + \mu^2\ ,$$

para estimar $\mu$ y $\sigma^2$ se deberá resolver el sistema

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i &= \widehat{\mu} \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \widehat{\mu}^2 + \widehat{\sigma}^2\ . \end{aligned}$$

Luego, se tiene

$$\widehat{\mu} = \delta_1(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

y

$$\widehat{\sigma}^2 = \delta_2(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \widehat{\mu})^2$$

que coinciden con los estimadores que habíamos propuesto en el ejemplo 1 de 2.4.

**Ejemplo 5:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $\Gamma(\alpha, \lambda)$. Consideremos $g_1(x) = x$ y $g_2(x) = x^2$. Como se tiene

$$E_{\alpha,\lambda}(g_1(X_1)) = \frac{\alpha}{\lambda} \quad \text{y} \quad E_{\alpha,\lambda}(g_2(X_1)) = \frac{\alpha(\alpha+1)}{\lambda^2}\ ,$$

para estimar $\alpha$ y $\lambda$ se deberá resolver el sistema

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i &= \frac{\widehat{\alpha}}{\widehat{\lambda}} \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \frac{\widehat{\alpha}(\widehat{\alpha}+1)}{\widehat{\lambda}^2} \, . \end{aligned}$$

Indiquemos por $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y por $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Entonces, despejando del sistema anterior, los estimadores de los momentos para $\lambda$ y $\alpha$ resultan ser

$$\widehat{\lambda} = \delta_1(X_1, X_2, \ldots, X_n) = \frac{\overline{X}}{\widehat{\sigma}^2}$$

y

$$\widehat{\alpha} = \delta_2(X_1, X_2, \ldots, X_n) = \frac{\overline{X}^2}{\widehat{\sigma}^2} \, .$$

**Estimación de $q(\boldsymbol{\theta})$.** Si lo que interesa estimar es una función de $\boldsymbol{\theta}$, $q(\boldsymbol{\theta})$ y esta función es continua, el método de los momentos consistirá en estimar primero $\boldsymbol{\theta}$ por $\widehat{\boldsymbol{\theta}}$ y luego $q(\boldsymbol{\theta})$ se estimará por $q(\widehat{\boldsymbol{\theta}})$. La justificación de esto reside en que si $\widehat{\boldsymbol{\theta}}$ está próximo a $\boldsymbol{\theta}$, entonces como $q$ es continua, $q(\widehat{\boldsymbol{\theta}})$ estará próxima a $q(\boldsymbol{\theta})$.

## 3.3 Método de máxima verosimilitud

Supongamos que se observa un vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ discreto o continuo cuya función de densidad discreta o continua pertenezca a una familia $p(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ y se quiera estimar $\boldsymbol{\theta}$.

En el caso discreto $p(\mathbf{x}, \boldsymbol{\theta})$ representa la probabilidad de observar el vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, cuando el valor del parámetro es $\boldsymbol{\theta}$. Es razonable pensar que si hemos observado el vector $\mathbf{x}$, este tendrá alta probabilidad. Luego se podría estimar $\boldsymbol{\theta}$ como el valor que hace máxima $p(\mathbf{x}, \boldsymbol{\theta})$. Un razonamiento análogo se puede hacer en el caso continuo, recordando que la probabilidad de un hipercubo con centro en $\mathbf{x}$ y de arista $\Delta$, cuando $\Delta$ es pequeño tiene probabilidad aproximadamente igual $p(\mathbf{x}, \boldsymbol{\theta})\ \Delta^n$. Esto sugiere la siguiente definición:

**Definición 1:** Diremos $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ es un estimador de máxima verosimilitud (E.M.V.) de $\boldsymbol{\theta}$, si se cumple

$$p(\mathbf{X}, \widehat{\boldsymbol{\theta}}(\mathbf{X})) = \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{X}, \boldsymbol{\theta})$$

**Ejemplo 1:** Supongamos que $\theta$ puede tomar valores $\theta = 1$ ó $\theta = 0$ y que $p(x, \theta)$ viene dado por

| $x$ | $\theta$ | |
|---|---|---|
| | 0 | 1 |
| 0 | 0.3 | 0.6 |
| 1 | 0.7 | 0.4 |
| $\Sigma$ | 1 | 1 |

Supongamos que se observe una muestra de tamaño 1 con valor $X$. Luego el estimador de máxima verosimilitud viene dado por

$$\widehat{\boldsymbol{\theta}}(\mathbf{X}) = \begin{cases} 1 & \text{si } \mathbf{X} = 0 \\ 0 & \text{si } \mathbf{X} = 1 \end{cases}$$

**Cómputo del E.M.V.:** Supongamos ahora que $\Theta$ es un subconjunto abierto de $\mathbb{R}^p$, que el soporte de $p(\mathbf{x}, \boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$ y que $p(\mathbf{x}, \boldsymbol{\theta})$ tiene derivadas parciales respecto a todas las componentes $\theta_i$.

Como la función $\ln(\mu)$ (logaritmo natural) es monótona creciente, maximizar $p(\mathbf{x}, \boldsymbol{\theta})$ será equivalente a maximizar $\ln p(\mathbf{x}, \boldsymbol{\theta})$. Luego el E.M.V. $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ debe verificar:

$$\frac{\partial \ln p(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_i} = 0 \qquad i = 1, 2, \ldots, p \,. \tag{3.2}$$

Hasta ahora hemos supuesto que $\mathbf{X}$ es un vector con una distribución arbitraria. Supongamos ahora que $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ es una muestra aleatoria de una distribución discreta o continua con densidad $p(\mathbf{x}, \boldsymbol{\theta})$. Luego se tiene

$$p(\mathbf{x}, \boldsymbol{\theta}) = p(x_1, x_2, \ldots, x_n, \boldsymbol{\theta}) = \prod_{j=1}^{n} p(x_j, \boldsymbol{\theta})$$

y bajo las condiciones dadas anteriormente, el sistema de ecuaciones (3.2) se transforma en

$$\sum_{i=1}^{n} \frac{\partial \ln p(x_i, \widehat{\boldsymbol{\theta}})}{\partial \theta_j} = 0 \qquad j = 1, 2, \ldots, p \; . \tag{3.3}$$

Supongamos que indicamos por $\psi_j(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial \ln p(x, \widehat{\boldsymbol{\theta}})}{\partial \theta_j}$, entonces (3.3) puede escribirse como

$$\sum_{i=1}^{n} \psi_j(\mathbf{x}_i, \boldsymbol{\theta}) = 0 \qquad j = 1, 2, \ldots, p \; .$$

Esta ecuación corresponde a la forma general de los denominados $M-$estimadores, que veremos más adelante.

Por supuesto que tanto (3.2) como (3.3) son condiciones necesarias pero no suficientes para que $\boldsymbol{\theta}$ sea un máximo. Para asegurarse que $\widehat{\boldsymbol{\theta}}$ es un máximo deberían verificarse las condiciones de segundo orden respectivas. Además debe verificarse que no se trata de un máximo relativo sino absoluto.

**Ejemplo 2:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $Bi(\theta, k)$, con $k$ conocido, luego cada variable $X_i$ tiene función de densidad

$$p(x, \theta) = \left( \begin{array}{c} k \\ x \end{array} \right) \theta^x \, (1 - \theta)^{k-x}$$

y

$$\frac{\partial \ln p(x, \theta)}{\partial \theta} = \frac{x}{\theta} - \frac{k - x}{1 - \theta} = \frac{x - k\theta}{\theta(1 - \theta)} \; .$$

Luego (3.3) se transforma en la ecuación

$$\sum_{i=1}^{n} \frac{X_i - k\widehat{\theta}}{\widehat{\theta}(1 - \widehat{\theta})} = 0 \; ,$$

y despejando $\widehat{\theta}$ resulta

$$\widehat{\theta}(X_1, X_2, \ldots, X_n) = \frac{1}{nk} \sum_{i=1}^{n} X_i \; .$$

**Ejemplo 3:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$. Busquemos los E.M.V. de $\mu$ y $\sigma^2$. La función de densidad de cada variable $X_i$ es

$$p(x, \mu, \sigma^2) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) \, e^{-\frac{1}{2\,\sigma^2} (x - \mu)^2} \; .$$

Por lo tanto,

$$\frac{\partial \ln p(x,\mu,\sigma^2)}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

y

$$\frac{\partial \ln p(x,\mu,\sigma^2)}{\partial \sigma^2} = -\frac{1}{2\,\sigma^2} + (\sigma^2)^2\,\frac{1}{2}\,(x-\mu)^2\ .$$

Luego el sistema (3.3) se transforma en el sistema

$$\begin{aligned} \sum_{i=1}^{n}(X_i-\widehat{\mu})/\widehat{\sigma}^2 &= 0 \\ \sum_{i=1}^{n} -\frac{1}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}(X_i-\widehat{\mu})^2 &= 0 \end{aligned}$$

que tiene como solución

$$\begin{aligned} \widehat{\mu}(X_1,X_2,\ldots,X_n) &= \sum_{i=1}^{n} X_i/n = \overline{X} \\ \widehat{\sigma}^2(X_1,X_2,\ldots,X_n) &= \sum_{i=1}^{n}(X_i-\overline{X})^2/n \end{aligned}$$

que son los mismos estimadores que encontramos por el método de los momentos.

**Ejemplo 4:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $\Gamma(\alpha,\lambda)$. La densidad de $X_i$ está dada por

$$p(x,\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\,\lambda^{\alpha}\,x^{\alpha-1}\,e^{-\lambda\,x}\ ,$$

con lo cual

$$\frac{\partial \ln p(x,\alpha,\lambda)}{\partial \alpha} = \ln\lambda + \ln x - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

y

$$\frac{\partial \ln p(x,\alpha,\lambda)}{\partial \lambda} = \frac{\alpha}{\lambda} - x\ ,$$

donde $\Gamma'(\alpha)$ indica la derivada de la función $\Gamma(\alpha)$. Luego el sistema (3.3) se transforma en el sistema

$$\begin{aligned} n \ln \widehat{\lambda} + \sum_{i=1}^{n} \ln(X_i) - n \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} &= 0 \\ \frac{n\,\widehat{\alpha}}{\widehat{\lambda}} - n\,\overline{X} &= 0\,, \end{aligned}$$

con $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Luego $\widehat{\lambda} = \frac{\widehat{\alpha}}{\overline{X}}$. Pero, este sistema no tiene una solución explícita ya que al reemplazar el valor de $\widehat{\lambda}$ obtenemos la ecuación no lineal

$$n\left(\ln \widehat{\alpha} - \ln(\overline{X})\right) + \sum_{i=1}^{n} \ln(X_i) - n \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} = 0\,,$$

que puede resolverse, por ejemplo mediante, el algoritmo de Newton-Raphson. Para iniciar el proceso, se puede tomar como estimador inicial el estimador de los momentos, por ejemplo.

En este caso, el estimador de máxima verosimilitud no coincide con el estimador de los momentos.

**Invarianza de los E.M.V.** Supongamos que $\boldsymbol{\lambda} = q(\boldsymbol{\theta})$ es una función *biunívoca* de $\Theta$ sobre $\Lambda$, donde $\Lambda \subset I\!R^p$. Luego la densidad $p(\mathbf{x}, \boldsymbol{\theta})$ se puede expresar en función de $\boldsymbol{\lambda}$ ya que $\boldsymbol{\theta} = q^{-1}(\boldsymbol{\lambda})$. Denominemos a la densidad de $\mathbf{X}$ como función de $\boldsymbol{\lambda}$ por $p^*(\mathbf{x}, \boldsymbol{\lambda})$. Claramente se tiene

$$p^*(\mathbf{x}, \boldsymbol{\lambda}) = p(\mathbf{x}, q^{-1}(\boldsymbol{\lambda}))$$

Luego se definen los E.M.V. $\widehat{\boldsymbol{\theta}}$ y $\widehat{\boldsymbol{\lambda}}$ por

$$p(\mathbf{x}, \widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{x}, \boldsymbol{\theta}) \tag{3.4}$$

y

$$p^*(\mathbf{x}, \widehat{\boldsymbol{\lambda}}) = \max_{\lambda \in \Lambda} p^*(\mathbf{x}, \boldsymbol{\lambda}) \tag{3.5}$$

El siguiente teorema muestra que los estimadores de máxima verosimilitud son invariantes por transformaciones biunívocas.

**Teorema 1:** *Si $\widehat{\boldsymbol{\theta}}$ es E.M.V. de $\boldsymbol{\theta}$, entonces $\widehat{\boldsymbol{\lambda}} = q(\widehat{\boldsymbol{\theta}})$ es E.M.V. de $\boldsymbol{\lambda}$.*

DEMOSTRACIÓN: Como $\widehat{\boldsymbol{\theta}}$ es E.M.V. de $\boldsymbol{\theta}$ se tendrá que (3.4) vale. Como $\widehat{\lambda} = q(\widehat{\boldsymbol{\theta}})$, (3.4) se puede escribir como

$$p(\mathbf{x}, q^{-1}(\widehat{\boldsymbol{\lambda}})) = \max_{\boldsymbol{\lambda} \in \Lambda} p(\mathbf{x}, q^{-1}(\boldsymbol{\lambda}))$$

pero, esta ecuación de acuerdo a la definición de $p^*$ es equivalente a

$$p^*(\mathbf{x}, \widehat{\boldsymbol{\lambda}}) = \max_{\boldsymbol{\lambda} \in \Lambda} p^*(\mathbf{x}, \boldsymbol{\lambda}) \ ,$$

luego $\widehat{\boldsymbol{\lambda}}$ satisface (3.5) y por lo tanto es un E.M.V. de $\boldsymbol{\lambda}$.

**Ejemplo 5:** De acuerdo al Teorema 1, en el ejemplo 2, el E.M.V. de $\lambda = q(\theta) = \ln \theta$ será

$$\widehat{\lambda} = \ln \widehat{\theta} = \ln \left( \frac{\overline{X}}{k} \right) .$$

En general, si $\boldsymbol{\lambda} = q(\boldsymbol{\theta})$, aunque $q$ no sea bunívoca, se define el estimador de máxima verosimilitud de $\boldsymbol{\lambda}$ por

$$\widehat{\boldsymbol{\lambda}} = q(\widehat{\boldsymbol{\theta}}) \ .$$

**Ejemplo 6:** Supongamos que en el ejemplo 3 interese encontrar el E.M.V. de $\lambda = q(\mu, \sigma^2) = \mu/\sigma^2$. Aunque esta transformación no es biunívoca, el E.M.V. de $\lambda$ será

$$\widehat{\lambda} = q(\widehat{\mu}, \widehat{\sigma}^2) = \frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2 / n} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

pues basta completar la transformación dada a una transformación biunívoca, tomando por ejemplo, $q_1(\mu, \sigma^2) = \mu$.

## 3.4 Método de cuadrados mínimos

Supongamos que $X_1, X_2, \ldots, X_n$ son variables aleatorias de la forma

$$X_i = S_i(\theta_1, \ldots, \theta_p) + \varepsilon_i \qquad 1 \leq i \leq n \tag{3.6}$$

donde $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ es un vector de parámetros desconocido, del cual lo único que se conoce es que está en un conjunto $\Theta \subset I\!R^p$ y $\varepsilon_i$ son variables aleatorias con

(i) $E(\varepsilon_i) = 0$

(ii) $\mathrm{Var}(\varepsilon_i) = \sigma^2$

(iii) $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son variables aleatorias independientes.

**Ejemplo 1:** Consideremos el ejemplo 3 de 2.4. Luego, en este caso, poniendo $\theta_1$ en lugar de $a$ y $\theta_2$ en lugar de $b$, se tiene

$$X_i = \theta_1 G_i + \theta_2 + \varepsilon_i \qquad 1 \leq i \leq n$$

donde las variables $\varepsilon_i$ satisfacen (i), (ii) y (iii).

Luego si llamamos:

$$S_i(\theta_1, \theta_2) = \theta_1 G_i + \theta_2 \qquad 1 \leq i \leq n$$

estamos en la situación descripta por la ecuación (3.6).

**Ejemplo 2:** Podemos generalizar el ejemplo 1 por la siguiente situación. Supongamos que la variable $X$ depende de otras dos variables $G$ y $H$ y que la forma de la dependencia es

$$X = u(G, H, \theta_1, \theta_2, \ldots, \theta_p) + \varepsilon$$

donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ se conoce que pertenece a un conjunto $\Theta \subset I\!R^p$, y donde $\varepsilon$ es una variable aleatoria que aglutina todos los otros factores que determina $X$ y que son desconocidos.

Por ejemplo se pueden tener

$$u_1(G, H, \boldsymbol{\theta}) = \theta_1 G + \theta_2 H + \theta_3$$

o

$$u_2(G, H, \boldsymbol{\theta}) = \theta_1 G^2 + \theta_2 H^2 + \theta_3 HG + \theta_4 H + \theta_5 G + \theta_6$$

o

$$u_3(G, H, \boldsymbol{\theta}) = \theta_1 e^{\theta_2 G} + \theta_3 e^{\theta_4 H} \ .$$

Supongamos que se hagan $n$ experimentos. En el experimento i-ésimo se fijan $G$ y $H$ iguales respectivamente a $G_i$ y $H_i$ y se observa un valor $X_i$. Luego se tendrá

$$X_i = u(G_i, H_i, \theta_1, \theta_2, \ldots, \theta_p) + \varepsilon_i \qquad 1 \leq i \leq n$$

donde se puede suponer que las $\varepsilon_i$ satisfacen (i), (ii) y (iii). Luego, si llamamos

$$S_i(\theta_1, \theta_2, \ldots, \theta_p) = u(G_i, H_i, \theta_1, \theta_2, \ldots, \theta_p)$$

obtenemos que las variables $X_i$ satisfacen (3.6).

Llamaremos *estimador de cuadrados mínimos* (E.C.M.) al valor $\widehat{\boldsymbol{\theta}}(X_1, X_2, \ldots, X_n)$ que hace mínima la expresión $\sum_{i=1}^{n}(X_i - S_i(\theta_1, \theta_2, \ldots, \theta_p))^2$, es decir si

$$\sum_{i=1}^{n}(X_i - S_i(\widehat{\boldsymbol{\theta}}))^2 = \min_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{n}(X_i - S_i(\boldsymbol{\theta}))^2 \ . \tag{3.7}$$

Este estimador tiene la siguiente justificación intuitiva: Se desea que $S_i(\theta_1 \ldots \theta_p)$ "ajuste" bien a $X_i$, y por lo tanto los términos residuales $\varepsilon_i$ deberían ser pequeños. Esto se logra minimizando la suma de los cuadrados de las desviaciones respectivas.

Se puede demostrar que si además de satisfacer (i), (ii) y (iii), los $\varepsilon_i$ tienen distribución normal, entonces el E.M.C. coincide con el E.M.V. Esto se verá en el problema 3 de 3.4.

**Computación de los E.C.M.:** Si $\Theta$ es abierto y si las funciones $S_i(\theta_1, \theta_2, \ldots, \theta_p)$ son derivables respecto a cada $\theta_i$, $\widehat{\boldsymbol{\theta}}$ deberá satisfacer el sistema de ecuaciones siguiente

$$\frac{\partial \sum_{i=1}^{n}(X_i - S_i(\widehat{\boldsymbol{\theta}}))^2}{\partial\theta_j} = 0 \qquad\qquad j = 1, 2, \ldots, p \ ,$$

que es equivalente a:

$$\sum_{i=1}^{n}(X_i - S_i(\widehat{\boldsymbol{\theta}}))\frac{\partial S_i(\widehat{\boldsymbol{\theta}})}{\partial\theta_j} = 0 \qquad\qquad j = 1, 2, \ldots, p \ .$$

Igual que en el caso de los E.M.V. estas condiciones son necesarias para el E.M.C. pero no son suficientes. También se deberán cumplir las condiciones de segundo orden, y se deberá verificar que se trata de un mínimo absoluto y no local.

**Ejemplo 3:** Volvemos al ejemplo 1. Luego se tiene

$$\frac{\partial S_i(\boldsymbol{\theta})}{\partial\theta_1} = G_i \qquad \text{y} \qquad \frac{\partial S_i(\boldsymbol{\theta})}{\partial\theta_2} = 1 \ .$$

Luego (3.7) se transforma en

$$\begin{aligned}\sum_{i=1}^{n}(X-\widehat{\theta}_1 G_i-\widehat{\theta}_2)G_i &= 0\\ \sum_{i=1}^{n}(X_i-\widehat{\theta}_1 G_i-\widehat{\theta}_2) &= 0\,.\end{aligned}$$

Es fácil ver la que la solución de este sistema viene dada por

$$\begin{aligned}\widehat{\theta}_1 &= \sum_{i=1}^{n}(X_i-\overline{X})(G_i-\overline{G})\Big/\sum_{i=1}^{n}(G_i-\overline{G})^2\,,\\ \widehat{\theta}_2 &= \overline{X}-\widehat{\theta}_1\overline{G}\,,\end{aligned}$$

donde

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad \text{y} \qquad \overline{G}=\frac{1}{n}\sum_{i=1}^{n}G_i\,.$$

Geométricamente la recta $X=\widehat{\theta}_1 G+\widehat{\theta}_2$ tiene la propiedad siguiente: Minimaza la suma de los cuadrados de las distancias de los puntos $(G_i, X_i)$ a la recta, si esta distancia se la mide paralelamente al eje de las $X$. Es decir si $X_i^*=\theta_1 G_1+\theta_2$, la recta $X=\widehat{\theta}_1 G+\widehat{\theta}_2$ hace mínimo $\sum_{i=1}^{n}(X_i-X_i^*)^2$.

Para un mayor desarrollo de los métodos de cuadrados mínimos, consultar Draper y Smith [2].

## 3.5 Criterios para medir la bondad de un estimador

Supongamos que se tenga una muestra $\mathbf{X}=(X_1, X_2, \ldots, X_n)$ de cuya distribución sólo se conoce que pertenece a la familia $\mathcal{F}=\{F(\mathbf{x},\boldsymbol{\theta})$ donde $\boldsymbol{\theta}\in\Theta\subset I\!R^p\}$. Supongamos además que se está interesado en estimar una función real $q(\boldsymbol{\theta})$. Para poder elegir el estimador $\delta(\mathbf{X})$ que se utilizará, se deberá dar un criterio para comparar dos estimadores cualesquiera. Esto se hará como sigue:

Es razonable pensar que dado un estimador $\delta(\mathbf{X})$ de $q(\boldsymbol{\theta})$, el error $\delta(\mathbf{X})-q(\boldsymbol{\theta})$ producirá un perjuicio o pérdida dado por un real no negativo, que dependerá por un lado del valor del estimador $\delta(\mathbf{X})$ y por otro del valor verdadero del vector $\boldsymbol{\theta}$ de parámetros.

Así llamaremos *función de pérdida* a una función $\ell(\boldsymbol{\theta}, d)$ no negativa que nos indica cuánto se pierde cuando el valor del estimador es "$d$" y el valor verdadero del vector de parámetros es $\boldsymbol{\theta}$. Entonces si usamos el estimador $\delta(\mathbf{X})$ la pérdida será

$$\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))$$

y esta pérdida será una variable aleatoria ya que depende de $\mathbf{X}$. Para evaluar *globalmente* el estimador $\delta(\mathbf{X})$ se puede utilizar el valor medio de esta pérdida, que indicará de acuerdo a la ley de los grandes números aproximadamente la pérdida promedio, si estimamos $q(\boldsymbol{\theta})$ muchas veces con vectores $\mathbf{X}$ independientes. Luego, definimos la función de pérdida media del estimador $\delta$ o *función de riesgo* $R(\delta, \boldsymbol{\theta})$ a

$$R(\delta, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\ell(\boldsymbol{\theta}, \delta(\mathbf{X})))$$

Un primer ejemplo de función de pérdida puede obtenerse tomando el error absoluto, es decir

$$\ell_1(\boldsymbol{\theta}, d) = |d - q(\boldsymbol{\theta})|$$

y en este caso, la pérdida media corresponde a un estimador $\delta(\mathbf{X})$ viene dada por

$$R_1(\delta, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(|\delta(\mathbf{X} - q(\boldsymbol{\theta})|)$$

Si consideramos como función de pérdida el cuadrado del error tenemos

$$\ell_2(\boldsymbol{\theta}, d) = (d - q(\boldsymbol{\theta}))^2$$

que es una función que desde el punto de vista matemático es más sencilla que $\ell_1$, ya que es derivable en todo punto.

La función de pérdida cuadrática fue la primera utilizada en Estadística, y aún hoy la más difundida. De ahora en adelante, salvo mención en contrario supondremos que la función de pérdida es $\ell_2$. La pérdida media, o riesgo, correspondiente está dada por

$$R_2(\delta, \boldsymbol{\theta}) = E(\delta(\mathbf{X}) - q(\boldsymbol{\theta}))^2$$

y será llamada en adelante error cuadrático medio, e indicada por $\mathrm{ECM}_{\boldsymbol{\theta}}(\delta)$. Luego

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta) = R_2(\delta, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\delta(\mathbf{X}) - q(\boldsymbol{\theta}))^2 \tag{3.8}$$

La función $\mathrm{ECM}_{\boldsymbol{\theta}}(\delta)$ nos proporciona un criterio para determinar si un estimador $\delta_1(\mathbf{X})$ de $q(\boldsymbol{\theta})$ es mejor que otro $\delta_2(\mathbf{X})$, basta verificar

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta_1) \leq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta_2) \qquad \forall\, \boldsymbol{\theta} \in \Theta$$

En este orden de ideas, un estimador óptimo $\delta^*$ podría definirse mediante la siguiente condición: Para cualquier otro estimador $\delta$ se tiene

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \leq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta) \qquad \forall\, \theta \in \Theta \tag{3.9}$$

Sin embargo, salvo en casos triviales no existirán tales estimadores óptimos. Para mostrar esto definamos para cada posible valor $\boldsymbol{\theta} \in \Theta$, el estimador constante $\delta_{\boldsymbol{\theta}}(\mathbf{X}) = q(\boldsymbol{\theta})$ que no depende del valor de la muestra. Luego si $\delta^*$ satisface (3.9), debe cumplirse:

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \leq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta_{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}((q(\boldsymbol{\theta}) - q(\boldsymbol{\theta}))^2) = 0 \quad \forall \boldsymbol{\theta} \in \Theta$$

Pero como $\mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \geq 0$ y $\ell_2(\boldsymbol{\theta}, d) = 0$ implica que $d = q(\boldsymbol{\theta})$, se obtiene

$$P_{\boldsymbol{\theta}}(\delta^*(\mathbf{X}) = q(\boldsymbol{\theta})) = 1 \quad \forall \boldsymbol{\theta} \in \Theta \tag{3.10}$$

(donde $P_{\boldsymbol{\theta}}(\Lambda)$ indica la probabilidad del evento $\Lambda$ cuando el valor de los parámetros está dado por el vector $\boldsymbol{\theta}$). La ecuación (3.10) significa que a partir de la muestra se puede estimar sin error $q(\boldsymbol{\theta})$. Esta situación sólo se da muy raramente, por ejemplo, cuando $q(\boldsymbol{\theta})$ es constante.

Otro ejemplo algo diferente de función de pérdida, corresponde a la función

$$\ell_3(\boldsymbol{\theta}, d) = I_{\{|q(\boldsymbol{\theta}) - d| > c\}}$$

donde $I_{\{|q(\boldsymbol{\theta}) - d| > c\}}$ es la función que vale 1 si $|q(\boldsymbol{\theta}) - d| > c$ y 0 en caso contrario. Esta pérdida da origen a la función de riesgo

$$R_3(\delta, \boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(|\delta(\mathbf{X}) - q(\boldsymbol{\theta})| > c)\ .$$

A diferencia de las anteriores, en este caso, $\ell_3(\boldsymbol{\theta}, d) = 0$ no implica implica $q(\boldsymbol{\theta}) = d$. Por otra parte, esta pérdida no es convexa como función de $d$ mientras que $\ell_1$ y $\ell_2$ lo son. En muchas situaciones, se podrán obtener procedimientos de estimación más efectivos para pérdidas convexas.

El estimador $\delta^*$ con E.C.M. mínimo uniformemente en $\boldsymbol{\theta}$ como se indica en (3.9) no existe, salvo en casos excepcionales, debido a que la clase de todos los posibles estimadores es muy amplia y contiene estimadores poco razonables como los $\delta_{\boldsymbol{\theta}}(\mathbf{X})$ definidos anteriormente. Por lo tanto, una manera de obtener estimadores óptimos consistirá en restringir primero la clase de los estimadores $\delta$ considerados, y luego buscar aquél con E.C.M. uniformemente menor dentro de esta clase. Otra forma de obtener estimadores óptimos consistirá en minimizar algún criterio general basado en la función de riesgo, como el máximo riesgo.

Antes de empezar el estudio de las clases de estimadores daremos una noción importante.

**Definición 1:** Se dice que un estimador $\delta(\mathbf{X})$ de $q(\boldsymbol{\theta})$ es *inadmisible* respecto de la pérdida $\ell(\boldsymbol{\theta}, d)$, si existe otro estimador $\delta'(\mathbf{X})$ mejor que él, es decir, si existe $\delta'(\mathbf{X})$ tal que

$$R(\delta', \boldsymbol{\theta}) \leq R(\delta, \boldsymbol{\theta}) \qquad \forall\, \boldsymbol{\theta} \in \Theta$$

El estimador $\delta(\mathbf{X})$ se dirá *admisible* si no es inadmisible, es decir, si no existe ningún otro estimador que sea uniformemente mejor que él.

El siguiente Teorema muestra la ventaja de utilizar pérdidas convexas.

**Teorema 1.** *Supongamos que $\ell(\boldsymbol{\theta}, d)$ es una pérdida estrictamente convexa en $d$ y que $\delta(\mathbf{X})$ es admisible para $q(\boldsymbol{\theta})$. Si $\delta'(\mathbf{X})$ es otro estimador de $q(\boldsymbol{\theta})$ con el mismo riesgo que $\delta(\mathbf{X})$ entonces $P_{\boldsymbol{\theta}}(\delta(\mathbf{X}) = \delta'(\mathbf{X})) = 1$.*
DEMOSTRACIÓN. Supongamos que $P_{\boldsymbol{\theta}}(\delta(\mathbf{X}) = \delta'(\mathbf{X})) < 1$ y sea $\delta^*(\mathbf{X}) = (\delta(\mathbf{X}) + \delta'(\mathbf{X}))/2$. Luego, por ser $\ell(\boldsymbol{\theta}, d)$ convexa se cumple

$$\ell(\boldsymbol{\theta}, \delta^*(\mathbf{X})) < \frac{\ell(\boldsymbol{\theta}, \delta(\mathbf{X})) + \ell(\boldsymbol{\theta}, \delta'(\mathbf{X}))}{2} \tag{3.11}$$

salvo si $\delta(\mathbf{X}) = \delta'(\mathbf{X})$. Luego, tomando esperanza en ambos miembros de (3.11) se obtiene

$$R(\delta^*, \boldsymbol{\theta}) < \frac{R(\delta, \boldsymbol{\theta}) + R(\delta', \boldsymbol{\theta})}{2} = R(\delta, \boldsymbol{\theta}) \tag{3.12}$$

lo que contradice el hecho de que $\delta(\mathbf{X})$ es admisible.

## 3.6 Estimadores insesgados

Una propiedad "razonable" que se puede exigir a un estimador está dada por la siguiente definición:

**Definición 1:** Se dice que $\delta(\mathbf{X})$ es un *estimador insesgado* para $q(\boldsymbol{\theta})$ si $E_{\boldsymbol{\theta}}(\delta(\mathbf{X})) = q(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta$.

Esto significa que si calculamos el estimador $\delta$ para varias muestras independientes, y luego promediamos los valores así obtenidos, entonces de

acuerdo a la ley de los grandes números el promedio converge al valor $q(\boldsymbol{\theta})$ que queremos estimar.

**Definición 2:** Si un estimador no es insesgado, se dice *sesgado*, definiéndose el *sesgo* del estimador como $E_{\boldsymbol{\theta}}(\delta(\mathbf{X})) - q(\boldsymbol{\theta})$.

Cuando $\delta(\mathbf{X})$ es un estimador insesgado, su ECM coincide con su varianza ya que

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta) = E_{\boldsymbol{\theta}}[(\delta(\mathbf{X}) - q(\boldsymbol{\theta}))^2] = E_{\boldsymbol{\theta}}[(\delta(\mathbf{X}) - E_{\boldsymbol{\theta}}(\delta(\mathbf{X})))^2] = \mathrm{Var}_{\boldsymbol{\theta}}(\delta(\mathbf{X})).$$

Para ilustrar estas definiciones veremos algunos ejemplos.

**Ejemplo 1:** Supongamos tener una variable $X$ de cuya distribución $F$ en la población sólo se sabe que tiene esperanza finita, es decir sólo se conoce que pertenece a $\mathcal{F}$, donde $\mathcal{F}$ es la familia de todas las distribuciones con esperanza finita. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de $F$ y supongamos que se quiere estimar $q_1(F) = E_F(X)$. Estamos frente a un problema de estimación no paramétrica, ya que la familia no puede indicarse con un número finito de parámetros. Un posible estimador para $q_1(F)$ es $\overline{X} = (1/n)\sum_{i=1}^n X_i$. El estimador $\overline{X}$ es insesgado ya que

$$E_F(\overline{X}) = E_F\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) = \frac{1}{n}\sum_{i=1}^n E_F(X_i) = E_F(X) = q_1(F)$$

$\overline{X}$ se denomina *media muestral.*

**Ejemplo 2:** Supongamos ahora que se conoce que la distribución $F$ de $X$ en la población pertenece a la familia $\mathcal{F}$ de todas las distribuciones que tienen segundo momento finito, es decir tales que $E_F(X^2) < \infty$. Supongamos que se quiere estimar $q_2(F) = \mathrm{Var}_F(X)$ a partir de una muestra aleatoria $X_1, X_2, \ldots, X_n$. Ya hemos visto que un estimador adecuado podría ser

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$$

Veremos que $\widehat{\sigma}^2$ no es un estimador insesgado de $q_2(F)$. Desarrollando el cuadrado del segundo miembro en la definición obtenemos

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n X_i^2 - n\overline{X}^2}{n}.$$

Luego, se tiene

$$E_F(\widehat{\sigma}^2) = E_F(X^2) - E_F(\overline{X}^2) \tag{3.13}$$

Por otro lado, se tiene

$$\mathrm{Var}_F(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}_F(X_i) = \frac{1}{n}\mathrm{Var}_F(X) \ .$$

Como

$$\mathrm{Var}_F(\overline{X}) = E_F(\overline{X}^2) - (E_F(\overline{X}))^2 \ ,$$

resulta

$$E_F(\overline{X}^2) = \mathrm{Var}_F(\overline{X}) + (E_F(\overline{X}))^2 = \frac{1}{n}\mathrm{Var}_F(X) + (E_F(X))^2 \tag{3.14}$$

y reemplazando (3.14) en (3.13) resulta

$$\begin{aligned} E_F(\widehat{\sigma}^2) &= E_F(X^2) - (E_F(X))^2 - \frac{1}{n}\mathrm{Var}_F(X) = \mathrm{Var}_F(X)(1-1/n) \\ &= \frac{n-1}{n}\ \mathrm{Var}_F(X) = \frac{n-1}{n} q_2(F). \end{aligned}$$

Esto prueba que $\widehat{\sigma}^2$ no es un estimador insesgado para $\mathrm{Var}_F(X)$, aunque el sesgo es $-\mathrm{Var}_F(X)/n$, y por lo tanto, tiende a 0 cuando $n$ tiende a infinito. El sesgo puede corregirse dividiendo $\widehat{\sigma}^2$ por $(n-1)/n$, obteniendo así el estimador insesgado

$$s^2 = \frac{n}{n-1}\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

que denominaremos *varianza muestral.*

**Ejemplo 3:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución de la cual se conoce únicamente que pertenece a la familia $N(\mu, \sigma^2)$ y supongamos que se quieran estimar $\mu$ y $\sigma^2$. Como se tiene

$$\mu = E_{\mu,\sigma^2}(X) \ ; \quad \sigma^2 = \mathrm{Var}_{\mu,\sigma^2}(X)$$

por lo visto en Ejemplos 1 y 2, resulta que $\overline{X}$ y $s^2$ son estimadores insesgados de $\mu$ y $\sigma^2$ respectivamente.

Si nos restringimos a la clase de los estimadores insesgados, se podrá encontrar frecuentemente, estimadores óptimos. Daremos la siguiente definición:

**Definición 2:** Se dirá que $\delta(\mathbf{X})$ es un *estimador insesgado de mínima varianza* para $q(\boldsymbol{\theta})$, uniformemente en $\boldsymbol{\theta} \in \Theta$ (IMVU) si:

(a) $\delta(\mathbf{X})$ es insesgado para $q(\boldsymbol{\theta})$

(b) dado otro estimador insesgado para $q(\boldsymbol{\theta})$, $\delta^*(\mathbf{X})$, se cumple $\mathrm{Var}_{\boldsymbol{\theta}}(\delta(\mathbf{X})) \leq \mathrm{Var}_{\boldsymbol{\theta}}(\delta^*(\mathbf{X})) \quad \forall\, \boldsymbol{\theta} \in \Theta$.

## 3.7 Estadísticos suficientes

Consideremos un vector aleatorio $\mathbf{X}$ de dimensión $n$ cuya distribución pertenece a una familia $\mathcal{F} = \{F(\mathbf{x}, \boldsymbol{\theta}) \text{ con } \boldsymbol{\theta} \in \Theta \subset I\!R^p\}$. El vector $\mathbf{X}$ interesa en cuanto nos provee información sobre el valor verdadero de $\boldsymbol{\theta}$. Puede ocurrir que una parte de la información contenida en $\mathbf{X}$ carezca de interés para el conocimiento de $\boldsymbol{\theta}$, y por consiguiente convenga eliminarla simplificando así la información disponible.

Al realizar esta simplificación, eliminando de $\mathbf{X}$ toda la información irrelevante, se obtendrá otro vector $\mathbf{T}$ que puede ser de dimensión menor que $n$.

Llamaremos *estadístico* a cualquier función medible $\mathbf{T} = r(\mathbf{X})$ con valores en un espacio euclídeo de dimensión finita.

Si la función $r$ no es biunívoca, del conocimiento de $\mathbf{T}$ no se podrá reconstruir el valor de $\mathbf{X}$, por lo que $\mathbf{T}$ conservará sólo una parte de la información que hay en $\mathbf{X}$. El estadístico $\mathbf{T}$ será llamado *suficiente* cuando conserve toda la información relevante para el conocimiento de $\boldsymbol{\theta}$. Esto se formalizará en la siguiente definición.

**Definición 1:** Sea $\mathbf{X}$ un vector aleatorio de dimensión $n$ cuya distribución es $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Se dice que un *estadístico* $\mathbf{T} = r(\mathbf{X})$ *es suficiente* para $\boldsymbol{\theta}$ si la distribución de $\mathbf{X}$ condicional a que $\mathbf{T} = \mathbf{t}$ es independiente de $\boldsymbol{\theta}$ para todo $\mathbf{t}$.

Esto puede interpretarse como afirmando que una vez conocido el valor $\mathbf{t}$ de $\mathbf{T}$, la distribución de $\mathbf{X}$ es independiente de $\boldsymbol{\theta}$ y por lo tanto no contiene información suplementaria sobre $\boldsymbol{\theta}$. En otros términos: una vez conocido el valor de $\mathbf{T}$ podemos olvidarnos del valor $\mathbf{X}$, ya que en $\mathbf{T}$ está toda la información que $\mathbf{X}$ tiene sobre $\boldsymbol{\theta}$.

**Ejemplo 1:** Supongamos que una máquina produce cierto artículo, existiendo la probabilidad $\theta$ de que lo produzca defectuoso. Supongamos además que se observa un lote de $n$ artículos producidos sucesivamente por la máquina,

de manera que la aparición de uno defectuoso resulte independiente del resultado obtenido para los restantes artículos del lote.

Consideremos las variable aleatorias $X_i$, $1 \leq i \leq n$, que valen 1 ó 0 según el i-ésimo artículo observado sea o no defectuoso. Entonces cada una de las variables $X_1, X_2, \ldots, X_n$ sigue una ley binomial $\mathrm{Bi}(\theta, 1)$, de modo que la función de probabilidad puntual conjunta es igual a

$$p(x_1, x_2, \ldots, x_n, \theta) = \theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{n-\Sigma_{i=1}^n x_i}$$

donde $x_i$ vale 0 ó 1.

Si queremos estimar el parámetro $\theta$, parece razonable pensar que sólo se deberá utilizar la cantidad total de artículos defectuosos del lote, ya que el orden en que han aparecido los mismos parece irrelevante para el conocimiento de $\theta$. Por lo tanto, es de esperar que el estadístico $T = \sum_{i=1}^n X_i$ sea suficiente.

Para ver si esta conjetura es correcta, calculemos la distribución de $\mathbf{X} = (X_1, \ldots, X_n)$ dado $T = t$:

$$p_{\mathbf{X}|T}(x_1, \ldots, x_n, \theta|t) = \frac{p_{\mathbf{X},T}(x_1, x_2, \ldots, x_n, t, \theta)}{p_T(t, \theta)} \tag{3.15}$$

El numerador de este cociente es la probabilidad conjunta:

$$\begin{aligned} & P_\theta(X_1 = x_1, \ldots, X_n = x_n, r(X_1, \ldots, X_n) = t) \\ = & \begin{cases} \theta^t(1-\theta)^{n-t} & \text{si } r(x_1, \ldots, x_n) = t \\ 0 & \text{si } r(x_1, \ldots, x_n) \neq t \end{cases} \end{aligned}$$

y como el estadístico $T = \sum_{i=1}^n X_i$ sigue una ley binomial $\mathrm{Bi}(\theta, n)$ el denominador de (3.15) vale

$$p_T(t, \theta) = \binom{n}{t} \theta^t(1-\theta)^{n-t}$$

Así resulta

$$p_{\mathbf{X}|T}(x_1, \ldots, x_n, \theta|t) = \begin{cases} 1/\binom{n}{t} & \text{si } r(x_1, \ldots, x_n) = t \\ 0 & \text{si } r(x_1, \ldots, x_n) \neq t\ . \end{cases}$$

De esta manera $p_{\mathbf{X}/T}$ es independiente de $\theta$ y por lo tanto el estadístico $T = \Sigma\, X_i$ es suficiente para $\theta$.

Una caracterización útil de los estadísticos suficientes es la proporcionada por el siguiente teorema:

**Teorema 1** (de factorización): *Sea* $\mathbf{X}$ *un vector aleatorio con función de densidad o función de probabilidad puntual* $p(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. *Entonces, el estadístico* $\mathbf{T} = r(\mathbf{X})$ *es suficiente para* $\boldsymbol{\theta}$ *si y sólo si existen dos funciones* $g$ *y* $h$ *tales que*

$$p(\mathbf{x}, \boldsymbol{\theta}) = g(r(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}) \tag{3.16}$$

DEMOSTRACIÓN: La haremos sólo para el caso discreto. Supongamos primero que existen dos funciones $g$ y $h$ tales que $p(\mathbf{x}, \boldsymbol{\theta})$ se factoriza según (3.16). Entonces la función de densidad conjunta vale

$$p_{\mathbf{XT}}(\mathbf{x}, \mathbf{t}, \boldsymbol{\theta}) = \begin{cases} g(\mathbf{t}, \boldsymbol{\theta}) h(\mathbf{x}) & \text{si } r(\mathbf{x}) = \mathbf{t} \\ 0 & \text{si } r(\mathbf{x}) \neq \mathbf{t} \end{cases}$$

y la densidad marginal $p_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\theta})$ está dada por

$$\begin{aligned} p_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\theta}) &= \sum_{r(\mathbf{x})=\mathbf{t}} p_{\mathbf{XT}}(\mathbf{x}, \mathbf{t}, \boldsymbol{\theta}) = \sum_{r(\mathbf{x})=\mathbf{t}} g(r(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}) \\ &= g(\mathbf{t}, \boldsymbol{\theta}) \sum_{r(\mathbf{x})=\mathbf{t}} h(\mathbf{x}) = g(\mathbf{t}, \boldsymbol{\theta}) h^*(\mathbf{t}) \end{aligned}$$

donde las sumatorias se realizan sobre todos los $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ tales que $r(\mathbf{x}) = \mathbf{t}$. Así resulta la función de densidad condicional

$$p_{\mathbf{X}|\mathbf{T}}(\mathbf{x}, \boldsymbol{\theta}|\mathbf{t}) = \begin{cases} h(\mathbf{x})/h^*(\mathbf{t}) & \text{si } r(\mathbf{x}) = \mathbf{t} \\ 0 & \text{si } r(\mathbf{x}) \neq \mathbf{t} \end{cases}$$

y por lo tanto la distribución de $\mathbf{X}$ dado $\mathbf{T} = \mathbf{t}$ es independiente de $\boldsymbol{\theta}$ para todo $\mathbf{t}$.

Recíprocamente, si suponemos que $\mathbf{T} = r(\mathbf{X})$ es suficiente para $\boldsymbol{\theta}$, se tiene

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}) &= P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x},\, \mathbf{T} = r(\mathbf{x})) = p_{\mathbf{XT}}(\mathbf{x}, r(\mathbf{x}), \boldsymbol{\theta}) \\ &= p_{\mathbf{X}|\mathbf{T}}(\mathbf{x}, \boldsymbol{\theta}|r(\mathbf{x})) p_{\mathbf{T}}(r(\mathbf{x}), \boldsymbol{\theta}) \end{aligned}$$

El primero de los factores del último miembro es por hipótesis independiente de $\boldsymbol{\theta}$ y por eso podemos llamarlo $h(\mathbf{x})$; mientras que el segundo –que depende de $\mathbf{x}$ a través de $\mathbf{t}$– puede denominarse $g(r(\mathbf{x}), \boldsymbol{\theta})$. El teorema queda demostrado. Para una demostración general, ver Teorema 8 y Corolario 1 de Lehmann [4]. También se puede ver Bahadur [1].

**Ejemplo 2:** Supongamos que las variables aleatorias $X_1, X_2, \ldots, X_n$ son independientes y que están uniformemente distribuídas en el intervalo $[\theta_1, \theta_2]$ de manera que su función de densidad conjunta vale

$$p(x_1, \ldots, x_n, \theta_1, \theta_2) = \begin{cases} (\theta_2 - \theta_1)^{-n} & \text{si } \theta_1 \leq x_i \leq \theta_2,\ \forall i,\ 1 \leq i \leq n \\ 0 & \text{en el resto de } I\!R^n \end{cases}$$

Si definimos los estadísticos

$$r_1(\mathbf{X}) = \min\{X_i :\ 1 \leq i \leq n\} \qquad \text{y} \qquad r_2(\mathbf{X}) = \max\{X_i :\ 1 \leq i \leq n\}$$

y si denotamos con $I_{[\theta_1,\theta_2]}(y)$ a la función característica del intervalo $[\theta_1, \theta_2]$ (que vale 1 para todo $y$ del intervalo y 0 fuera del mismo), resulta:

$$p(x_1, \ldots, x_n, \theta_1, \theta_2) = (\theta_2 - \theta_1)^{-n} I_{[\theta_1,\theta_2]}(r_1(x_1, \ldots, x_n)) I_{[\theta_1,\theta_2]}(r_2(x_1, \ldots, x_n))$$

Por lo tanto la función de densidad $p(\mathbf{x}, \boldsymbol{\theta})$ se factoriza como en (3.16) con $h(\mathbf{x}) = 1$. La función $g$ que depende de $\mathbf{X}$ a través de $r_1(\mathbf{x})$ y $r_2(\mathbf{x})$ vale en este caso

$$g(r_1(\mathbf{x}), r_2(\mathbf{x}), \boldsymbol{\theta}) = (\theta_2 - \theta_1)^{-n} I_{[\theta_1,\theta_2]}(r_1(\mathbf{x})) I_{[\theta_1,\theta_2]}(r_2(\mathbf{x}))$$

Esto demuestra que el estadístico

$$\mathbf{T} = (r_1(\mathbf{X})\ ,\ r_2(\mathbf{X}))$$

es suficiente para $\theta_1$ y $\theta_2$.

El siguiente resultado es Corolario inmediato del Teorema 1.

**Corolario.** *Sea* $\mathbf{X}$ *un vector aleatorio con función de densidad o función de probabilidad puntual* $p(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. *Supongamos que la familia* $\{p(\mathbf{x}, \boldsymbol{\theta})\}$ *tiene soporte común, independiente de* $\boldsymbol{\theta}$. *Entonces, una condición necesaria y suficiente para que* $\mathbf{T}$ *sea suficiente para* $\boldsymbol{\theta}$ *es que fijados* $\boldsymbol{\theta}_1$ *y* $\boldsymbol{\theta}_2$ *el cociente* $\frac{p(\mathbf{x}, \boldsymbol{\theta}_1)}{p(\mathbf{x}, \boldsymbol{\theta}_2)}$ *sea función de* $\mathbf{T}$.

El siguiente Teorema muestra que una función biunívoca de un estadístico suficiente es también un estadístico suficiente. Esta propiedad es intuitivamente razonable: si $\mathbf{T}$ contiene toda la información relevante acerca de $\boldsymbol{\theta}$, y $\mathbf{T}^*$ es una función biunívoca de $\mathbf{T}$, entonces también $\mathbf{T}^*$ la contiene ya que el vector $\mathbf{T}$ puede reconstruirse a partir del vector $\mathbf{T}^*$.

**Teorema 2:** *Si* $\mathbf{X}$ *es un vector aleatorio con una distribución* $F(\mathbf{x}, \boldsymbol{\theta})$*, con* $\boldsymbol{\theta} \in \Theta$ *si* $\mathbf{T} = r(\mathbf{X})$ *es un estadístico suficiente para* $\boldsymbol{\theta}$ *y si* $m$ *es una función biunívoca de* $\mathbf{T}$ *entonces el estadístico* $\mathbf{T}^* = m(\mathbf{T})$ *también es suficiente para* $\boldsymbol{\theta}$*.*

DEMOSTRACIÓN: Apliquemos el teorema de factorización a la función de densidad del vector $\mathbf{X}$:

$$p(\mathbf{x}, \boldsymbol{\theta}) = g(r(\mathbf{x}), \boldsymbol{\theta})h(\mathbf{x}) = g(m^{-1}(m(r(\mathbf{x})), \boldsymbol{\theta})h(\mathbf{x})$$

El primer factor del último miembro es una función $g^*(r^*(\mathbf{x}), \boldsymbol{\theta})$, donde $r^*(\mathbf{x}) = m(r(\mathbf{x}))$, y esto prueba que $\mathbf{T}^* = r^*(\mathbf{X})$ es suficiente para $\boldsymbol{\theta}$.

## 3.8 Estadísticos minimales suficientes

De la noción intuitiva de suficiencia, se deduce que si $\mathbf{T}$ es suficiente para $\boldsymbol{\theta}$ y $\mathbf{T} = H(\mathbf{U})$ entonces $\mathbf{U}$ es suficiente para $\boldsymbol{\theta}$, ya que el conocimiento de $\mathbf{U}$ permite conocer $\mathbf{T}$ que es el que contiene toda la información relevante sobre $\boldsymbol{\theta}$. Más aún, salvo que $H$ sea biunívoca $\mathbf{T}$ da una mayor reducción de la muestra original que $\mathbf{U}$. Este hecho motiva la siguiente definición.

**Definición 1:** Sea $\mathbf{X}$ un vector aleatorio de dimensión $n$ cuya distribución es $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Se dice que un *estadístico* $\mathbf{T} = r(\mathbf{X})$ *es minimal suficiente* para $\boldsymbol{\theta}$ si dado cualquier otro estadístico $\mathbf{U} = g(\mathbf{X})$ suficiente para $\boldsymbol{\theta}$ existe una función $H$ tal que $\mathbf{T} = H(\mathbf{U})$.

En muchas situaciones, es fácil construir estadísticos minimal suficientes. Sea $S(\boldsymbol{\theta}) = \{\mathbf{x} : p(\mathbf{x}, \boldsymbol{\theta}) > 0\}$, $S(\boldsymbol{\theta})$ se llama el soporte de la densidad o de la probabilidad puntual $p(\mathbf{x}, \boldsymbol{\theta})$, según corresponda. Para simplificar, supondremos que las posibles distribuciones del vector $\mathbf{X}$ tienen todas el mismo soporte, es decir, que el conjunto $S(\boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$.

**Teorema 1.** *Supongamos que* $\mathbf{X}$ *tiene una distribución perteneciente a una familia finita de distribuciones* $\mathcal{F} = \{F(\mathbf{x}, \boldsymbol{\theta}_i) \quad 1 \leq i \leq k\}$ *con densidades o probabilidades puntuales* $p(\mathbf{x}, \boldsymbol{\theta}_i)$*,* $1 \leq i \leq k$ *todas con el mismo soporte. Entonces el estadístico*

$$\mathbf{T} = r(\mathbf{x}) = \left(\frac{p(\mathbf{x}, \boldsymbol{\theta}_2)}{p(\mathbf{x}, \boldsymbol{\theta}_1)}, \ldots, \frac{p(\mathbf{x}, \boldsymbol{\theta}_k)}{p(\mathbf{x}, \boldsymbol{\theta}_1)}\right)$$

*es minimal suficiente.*

DEMOSTRACIÓN. Obviamente, para todo $1 \leq i < j \leq k$ el cociente $p(\mathbf{x}, \boldsymbol{\theta}_i)/p(\mathbf{x}, \boldsymbol{\theta}_j)$ es función de $\mathbf{T}$. Por lo tanto, por el Corolario del teorema de Factorización, $\mathbf{T}$ es suficiente.
Sea ahora $\mathbf{U}$ un estadístico suficiente para $\boldsymbol{\theta}$. Entonces, utilizando el Corolario anterior se cumple que para todo $2 \leq i \leq k$, el cociente $\frac{p(\mathbf{x}, \boldsymbol{\theta}_i)}{p(\mathbf{x}, \boldsymbol{\theta}_1)}$ es una función de $\mathbf{U}$. Luego, $\mathbf{T}$ es función de $\mathbf{U}$ y $\mathbf{T}$ es minimal suficiente.

En muchas situaciones, se pueden obtener estadísticos minimales suficientes combinando el Teorema 1 con el siguiente Teorema.

**Teorema 2.** *Supongamos que* $\mathbf{X}$ *tiene una distribución perteneciente a una familia de distribuciones* $\mathcal{F} = \{F(\mathbf{x}, \boldsymbol{\theta}) \quad \boldsymbol{\theta} \in \Theta\}$ *con densidades o probabilidades puntuales* $p(\mathbf{x}, \boldsymbol{\theta})$*, todas con el mismo soporte. Sea*

$$\mathcal{F}_0 = \{F(\mathbf{x}, \boldsymbol{\theta}) \quad \boldsymbol{\theta} \in \Theta_0 \subset \Theta\} \subset \mathcal{F}.$$

*Supongamos además que* $\mathbf{T} = r(\mathbf{X})$ *es un estadístico minimal suficiente para* $\boldsymbol{\theta} \in \Theta_0$ *y suficiente para* $\boldsymbol{\theta} \in \Theta$*, entonces* $\mathbf{T}$ *es minimal suficiente para* $\boldsymbol{\theta} \in \Theta$*.*
DEMOSTRACIÓN. Sea $\mathbf{U}$ un estadístico suficiente para $\boldsymbol{\theta}$, entonces $\mathbf{U}$ es suficiente para $\boldsymbol{\theta} \in \Theta_0$. Por lo tanto, $\mathbf{T}$ es función de $\mathbf{U}$, con lo cual $\mathbf{T}$ es minimal suficiente.

**Ejemplo 1.** Sean $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $Bi(\theta, 1)$, $0 < \theta < 1$. Hemos visto que $T = \sum_{i=1}^{n} X_i$ es suficiente para $\theta \in (0, 1)$. Queremos ver que es minimal suficiente.

Para ello consideremos la familia finita $\mathcal{F}_0 = \{Bi(1/4, 1), Bi(3/4, 1)\}$. Luego, un estadístico minimal suficiente para esta familia está dado por

$$U = g(\mathbf{x}) = \frac{p(\mathbf{x}, \frac{3}{4})}{p(\mathbf{x}, \frac{1}{4})} = 3^{2T-n}$$

que es una función biunívoca de $T$. Por lo tanto, $T$ es un estadístico minimal suficiente para $\mathcal{F}_0$ y suficiente para $\theta \in (0, 1)$, con lo cual es minimal suficiente para $\theta \in (0, 1)$.

## 3.9 Estimadores basados en estadísticos suficientes

Supongamos que $\mathbf{X}$ es un vector correspondiente a una muestra de una distribución que pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Supongamos que $\mathbf{T} = r(\mathbf{X})$ es un estadístico suficiente para $\boldsymbol{\theta}$. Luego de acuerdo al concepto

intuitivo que tenemos de estadístico suficiente, para estimar una función $q(\boldsymbol{\theta})$ deberán bastar estimadores que dependan sólo de $\mathbf{T}$, ya que en $\mathbf{T}$ está toda la información que $\mathbf{X}$ contiene sobre el parámetro $\boldsymbol{\theta}$. Esto es justamente lo que afirma el siguiente teorema.

**Teorema 1** (Rao–Blackwell): *Sea* $\mathbf{X}$ *un vector de una distribución perteneciente a la familia* $F(\mathbf{x}, \boldsymbol{\theta})$ *con* $\boldsymbol{\theta} \in \Theta$. *Sea* $\mathbf{T}$ *un estadístico suficiente para* $\boldsymbol{\theta}$ *y* $\delta(\mathbf{X})$ *un estimador de* $q(\boldsymbol{\theta})$. *Definamos un nuevo estimador*

$$\delta^*(\mathbf{T}) = E(\delta(\mathbf{X})|\mathbf{T}).$$

*Luego se tiene*

(i) $\mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \leq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta), \ \ \forall\, \theta \in \Theta$

(ii) *La igualdad en* (i) *se satisface si y sólo si*

$$P_{\boldsymbol{\theta}}(\delta^*(\mathbf{T}) = \delta(\mathbf{X})) = 1 \quad \forall\, \boldsymbol{\theta} \in \Theta$$

(iii) *Si* $\delta(\mathbf{X})$ *es insesgado, entonces* $\delta^*(\mathbf{T})$ *también lo es.*

Demostración: Podemos escribir

$$\begin{aligned} \mathrm{ECM}_{\boldsymbol{\theta}}(\delta) &= E_{\boldsymbol{\theta}}((\delta(\mathbf{X}) - q(\boldsymbol{\theta}))^2) \\ &= E_{\boldsymbol{\theta}}([(\delta^*(\mathbf{T}) - q(\boldsymbol{\theta})) + (\delta(\mathbf{X}) - \delta^*(T))]^2) \\ &= E_{\boldsymbol{\theta}}((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))^2) + E_{\boldsymbol{\theta}}((\delta(\mathbf{X}) - \delta^*(\mathbf{T}))^2) \\ &+ 2\, E_{\boldsymbol{\theta}}((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) \end{aligned} \tag{3.17}$$

Luego, usando

$$\begin{aligned} E_{\boldsymbol{\theta}}((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) &= E_{\boldsymbol{\theta}}[E((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))|\mathbf{T}] \\ &= E_{\boldsymbol{\theta}}[(\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))E(\delta(\mathbf{X}) - \delta^*(\mathbf{T})|\mathbf{T}] \end{aligned}$$

y

$$E_{\boldsymbol{\theta}}(\delta(\mathbf{X}) - \delta^*(\mathbf{T})|\mathbf{T}) = E(\delta(\mathbf{X})|\mathbf{T}) - \delta^*(\mathbf{T}) = \delta^*(\mathbf{T}) - \delta^*(\mathbf{T}) = 0\,,$$

se obtiene

$$E_{\boldsymbol{\theta}}((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) = 0\,.$$

Luego (3.17) se transforma en

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta) = \mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) + E_{\boldsymbol{\theta}}((\delta(\mathbf{X}) - \delta^*(\mathbf{T}))^2)$$

y resulta

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta) \geq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \ .$$

Además igualdad se cumple sólo si $P_{\boldsymbol{\theta}}(\delta(\mathbf{X}) = \delta^*(\mathbf{T})) = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta$.

Luego ya se ha demostrado (i) y (ii). Para mostrar (iii) supongamos que $\delta$ es insesgado, luego se tiene

$$E_{\boldsymbol{\theta}}(\delta^*(\mathbf{T})) = E_{\boldsymbol{\theta}}(E(\delta(\mathbf{X})|\mathbf{T})) = E_{\boldsymbol{\theta}}(\delta(\mathbf{X})) = q(\boldsymbol{\theta})$$

Luego se cumple (iii).

**Observación:** El estimador $\delta^*(\mathbf{T}) = E(\delta(\mathbf{X})|\mathbf{T})$ es realmente un estimador ya que depende sólo de $\mathbf{T}$ (y por lo tanto de $\mathbf{X}$) y no de $\boldsymbol{\theta}$, ya que por ser $\mathbf{T}$ un estadístico suficiente la distribución de $\delta(\mathbf{X})$ condicional $\mathbf{T} = \mathbf{t}$ es independiente de $\boldsymbol{\theta}$, por lo tanto lo mismo sucede con la esperanza condicional.

**Ejemplo 1:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $Bi(\theta, 1)$. Luego $\delta(X_1, \ldots, X_n) = X_1$ es un estimador insesgado de $\theta$. Un estadístico suficiente para $\theta$ es $T = \sum_{i=1}^n X_i$ (ver ejemplo 1 de 3.7). Por lo tanto, de acuerdo al teorema de Rao–Blackwell, $\delta^*(T) = E(\delta(X_1, \ldots, X_n)|T)$ será otro estimador insesgado de $\theta$ y $\mathrm{Var}_\theta(\delta^*) \leq \mathrm{Var}_\theta(\delta)$. Vamos a calcular entonces $\delta^*(T)$.

Por ser $X_1, X_2, \ldots, X_n$ idénticamente distribuídas y como $T$ es invariante por permutaciones entre $X_1, X_2, \ldots, X_n$, la distribución conjunta de $(X_i, T)$ es la misma para todo $i$. Por lo tanto, $E(X_i|T)$ será independiente de $i$ (ver Problema 1 de 3.9). Luego

$$E(X_i|T) = E(X_1|T) = \delta^*(T) \qquad 1 \leq i \leq n \ .$$

Sumando en $i$ se tiene

$$\sum_{i=1}^n E(X_i|T) = n\, \delta^*(T) \ .$$

Pero además vale que

$$\sum_{i=1}^n E(X_i|T) = E\Big(\sum_{i=1}^n X_i|T\Big) = E(T|T) = T \ ,$$

luego

$$\delta^*(T) = \frac{T}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

Es fácil ver que

$$\mathrm{Var}_\theta(\delta^*(T)) \leq \mathrm{Var}_\theta(\delta(X))$$

ya que

$$\mathrm{Var}_\theta(\delta^*(T)) = \theta(1-\theta)/n \quad \text{y} \quad \mathrm{Var}_\theta(\delta(X)) = \theta(1-\theta) \ .$$

## 3.10 Familias exponenciales

**Definición:** Se dice que una familia de distribuciones continuas o discretas en $\mathbb{R}^q$, $F(\mathbf{x}, \boldsymbol{\theta})$, donde $\mathbf{x} = (x_1, \ldots, x_q)$ y $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ es una *familia exponencial a k parámetros* si la correspondiente función de densidad discreta o continua se puede escribir como

$$p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{\sum_{i=1}^{k} c_i(\boldsymbol{\theta}) r_i(\mathbf{x})} h(\mathbf{x}) \tag{3.18}$$

donde $c_1(\boldsymbol{\theta}), \ldots, c_k(\boldsymbol{\theta})$ son funciones de $\Theta$ en $\mathbb{R}$, $A(\boldsymbol{\theta})$ es una función de $\Theta$ en $\mathbb{R}^+$ (reales no negativos), $r_1(\mathbf{x}), \ldots, r_k(\mathbf{x})$ son funciones de $\mathbb{R}^q$ en $\mathbb{R}$ y $h(\mathbf{x})$ es una función de $\mathbb{R}^q$ en $\mathbb{R}^+$.

**Ejemplo 1:** Sea la familia $Bi(\theta, n)$ con $n$ fijo y $\theta$ en (0,1). Luego

$$\begin{aligned} p(x,\theta) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} = (1-\theta)^n \left(\tfrac{\theta}{1-\theta}\right)^n \binom{n}{x} \quad x = 0, 1, \ldots, n \\ &= (1-\theta)^n e^{x \ln(\theta/(1-\theta))} \binom{n}{x} \end{aligned}$$

Luego esta familia es exponencial a un parámetro con $A(\theta) = (1-\theta)^n$; $r(x) = x$; $c(\theta) = \ln(\theta/(1-\theta))$ y $h(x) = \binom{n}{x}$.

**Ejemplo 2:** Sea la familia $N(\mu, \sigma^2)$ con $\mu \in \mathbb{R}$ y $\sigma^2$ real positivo. Luego, su densidad viene dada por

$$p(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\,\sigma^2}x^2+(\frac{\mu}{\sigma^2})x-\frac{\mu^2}{2\sigma^2}}$$

$$= \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2)}} e^{(-\frac{1}{2\,\sigma^2})x^2+\frac{\mu}{\sigma^2}x} \tag{3.19}$$

Luego esta es una familia exponencial a dos parámetros con $A(\mu,\sigma^2) = e^{-\mu^2/2\sigma^2}/\sqrt{2\pi\sigma^2}$; $c_1(\mu,\sigma^2) = (-1/2\,\sigma^2)$; $c_2(\mu,\sigma^2) = \mu/\sigma^2$; $r_1(x) = x^2$; $r_2(x) = x$; $h(x) = 1$.

**Ejemplo 3:** Sea la familia $P(\lambda)$. Se puede mostrar que es exponencial a un parámetro. Ver problema 2.i) de 3.10.

**Ejemplo 4:** Sea la familia $\varepsilon(\lambda)$. Se puede mostrar que es exponencial a un parámetro. Ver problema 2.ii) de 3.10.

**Ejemplo 5:** Sea la familia de distribuciones normales bivariadas $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$. Es exponencial a 5 parámetros. Ver problema 2.iii) de 3.10.

**Teorema 1:** *Una familia exponencial a k parámetros cuya función de densidad viene dada por (3.18) tiene como estadístico suficiente para* $\boldsymbol{\theta}$ *el vector* $\mathbf{T} = \mathbf{r}(\mathbf{X}) = (r_1(\mathbf{X}),\ldots,r_k(\mathbf{X}))$.

Demostración. Inmediata a partir del Teorema 1 de 3.9.

El siguiente teorema establece la propiedad más importante de las familias exponenciales.

**Teorema 2:** *Sea* $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ *una muestra aleatoria de una distribución que pertenece a una familia exponencial a k parámetros, cuya función de densidad viene dada por (3.18). Luego la distribución conjunta de* $\mathbf{X}_1,\ldots,\mathbf{X}_n$ *también pertenece a una familia exponencial a k parámetros y el estadístico suficiente para* $\boldsymbol{\theta}$ *es el vector*

$$\mathbf{T}^* = (T_1^*,\ldots,T_k^*),\ \textit{donde}\ T_i^* = \sum_{j=1}^{n} r_i(\mathbf{X}_j), \quad 1 \le i \le k$$

Demostración: Es inmediata, ya que por (3.18) se tiene

$$p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\boldsymbol{\theta}) = \textstyle\prod_{j=1}^{n} p(\mathbf{x}_i,\boldsymbol{\theta})$$

$$= (A(\boldsymbol{\theta}))^n e^{c_1(\boldsymbol{\theta})\Sigma_{i=1}^n r_1(\mathbf{x}_j)+\cdots+c_k(\boldsymbol{\theta})\Sigma_{i=1}^n r_k(\mathbf{x}_j)} \textstyle\prod_{j=1}^{n} h(\mathbf{x}_j)$$

$$= A^*(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta})r_1^*(\mathbf{x}_1,\ldots,\mathbf{x}_n)+\cdots+c_k(\boldsymbol{\theta})r_k^*(\mathbf{x}_1,\ldots,\mathbf{x}_n)} h^*(\mathbf{x}_1,\ldots,\mathbf{x}_n)$$

donde $A^*(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^n$; $r_i^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{j=1}^n r_i(\mathbf{x}_j)$, $h^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^n h(\mathbf{x}_j)$, y por lo tanto el Teorema 2 queda demostrado.

Este último Teorema nos afirma que para familias exponenciales de $k$ parámetros, cualquiera sea el tamaño de la muestra, siempre existe un estadístico suficiente de sólo $k$ componentes. Es decir, que toda la información se puede resumir en $k$ variables aleatorias. Se puede mostrar que esta propiedad bajo condiciones generales caracteriza a las familias exponenciales. Para esta caracterización se puede consultar Sección 2.5 de Zacks [7] y Dynkin [3].

**Ejemplo 3:** Volvamos al ejemplo 1. Supongamos que tomamos una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución $Bi(\theta, n)$ con $n$ fijo. Luego la distribución conjunta de la muestra pertenecerá a una familia exponencial a un parámetro con estadístico suficiente $T = \sum_{i=1}^n X_i$.

**Ejemplo 4:** Sea $X_1, \ldots, X_n$ una muestra de una distribución perteneciente a la familia $N(\mu, \sigma^2)$. Luego, de acuerdo a lo visto en el ejemplo 2 y al teorema 2, la distribución conjunta de $X_1, X_2, \ldots, X_n$ pertenece a una familia exponencial a dos parámetros y con estadístico suficiente $T = \left( \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i \right)$.

El siguiente teorema establece que las familias de distribuciones de los estadísticos suficientes de una familia exponencial a $k$ parámetros también forma una familia exponencial a $k$ parámetros.

**Teorema 3:** *Sea* $\mathbf{X}$ *un vector cuya distribución pertenece a una familia exponencial a* $k$ *parámetros cuya función de densidad satisface (3.18). Luego la función de densidad de los estadísticos suficientes* $\mathbf{T} = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$ *es de la forma*

$$p_{\mathbf{T}}(t_1, t_2, \ldots, t_k, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta}) t_1 + \cdots + c_k(\boldsymbol{\theta}) t_k} h^*(t_1, \ldots, t_k)$$

*Por lo tanto la familia de distribuciones de* $\mathbf{T}$ *también forma una familia exponencial a* $k$ *parámetros.*

DEMOSTRACIÓN: Sólo se hará para el caso discreto. Para el caso general se puede consultar Lema 8 de 2.7 en Lehmann [4]. En el caso particular elegido se tiene:

$$p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{\sum_{j=1}^k c_j(\boldsymbol{\theta}) r_j(\mathbf{x})} h(\mathbf{x})$$

Luego si $\mathbf{T} = r(\mathbf{x}) = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$ y si $\mathbf{t} = (t_1, \ldots, t_k)$, se tendrá

$$\begin{aligned} p_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\theta}) &= \sum_{\{\mathbf{x}:\, r(\mathbf{x})=\mathbf{t}\}} p(\mathbf{x}, \boldsymbol{\theta}) = \sum_{\{\mathbf{x}:\, r(\mathbf{x})=\mathbf{t}\}} A(\boldsymbol{\theta}) e^{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) r_j(\mathbf{x})} h(\mathbf{x}) \\ &= A(\boldsymbol{\theta}) e^{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) t_j} \sum_{\{\mathbf{x}:\, r(\mathbf{x})=\mathbf{t}\}} h(\mathbf{x}) = A(\boldsymbol{\theta}) e^{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) t_j} h^*(\mathbf{t}) \end{aligned}$$

con $h^*(\mathbf{t}) = \sum_{\{\mathbf{x}: r(\mathbf{x})=\mathbf{t}\}} h(\mathbf{x})$.

El siguiente lema es de carácter técnico y nos será útil en lo que sigue.

**Lema 1:** *Sea $\mathbf{X} = (X_1, \ldots, X_q)$ un vector aleatorio cuya distribución pertenece a una familia exponencial a un parámetro discreta o continua con densidad dada por $p(\mathbf{x}, \boldsymbol{\theta}) = A(\theta) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x})$; con $\theta \in \Theta$, donde $\Theta$ es un abierto en $\mathbb{R}$ y $c(\theta)$ infinitamente derivable. Luego, si $m(\mathbf{x})$ es un estadístico tal que*

$$\int \ldots \int |m(\mathbf{x})| p(\mathbf{x}, \theta) dx_1 \ldots dx_q < \infty \quad \forall\, \theta \in \Theta$$

*o*

$$\Sigma_{x_1} \ldots \Sigma_{x_q} |m(\mathbf{x})| p(\mathbf{x}, \theta) < \infty$$

*según sea $\mathbf{X}$ continua o discreta, entonces las expresiones*

$$\int \ldots \int m(\mathbf{x}) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x}) dx_1 \ldots dx_q \quad o \quad \Sigma_{x_1} \ldots \Sigma_{x_q} m(\mathbf{x}) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x})$$

*según corresponda, son infinitamente derivables y se puede derivar dentro de los signos integral o sumatoria, respectivamente.*

Demostración: No se dará en este curso, puede consultarse en el Teorema 9 de 2.7 de Lehmann [4].

**Teorema 4:** *Sea $\mathbf{X} = (X_1, \ldots, X_q)$ un vector aleatorio cuya distribución pertenece a una familia exponencial a un parámetro con densidad dada por $p(\mathbf{x}, \theta) = A(\theta) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x})$ con $\theta \in \Theta$, donde $\Theta$ es un abierto en $\mathbb{R}$ y $c(\theta)$ es infinitamente derivable. Luego se tiene:*

*(i) $A(\theta)$ es infinitamente derivable.*

*(ii)*

$$E_\theta(r(\mathbf{X})) = -\frac{A'(\theta)}{A(\theta) c'(\theta)}$$

*(iii)*

$$Var_\theta(r(\mathbf{x})) = \frac{\frac{\partial E_\theta(r(\mathbf{x}))}{\partial \theta}}{c'(\theta)}$$

DEMOSTRACIÓN: Supongamos que $\mathbf{X}$ sea continuo. El caso discreto es totalmente similar. Como

$$\int \dots \int A(\theta) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x}) dx_1 \dots dx_q = 1$$

se tiene

$$\frac{1}{A(\theta)} = \int \dots \int e^{c(\theta) r(\mathbf{x})} h(\mathbf{x}) dx_1 \dots dx_q$$

Como el segundo miembro de esta igualdad satisface las condiciones del Lema 1 con $m(\mathbf{x}) = 1$, resulta infinitamente derivable y luego también $A(\theta)$, con lo cual queda demostrado (i).

Por otro lado se tiene

$$A(\theta) \int \dots \int e^{c(\theta) r(x)} h(x) dx_1 \dots dx_q = 1 \quad \forall\, \theta \in \Theta$$

y usando el Lema 1 que nos permite derivar dentro del signo integral resulta

$$A'(\theta) \int \dots \int e^{c(\theta) r(\mathbf{x})} h(\mathbf{x}) dx_1 \dots dx_q +$$
$$A(\theta) c'(\theta) \int \dots \int r(\mathbf{x}) e^{c(\theta) r(\mathbf{x})} dx_1 \dots dx_q = 0$$

y esta última ecuación se puede escribir

$$\frac{A'(\theta)}{A(\theta)} + c'(\theta) E_\theta(r(\mathbf{x})) = 0$$

y luego

$$E_\theta(r(\mathbf{x})) = -\frac{A'(\theta)}{c'(\theta) A(\theta)}$$

y se ha demostrado (ii).

(iii) se deja para resolver en el Problema 3 de 3.10.

## 3.11 Estadísticos completos

Sea $\mathbf{X}$ un vector aleatorio cuya distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Hasta ahora hemos visto que tomando estimadores insesgados de una función $q(\boldsymbol{\theta})$ basados en estadísticos suficientes se logra mejorar la estimación. Lo que no conocemos es si puede haber más de un estimador insesgado, basado en un estadístico suficiente $\mathbf{T}$ dado. Veremos que bajo ciertas condiciones hay uno solo.

**Definición 1:** Sea $\mathbf{X}$ un vector aleatorio cuya distribución pertenece a una familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Un estadístico $\mathbf{T} = r(\mathbf{X})$ se dice completo si $E_\theta(g(\mathbf{T})) = 0$ para todo $\boldsymbol{\theta}$ implica que $P_{\boldsymbol{\theta}}(g(\mathbf{T}) = 0) = 1$ para todo $\boldsymbol{\theta} \in \Theta$

**Ejemplo 1:** Sea $X$ una variable aleatoria con distribución $Bi(\theta, k)$ con $k$ fijo y $0 \leq \theta \leq 1$. Sea $g$ tal que $E_\theta(g(X)) = 0$, para todo $\theta$. Mostraremos que $g(x) = 0$, $x = 0, 1, \ldots, k$. Tenemos

$$E_\theta(g(X)) = \sum_{x=0}^{k} g(x) \binom{k}{x} \theta^x (1-\theta)^{k-x} = 0 \quad \forall\, \theta \in [0,1] \tag{3.20}$$

Sea $\lambda = \theta/(1-\theta)$; luego cuando $\theta \in [0,1]$, $\lambda$ toma los valores en $\mathbb{R}^+$ (reales no negativos).

Poniendo (3.20) en función de $\lambda$ resulta

$$(1-\theta)^k \sum_{x=0}^{k} g(x) \binom{k}{x} \lambda^x = 0 \quad \forall\, \lambda \in \mathbb{R}^+$$

Luego

$$Q(\lambda) = \sum_{x=0}^{k} g(x) \binom{k}{x} \lambda^x = 0 \quad \forall\, \lambda \in \mathbb{R}^+$$

Pero $Q(\lambda)$ es un polinomio de grado $k$ con infinitas raíces, luego todos sus coeficientes deben ser 0. Por lo tanto,

$$g(x) \binom{k}{x} = 0 \qquad x = 0, 1, \ldots, k\,,$$

y entonces

$$g(x) = 0 \qquad x = 0, 1, \ldots, k\,.$$

Con lo que queda probado que $T(X) = X$ es un estadístico completo.

**Ejemplo 2:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución que pertenece a la familia $Bi(\theta, k)$. Sea $T = r(X_1, \ldots, X_n) = X_1 + X_2 + \cdots + X_n$. Luego $T$ es un estadístico suficiente y tiene distribución $Bi(\theta, nk)$, por lo tanto de acuerdo a lo visto en el ejemplo 1 es completo.

**Ejemplo 3:** Consideremos una variable $X$ con distribución $U[0, \theta]$, $\theta \in I\!R^+$. Sea $T = X$. Luego se puede demostrar que $T$ es un estadístico completo. La demostración de este hecho está fuera de los alcances de este curso. De todos modos, veremos una proposición más débil relacionada con completitud. Sea $g$ de $I\!R^+$ en $I\!R$ una función continua. Luego veremos que si $E_\theta(g(X)) = 0$ para todo $\theta$ en $I\!R^+$, entonces $g(x) = 0$

$$E_\theta(g(X)) = \frac{1}{\theta)} \int_0^\theta g(x)dx = 0, \quad \forall\, \theta \geq 0,$$

luego

$$\int_0^\theta g(x)dx = 0, \quad \forall\, \theta \in I\!R^+$$

Sea $G(\theta) = \int_0^\theta g(x)dx$, entonces se tiene

$$G(\theta) = 0 \qquad \forall\, \theta \in I\!R^+$$

Usando el Teorema Fundamental del Cálculo Integral se tiene que

$$\frac{\partial G(\theta)}{\partial \theta} = g(\theta) = 0 \quad \forall\, \theta \in I\!R^+$$

Lo que faltaría ver es que en el caso en que $g$ no es continua, $E_\theta(g(X)) = 0 \quad \forall\, \theta \in I\!R^+$ implica $g(x) = 0$ con probabilidad 1.

El siguiente teorema muestra que bajo condiciones muy generales el estadístico suficiente correspondiente a una familia exponencial es completo.

**Teorema 1:** *Sea una familia exponencial a k parámetros, discreta o continua con función de densidad dada por*

$$p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta}) r_1(\mathbf{x}) + \ldots + c_k(\boldsymbol{\theta}) r_k(\mathbf{x})} h(\mathbf{x})$$

*y sea* $\Lambda = \{\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k) : \lambda_i = c_i(\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta\}$.

*a) Si* $\Lambda$ *contiene* $k + 1$ *puntos* $\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(k+1)}$ *tales que* $\{\boldsymbol{\lambda}^{(j)} - \boldsymbol{\lambda}^{(1)}, 2 \leq j \leq k + 1\}$ *son linealmente independientes, entonces el estadístico suficiente* $\mathbf{T} = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$ *es minimal suficiente.*

*b) Si $\Lambda$ un conjunto que contiene una esfera en $\mathbb{R}^k$, entonces estadístico suficiente $\mathbf{T} = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$ es completo.*

DEMOSTRACIÓN: a) Como $\mathbf{T}$ es suficiente para $\mathcal{F} = \{p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta}) r_1(\mathbf{x}) + \ldots + c_k(\boldsymbol{\theta}) r_k(\mathbf{x})} h(\mathbf{x}) \quad \boldsymbol{\theta} \in \Theta\}$, de acuerdo al Teorema 2 de la sección 3.8 bastará probar que $\mathbf{T}$ es minimal suficiente para una subfamilia finita de $\mathcal{F}$. Sean $\boldsymbol{\theta}^{(j)}$ , $1 \leq j \leq k+1$, tales que

$$\boldsymbol{\lambda}^{(j)} = (\lambda_1^{(j)}, \ldots, \lambda_k^{(j)}) = \left(c_1(\boldsymbol{\theta}^{(j)}), \ldots, c_k(\boldsymbol{\theta}^{(j)})\right) .$$

Consideremos la subfamilia

$$\begin{aligned} \mathcal{F}_0 = \{p(\mathbf{x}, \boldsymbol{\theta}^{(j)}) &= A(\boldsymbol{\theta}^{(j)}) e^{\sum_{i=1}^k c_i(\boldsymbol{\theta}^{(j)}) r_i(\mathbf{x})} h(\mathbf{x}) \\ &= A(\boldsymbol{\theta}^{(j)}) e^{\sum_{i=1}^k \lambda_i^{(j)} r_i(\mathbf{x})} h(\mathbf{x}) \; 1 \leq j \leq k+1\} . \end{aligned}$$

Luego, por el Teorema 1 de la sección 3.8 un estadístico minimal suficiente para $\mathcal{F}_0$ está dado por

$$\begin{aligned} \mathbf{T}^* &= r^*(\mathbf{x}) = \left( \frac{p(\mathbf{x}, \boldsymbol{\theta}^{(2)})}{p(\mathbf{x}, \boldsymbol{\theta}^{(1)})}, \ldots, \frac{p(\mathbf{x}, \boldsymbol{\theta}^{(k+1)})}{p(\mathbf{x}, \boldsymbol{\theta}^{(1)})} \right) \\ &= \left( \frac{A(\boldsymbol{\theta}^{(2)}) e^{\lambda_1^{(2)} r_1(\mathbf{x}) + \ldots + \lambda_k^{(2)} r_k(\mathbf{x})}}{A(\boldsymbol{\theta}^{(1)}) e^{\lambda_1^{(1)} r_1(\mathbf{x}) + \ldots + \lambda_k^{(1)} r_k(\mathbf{x})}}, \ldots, \frac{A(\boldsymbol{\theta}^{(k+1)}) e^{\lambda_1^{(k+1)} r_1(\mathbf{x}) + \ldots + \lambda_k^{(k+1)} r_k(\mathbf{x})}}{A(\boldsymbol{\theta}^{(1)}) e^{\lambda_1^{(1)} r_1(\mathbf{x}) + \ldots + \lambda_k^{(1)} r_k(\mathbf{x})}} \right) \end{aligned}$$

que es equivalente a

$$\mathbf{T}^{**} = r^{(**)}(\mathbf{x}) = \left( \sum_{i=1}^k [\lambda_i^{(2)} - \lambda_i^{(1)}] r_i(\mathbf{x}), \ldots, \sum_{i=1}^k [\lambda_i^{(k+1)} - \lambda_i^{(1)}] r_i(\mathbf{x}) \right) .$$

Como $\mathbf{T}^{**} = M\mathbf{T}$ donde la matriz $M \in \mathbb{R}^{k \times k}$ es no singular, ya que su $j-$ésima columna es el vector $\boldsymbol{\lambda}^{(j+1)} - \boldsymbol{\lambda}^{(1)}$, $\mathbf{T}$ es equivalente a $\mathbf{T}^{**}$ y por lo tanto, es minimal suficiente para $\mathcal{F}_0$, de donde se obtiene el resultado.
b) Para una demostración general se puede ver Teorema 1 de Sección 4.3 de Lehmann [4]. En este curso sólo se demostrará para el caso que $k = 1$, y que $T = r(\mathbf{X})$ toma un número finito de valores racionales. De acuerdo al teorema 3, en este caso la función de densidad de $T$ será de la forma:

$$p(t, \theta) = A(\boldsymbol{\theta}) e^{c(\boldsymbol{\theta}) t} h(t)$$

Supongamos que los posibles valores de $T$ que tienen probabilidad positiva es el conjunto $A = \{t_1, t_2, \ldots, t_r\} \cup \{-t'_1, -t'_2, \ldots, -t'_s\}$ donde los $t_i$ y los $t'_j$ son racionales no negativos.

Sea $v$ un múltiplo común de los denominadores de todos los racionales $t_i$ y $t'_j$ y sean $w_i = vt_i \quad 1 \leq i \leq r$ y $w'_i = vt'_i$, $1 \leq i \leq s$. Luego los $w_i$ y los $w'_i$ son naturales. Finalmente sea $w = \max_{1 \leq i \leq s} w'_i$, $z_i = w_i + w, \quad 1 \leq i \leq r$ y $z'_i = -w'_i + w$, $1 \leq i \leq s$. Luego los $z_i$ y los $z'_i$ son naturales y todos diferentes.

Supongamos que

$$E_{\boldsymbol{\theta}}(g(T)) = 0 \qquad \forall\, \boldsymbol{\theta} \in \Theta$$

luego

$$\sum_{i=1}^{r} g(t_i)p(t_i, \boldsymbol{\theta}) + \sum_{i=1}^{s} g(-t'_i)p(-t'_i, \boldsymbol{\theta}) = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta$$

con lo cual

$$\sum_{i=1}^{r} g(t_i)A(\boldsymbol{\theta})e^{c(\boldsymbol{\theta})t_i}h(t_i) + \sum_{i=1}^{s} g(-t'_i)A(\boldsymbol{\theta})e^{-c(\boldsymbol{\theta})t'_i}h(-t'_i) = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta\,,$$

de donde se obtiene

$$\sum_{i=1}^{r} g(t_i)h(t_i)(e^{c(\boldsymbol{\theta})/v})^{t_i v} + \sum_{i=1}^{s} g(-t'_i)h(-t'_i)(e^{c(\boldsymbol{\theta})/v})^{-t'_i v} = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta\,.$$

Llamando $\lambda = e^{c(\boldsymbol{\theta})/v}$ resulta que como hay infinitos posibles valores de $c(\boldsymbol{\theta})$, el conjunto $\Lambda$ de posibles valores de $\lambda$, también es infinito. Luego tenemos

$$\sum_{i=1}^{r} g(t_i)h(t_i)\lambda^{w_i} + \sum_{i=1}^{s} g(-t'_i)h(-t'_i)\lambda^{-w'_i} = 0 \quad \forall\, \lambda \in \Lambda$$

Multiplicando por $\lambda^w$ la última ecuación resulta

$$P(\lambda) = \sum_{i=1}^{r} g(t_i)h(t_i)\lambda^{z_i} + \sum_{i=1}^{s} g(t'_i)h(-t'_i)\lambda^{z'_i} = 0 \quad \forall\, \lambda \in \Lambda$$

Luego el polinomio $P(\lambda)$ tiene infinitas raíces y por lo tanto, todos los coeficientes deben ser 0, es decir, $g(t_i)h(t_i) = 0, \quad 1 \leq i \leq r$ y $g(-t'_i)h(-t'_i) = 0, \quad 1 \leq i \leq s$. Como $h(t_i) > 0, \quad 1 \leq i \leq r$ y $h(-t'_i) > 0, \quad 1 \leq i \leq s$,

resulta que $g(t_i) = 0 \quad 1 \leq i \leq r$ y $g(-t_i') = 0 \quad 1 \leq i \leq s$. Con lo cual, $P_{\boldsymbol{\theta}}(g(T) = 0) = 1$ para todo $\boldsymbol{\theta} \in \Theta$.

**Ejemplo 4:** Sea $X_1$ una variable $N(\mu, \sigma_1^2)$ y $X_2$ independiente de $X_1$ una variable $N(\mu, \sigma_2^2)$, luego si $\boldsymbol{\theta} = (\mu, \sigma_1^2, \sigma_2^2)$ la densidad de $\mathbf{X} = (X_1, X_2)$ puede escribirse como

$$p(x_1, x_2, \boldsymbol{\theta}) = \frac{1}{2\,\pi\sigma_1\sigma_2} e^{-\mu^2(\frac{1}{2\,\sigma_1^2}-\frac{1}{2\sigma_2^2})} e^{(-\frac{1}{2\,\sigma_1^2})x_1^2+(-\frac{1}{2\,\sigma_2^2})x_2^2+(\frac{\mu}{\sigma_1^2})x_1+(\frac{\mu}{\sigma_2^2})x_2}$$

Por lo tanto es una familia exponencial a 4 parámetros, pero no satisface la condición del Teorema 1 ya que el conjunto

$$\Lambda = \{\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4) \text{ con } \lambda_1 = -\frac{1}{2\,\sigma_1^2}\,, \lambda_2 = -\frac{1}{2\,\sigma_2^2}\;\lambda_3 = \frac{\mu}{\sigma_1^2}\;\lambda_4 = \frac{\mu}{\sigma_2^2}\}\,,$$

está en una superficie de dimensión 3, ya que depende de 3 parámetros, $\sigma_1^2$, $\sigma_2^2$ y $\mu$, y por lo tanto no contiene ninguna esfera de $\mathbb{R}^4$. Como el Teorema 1 de la sección 3.11 da un condición suficiente pero no necesaria para completitud, no se deduce que $\mathbf{T} = (X_1, X_2, X_1^2, X_2^2)$ no sea completo. Sin embargo, dado que $E_{\mu,\sigma_1^2,\sigma_2^2}(X_1 - X_2) = \mu - \mu = 0$ y $X_1 - X_2$ no es igual a 0 resulta que $\mathbf{T}$ no es completo.

El Teorema 1 nos permite, sin embargo, deducir que $\mathbf{T}$ es minimal suficiente.

Por lo tanto, hemos visto un estadístico minimal suficiente no necesariamente es completo. El siguiente resultado establece la recíproca.

**Teorema 2:** *Sea* $\mathbf{T}$ *un estadístico suficiente y completo para* $\boldsymbol{\theta}$*. Si existe un estadístico minimal suficiente para* $\boldsymbol{\theta}$ *entonces* $\mathbf{T}$ *es minimal suficiente.*
DEMOSTRACIÓN. La haremos sólo en el caso en que el estadístico minimal suficiente y el estadístico suficiente y completo $T$ tienen dimensión 1. Sea $U$ el estadístico minimal suficiente para $\boldsymbol{\theta}$, luego por ser $T$ suficiente se cumple que $U = m(T)$. Queremos ver que $m$ es biunívoca.

Sea $\psi(t)$ la función arcotangente. Luego $\psi : \mathbb{R} \to [0, 2\pi]$ es una función estrictamente creciente y acotada. Por lo tanto, $E_{\boldsymbol{\theta}}(\psi(T)) < \infty$ y bastará mostrar que $\psi(T)$ es función de $U$.

Definamos $\eta(U) = E(\psi(T)|U)$. Como $U$ es suficiente $\eta(U)$ es un estadístico. Luego, si

$$g(T) = \psi(T) - \eta[m(T)] = \psi(T) - \eta(U)$$

se cumple que $E_{\boldsymbol{\theta}}[g(T)] = 0$ para todo $\boldsymbol{\theta} \in \Theta$. Por lo tanto, $P_{\boldsymbol{\theta}}(\psi(T) = \eta(U)) = 1$ para todo $\boldsymbol{\theta} \in \Theta$, y entonces $T$ es equivalente a $U$.

El siguiente Teorema es útil en muchas situaciones, donde probar independencia entre estadísticos puede resultar laborioso.

**Teorema 3:** (Teorema de Basu) *Sea* $\mathbf{T}$ *un estadístico suficiente y completo para* $\boldsymbol{\theta}$. *Sea* $\mathbf{U} = g(\mathbf{X})$ *un estadístico cuya distribución no depende de* $\boldsymbol{\theta}$ *entonces* $\mathbf{U}$ *es independiente de* $\mathbf{T}$.
DEMOSTRACIÓN. Sea $A$ un suceso, como $\mathbf{U}$ tiene distribución independiente de $\boldsymbol{\theta}$, $p_A = P(\mathbf{U} \in A)$ no depende de $\boldsymbol{\theta}$.
Sea $\eta_A(\mathbf{t}) = P(\mathbf{U} \in A|\mathbf{T} = \mathbf{t})$. Como $\mathbf{T}$ es suficiente $\eta_A(\mathbf{T})$ es un estadístico. Por otra parte, $E_{\boldsymbol{\theta}}(\eta_A(\mathbf{T}) - p_A) = 0$ para todo $\boldsymbol{\theta} \in \Theta$, con lo cual la completitud de $\mathbf{T}$ implica que $P_{\boldsymbol{\theta}}(\eta_A(\mathbf{T}) = p_A) = 1$ para todo $\boldsymbol{\theta} \in \Theta$ y por lo tanto, $\mathbf{U}$ es independiente de $\mathbf{T}$.

## 3.12 Estimadores insesgados de mínima varianza uniformemente

El siguiente teorema nos da un método para construir estimadores IMVU cuando se conoce un estadístico que es a la vez suficiente y completo.

**Teorema 1** (Lehmann-Scheffé): *Sea* $\mathbf{X}$ *un vector aleatorio de cuya distribución pertenece a la familia* $F(\mathbf{x}, \boldsymbol{\theta})$ *con* $\boldsymbol{\theta} \in \Theta$. *Sea* $\mathbf{T}$ *un estadístico suficiente y completo. Luego dada una función* $q(\boldsymbol{\theta})$ *de* $\Theta$ *en* $\mathbb{R}$, *se tiene que*

(i) *Existe a lo sumo un estimador insesgado de* $q(\boldsymbol{\theta})$, *basado en* $\mathbf{T}$.

(ii) *Si* $\delta(\mathbf{T})$ *es un estimador insesgado de* $q(\boldsymbol{\theta})$, *entonces* $\delta(\mathbf{T})$ *es IMVU.*

(iii) *Si* $\delta(\mathbf{X})$ *es un estimador insesgado para* $q(\boldsymbol{\theta})$, *luego* $\delta^*(\mathbf{T}) = E(\delta(\mathbf{X})|\mathbf{T})$ *es un estimador IMVU para* $q(\boldsymbol{\theta})$.

DEMOSTRACIÓN:

(i) Sean $\delta_1(\mathbf{T})$ y $\delta_2(\mathbf{T})$ dos estimadores insesgados de $q(\boldsymbol{\theta})$. Luego

$$E_{\boldsymbol{\theta}}(\delta_1(\mathbf{T}) - \delta_2(\mathbf{T})) = q(\boldsymbol{\theta}) - q(\boldsymbol{\theta}) = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta$$

luego como $\mathbf{T}$ es completo

$$P_{\boldsymbol{\theta}}(\delta_1(\mathbf{T}) - \delta_2(\mathbf{T}) = 0) = 1, \ \forall \boldsymbol{\theta} \in \Theta$$

(ii) Sea $\delta(\mathbf{T})$ un estimador insesgado de $q(\boldsymbol{\theta})$, y sea $\delta_1(\mathbf{X})$ otro estimador insesgado. Si llamamos $\delta_1^*(\mathbf{T}) = E(\delta_1(\mathbf{X})|\mathbf{T})$ sabemos por el Teorema 1 de la sección 3.9 que $\delta_1^*(\mathbf{T})$ es insesgado y

$$\mathrm{Var}_{\boldsymbol{\theta}}(\delta_1^*) \leq \mathrm{Var}_{\boldsymbol{\theta}}(\delta_1) \quad \forall\, \boldsymbol{\theta} \in \Theta \tag{3.21}$$

Pero de acuerdo a (i) se tiene que $\delta_1^*(\mathbf{T}) = \delta(\mathbf{T})$ con probabilidad 1. Luego

$$\mathrm{Var}_{\boldsymbol{\theta}}(\delta_1^*) = \mathrm{Var}_{\theta}(\delta)$$

y luego de 3.21 resulta que

$$\mathrm{Var}_{\boldsymbol{\theta}}(\delta) \leq \mathrm{Var}_{\boldsymbol{\theta}}(\delta_1)$$

y (ii) queda demostrado.

(iii) Como $\delta^*(\mathbf{T})$ es por el Teorema 1 de la sección 3.9 insesgado, de (ii) se deduce que es un estimador IMVU para $q(\boldsymbol{\theta})$.

De acuerdo al punto (ii) de este teorema, en el caso de tener un estadístico suficiente y completo $\mathbf{T}$, cualquier estimador insesgado basado en $\mathbf{T}$ es un estimador IMVU. El punto (iii) nos indica cómo construir un estimador IMVU de $q(\boldsymbol{\theta})$ a partir de cualquier estimador insesgado.

**Teorema 2:** *Sea* $\mathbf{X}$ *un vector aleatorio cuya distribución pertenece a una familia exponencial a* $k$ *parámetros con función de densidad dada por*

$$p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta}) r_1(\mathbf{x}) + \cdots + c_k(\boldsymbol{\theta}) r_k(\mathbf{x})} h(\mathbf{x})$$

*donde* $\boldsymbol{\theta}$ *toma valores en el conjunto* $\Theta$*. Supongamos además que*

$$\Lambda = \{\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k) : \lambda_i = c_i(\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta\}$$

*contiene una esfera en* $\mathbb{R}^k$*. Sea* $\mathbf{T} = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$*, luego si* $\delta(\mathbf{T})$ *es un estimador insesgado de* $q(\boldsymbol{\theta})$*, entonces* $\delta(\mathbf{T})$ *es un estimador IMVU para* $q(\boldsymbol{\theta})$*.*

Demostración: Inmediata a partir de los Teoremas 3 de sección 3.10 y 1 de sección 3.12.

**Ejemplo 1:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $Bi(\theta, k)$ con $k$ fijo. Luego, la distribución conjunta de la muestra viene dada por

$$\begin{aligned} p(x_1, x_2, \ldots, x_n, \theta) &= \binom{k}{x_1}\binom{k}{x_2}\ldots\binom{k}{x_n}\theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{nk-\Sigma_{i=1}^n x_i} \\ &= (1-\theta)^{nk}\, e^{(\Sigma_{i=1}^n x_i)\ln(\theta/(1-\theta))}\binom{k}{x_1}\binom{k}{x_2}\ldots\binom{k}{x_n} \end{aligned}$$

Esta familia constituye una familia exponencial, con estadístico suficiente $T = \sum_{i=1}^n X_i$. Por otro lado $c(\theta) = \ln\theta/(1-\theta)$ toma todos los posibles valores de $\mathbb{R}$ cuando $\theta$ varía en el intervalo (0,1). Luego $T$ es un estadístico suficiente y completo. Como $\delta(T) = T/nk$ es un estimador insesgado de $\theta$, resulta un estimador IMVU de $\theta$.

**Ejemplo 2:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $N(\mu, \sigma^2)$. Luego usando (3.19) resulta que la distribución conjunta de la muestra viene dada por

$$p(x_1, \ldots, x_n, \mu, \sigma^2) = \frac{1}{2\pi\sigma^2)^{n/2}} e^{-\frac{n\,\mu^2}{2\sigma^2}}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n X_i}$$

Luego constituye una familia exponencial a dos parámetros con estadístico suficiente $\mathbf{T} = \Big(\sum_{i=1}^n X_i^2,\ \sum_{i=1}^n X_i\Big)$. Mostraremos ahora que $\mathbf{T}$ es completo. Bastará mostrar que

$$\Lambda = \{(\lambda_1, \lambda_2) : \lambda_1 = -\frac{1}{2\,\sigma^2},\ \lambda_2 = \frac{\mu}{\sigma^2},\ \lambda \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\}$$

contiene una esfera.

Mostraremos que $\Lambda$ contiene todo $(\lambda_1, \lambda_2) \in \mathbb{R}^2$ con $\lambda_1 < 0$.

Sea $(\lambda_1, \lambda_2)$ con $\lambda_1 < 0$, tenemos que mostrar que viene de un par $(\mu, \sigma^2)$ con $\sigma^2 > 0$. Para ver esto basta tomar $\sigma^2 = -1/2\,\lambda_1$ y $\mu = \lambda_2\sigma^2 = -\lambda_2/2\lambda_1$. Luego $\mathbf{T}$ es completo.

Como $\overline{X}$ es un estimador insesgado de $\mu$, y como depende de $\mathbf{T}$, resulta que es IMVU de $\mu$.

Por otro lado $s^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1) = \Big(\sum_{i=1}^n X_i^2 - n\overline{X}^2\Big)/(n-1)$ es un estimador insesgado de $\sigma^2$ y además depende de $\mathbf{T}$, luego es IMVU para $\sigma^2$.

**Ejemplo 3:** Sea $X_1$ una variable $N(\mu, \sigma_1^2)$ y $X_2$ independiente de $X_1$ una variable $N(\mu, \sigma_2^2)$. Vimos en el Ejemplo 4 de la sección 3.11 que

$\mathbf{T} = (X_1, X_2, X_1^2, X_2^2)$ era minimal suficiente pero no era completo. Se puede mostrar que en este caso no hay ningún estimador IMVU (ver Problema 7 de 3.11).

**Ejemplo 4:** El siguiente ejemplo muestra que no siempre existen estimadores IMVU. Volvamos al ejemplo 1 y supongamos que se quiera estimar $q(\theta)$. Como $T = \sum_{i=1}^{n} X_i$ es un estadístico suficiente, un estimador IMVU de $q(\theta)$ deberá estar basado en $T$. Supongamos que $\delta(T)$ es un estimador IMVU para $q(\theta)$. Como $T$ tiene distribución $Bi(\theta, kn)$ y $\delta(T)$ es insesgado se tendrá

$$q(\theta) = E_\theta(\delta(T)) = \sum_{i=0}^{kn} \delta(t) \binom{kn}{t} \theta^r (1-\theta)^{kn-t}$$

Luego una condición necesaria para que $q(\theta)$ tenga un estimador IMVU es que sea un polinomio de grado menor o igual a $kn$. Se puede mostrar que es también una condición suficiente aunque no lo demostraremos.

Por lo tanto no existen estimadores IMVU, por ejemplo, para $e^\theta$, $\ln\theta$, sen $\theta$. Esto no quiere decir que no existen buenos estimadores. Si $q(\theta)$ es continua, un buen estimador será $\delta(T) = q(T/nk)$ ya que $T/nk$ es un estimador IMVU de $\theta$.

**Ejemplo 5:** En este ejemplo veremos que un estimador IMVU puede ser mejorado en su error cuadrático medio por otro estimador no insesgado. Volvamos al ejemplo 2 y supongamos que se desea estimar $\sigma^2$. Hemos visto que un estimador IMVU para $\sigma^2$ es $s^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$, sin embargo veremos que $s^2$ no es admisible.
Sea $\widehat{\sigma}_c^2 = c\,U$ donde $U = \sum_{i=1}^{n}(X_i - \overline{X})^2$. Luego, $s^2 = \widehat{\sigma}^2_{\frac{1}{n-1}}$. Sabemos que $U/\sigma^2$ tiene distribución $\chi^2_{n-1}$, por lo tanto, $E_{\sigma^2}(U) = (n-1)\,\sigma^2$ y $\mathrm{Var}_{\sigma^2}(U) = 2\,(n-1)\,\sigma^4$. Con lo cual,

$$\begin{aligned}
\mathrm{ECM}_{\sigma^2}(\widehat{\sigma}_c^2) &= E_{\sigma^2}\left[(\widehat{\sigma}_c^2 - \sigma^2)^2\right] \\
&= \mathrm{Var}_{\sigma^2}(\widehat{\sigma}_c^2) + \left[E_{\sigma^2}(\widehat{\sigma}_c^2) - \sigma^2\right]^2 \\
&= c^2 \mathrm{Var}_{\sigma^2}(U) + \left[c\,E_{\sigma^2}(U) - \sigma^2\right]^2 \\
&= 2\,c^2(n-1)\,\sigma^4 + \left[c\,(n-1)\,\sigma^2 - \sigma^2\right]^2 \\
&= \sigma^4\left[c^2(n+1)(n-1) - 2(n-1)c + 1\right]
\end{aligned}$$

Luego, el ECM de $\widehat{\sigma}_c^2$ es un polinomio de grado 2 en $c$ que alcanza su mínimo cuando $c = 1/(n+1)$. Por lo tanto, $U/(n+1)$ tiene menor ECM que el estimador IMVU $s^2$.

Cómo caraterizamos los estimadores IMVU cuando no existe un estadístico suficiente y completo?

**Lema 1:** *Sea $\delta_0$ un estimador insesgado de $q(\theta)$. Dado cualquier otro estimador $\delta$ insesgado de $q(\theta)$, se cumple que $\delta = \delta_0 - U$ con $E_\theta(U) = 0$ $\forall \theta \in \Theta$.*

Luego como $ECM_\theta(\delta) = Var_\theta(\delta) = Var_\theta(\delta_0 - U) = E_\theta\{(\delta_0 - U)^2\} - q(\theta)^2$, para encontrar el estimador IMVU basta minimizar $E_\theta\{(\delta_0 - U)^2\}$, o sea, basta encontrar la proyección de $\delta_0$ sobre el espacio de los estimadores del 0.

**Teorema 3:** *Supongamos que $\mathbf{X}$ es un vector aleatorio de cuya distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Sea $\Delta = \{\delta(\mathbf{X}) : E_\theta \delta^2(\mathbf{X}) < \infty\}$. Sea $\mathcal{U} = \{\{\delta(\mathbf{X}) \in \Delta : E_\theta \delta(\mathbf{X})\} = 0 \ \forall \theta \in \Theta$. Una condición necesaria y suficiente para que $\delta \in \Delta$, insesgado, sea IMVU para $q(\theta)$ es que $E_\theta(\delta U) = 0$, $\forall \theta \in \Theta$, $\forall U \in \mathcal{U}$.*

## 3.13 Desigualdad de Rao–Cramer

En esta sección mostraremos que bajo hipótesis muy generales, la varianza de un estimador insesgado no puede ser inferior a cierta cota.

Supongamos que $\mathbf{X} = (X_1, \ldots, X_n)$ es un vector aleatorio de cuya distribución pertenece a la familia de distribuciones discreta o continua con densidad $p(\mathbf{x}, \theta)$, con $\theta \in \Theta$; donde $\Theta$ es un conjunto abierto de $\mathbb{R}$. Supongamos además que se cumplen las siguientes condiciones (en lo que sigue suponemos que $\mathbf{X}$ es continuo, para el caso discreto habrá que reemplazar todos los signos $\int$ por $\sum$):

(A) El conjunto $S = \{\mathbf{x} : p(\mathbf{x}, \theta) > 0\}$ es independiente de $\theta$.

(B) Para todo $\mathbf{x}$, $p(\mathbf{x}, \theta)$ es derivable respecto de $\theta$.

(C) Si $h(\mathbf{X})$ es un estadístico tal que $E_\theta[|h(\mathbf{X})|] < \infty$ para todo $\theta \in \Theta$ entonces se tiene

$$\frac{\partial}{\partial \theta}\left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x}\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\mathbf{x}) \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta} d\mathbf{x}$$

donde $d\mathbf{x} = (dx_1, \ldots, dx_n)$ (o sea se puede derivar dentro del signo integral)

(D)
$$0 < I(\theta) = E_\theta \left[ \left( \frac{\partial \ln p(\mathbf{X}, \theta)}{\partial \theta} \right)^2 \right] < \infty$$

$I(\theta)$ se denomina *número de información de Fisher.*

**Lema 1:** *Supongamos que se cumplan las condiciones A, B, C y D. Sea* $\psi(\mathbf{x}, \theta) = \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta}$. *Entonces,*

*(i)* $E_\theta \psi(\mathbf{X}, \theta) = 0$ *y* $Var_\theta \psi(\mathbf{X}, \theta) = I(\theta)$.

*(ii) Si además existe la derivada segunda de* $p(\mathbf{x}, \theta)$ *respecto de* $\theta$ *y si para todo estadístico* $h(\mathbf{X})$ *tal que,* $E_\theta[|h(\mathbf{X})|] < \infty$ *para todo* $\theta \in \Theta$, *se cumple que*

$$\frac{\partial^2}{\partial \theta^2} \left[ \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} \right] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(\mathbf{x}) \frac{\partial^2 p(\mathbf{x}, \theta)}{\partial \theta^2} d\mathbf{x} \tag{3.22}$$

*entonces*
$$I(\theta) = -E_\theta \frac{\partial^2 \ln p(\mathbf{X}, \theta)}{\partial \theta^2} = -E_\theta \frac{\partial \psi(\mathbf{X}, \theta)}{\partial \theta}$$

DEMOSTRACIÓN: (i) Por ser $p(\mathbf{x}, \theta)$ una densidad, si $S$ es el conjunto definido en la condición (A) se tiene

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(\mathbf{x}, \theta) d\mathbf{x} = \int_S \ldots \int p(\mathbf{x}, \theta) d\mathbf{x} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(\mathbf{x}, \theta) I_S(\mathbf{x}) d\mathbf{x} = 1$$

donde $I_S$ es la función indicadora del conjunto $S$.
Luego aplicando la condición (C) a $h(\mathbf{x}) = I_S(\mathbf{x})$ se obtiene derivando ambos miembros que

$$\int_{\infty}^{\infty} \ldots \int_{\infty}^{\infty} \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta} \, I_S(\mathbf{x}) d\mathbf{x} = 0,$$

y por lo tanto

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \left[ \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta} \Big/ p(\mathbf{x}, \theta) \right] I_S(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} = 0 \,.$$

Esta última ecuación es equivalente a

$$\int_{\infty}^{\infty} \cdots \int_{\infty}^{\infty} \left[ \frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta} \right] I_S(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} = 0$$

la cual implica

$$E_\theta \psi(\mathbf{X}, \theta) = E_\theta \left( \frac{\partial \ln p(\mathbf{X}, \theta)}{\partial \theta} \right) = 0 \tag{3.23}$$

Como $I(\theta) = E_\theta \psi^2(\mathbf{X}, \theta)$, (3.23) implica que $\mathrm{Var}_\theta \psi(\mathbf{X}, \theta) = I(\theta)$
(ii) De la igualdad

$$\frac{\partial^2 \ln p(\mathbf{x}, \theta)}{\partial \theta^2} = \frac{\frac{\partial^2 p(\mathbf{x}, \theta)}{\partial \theta^2}}{p(\mathbf{x}, \theta)} - \psi^2(\mathbf{x}, \theta)$$

se obtiene que

$$E_\theta \frac{\partial^2 \ln p(\mathbf{X}, \theta)}{\partial \theta^2} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^2 p(\mathbf{x}, \theta)}{\partial \theta^2} d\mathbf{x} - E_\theta \psi^2(\mathbf{X}, \theta)\,. \tag{3.24}$$

Utilizando (3.22) con $h(\mathbf{x}) = I_S(\mathbf{x})$ se obtiene que el primer término del lado derecho de (3.24) es igual a cero, de donde el resultado.

**Teorema 1** (Rao–Cramer): *Bajo las condiciones A, B, C y D si $\delta(\mathbf{X})$ es un estimador insesgado de $q(\theta)$ tal que $E_\theta \delta^2(\mathbf{X}) < \infty$ se tiene*

(i)
$$\mathrm{Var}_{\,\theta}(\delta(\mathbf{X})) \geq \frac{|q'(\theta)|^2}{I(\theta)}$$

(ii) (i) *vale como igualdad si y sólo si $\delta(\mathbf{X})$ es estadístico suficiente de una familia exponencial, es decir si y sólo si*

$$p(\mathbf{x}, \theta) = A(\theta) e^{c(\theta)\delta(\mathbf{x})} h(\mathbf{x}) \tag{3.25}$$

Demostración: (i) Sea $\psi(\mathbf{x}, \theta) = \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta}$. Por el Lema 1 tenemos que

$$E_\theta \psi(\mathbf{X}, \theta) = 0 \quad \text{y} \quad \mathrm{Var}_\theta \psi(\mathbf{X}, \theta) = I(\theta). \tag{3.26}$$

Por otro lado, como $\delta(\mathbf{X})$ es insesgado se tiene

$$E_\theta(\delta(\mathbf{X})) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta(\mathbf{x}) p(\mathbf{x}, \theta) I_S(\mathbf{x}) d\mathbf{x} = q(\theta)$$

y luego aplicando la hipótesis C, tomando $h(\mathbf{X}) = \delta(\mathbf{X})I_S(\mathbf{X})$ se obtiene derivando ambos miembros que

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \delta(\mathbf{x}) \frac{\partial p(\mathbf{x},\theta)}{\partial \theta} \, I_S(\mathbf{x}) d\mathbf{x} = q'(\theta)$$

de donde

$$\begin{aligned} E_\theta \left[\delta(\mathbf{X})\psi(\mathbf{X},\theta)\right] &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \delta(\mathbf{x})\psi(\mathbf{x},\theta)p(\mathbf{x},\theta)d\mathbf{x} \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \delta(\mathbf{x}) \frac{\partial \ln p(\mathbf{x},\theta)}{\partial \theta} \, I_S(\mathbf{x})p(\mathbf{x},\theta)d\mathbf{x} \\ &= q'(\theta) \end{aligned} \tag{3.27}$$

Teniendo en cuenta (3.26), (3.27) se puede escribir como

$$\mathrm{Cov}\left(\delta(\mathbf{X}),\, \psi(\mathbf{X},\theta)\right) = q'(\theta) \tag{3.28}$$

De acuerdo a la desigualdad de Cauchy–Schwartz, $\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X), \mathrm{Var}(Y)$, y vale la igualdad si y sólo si $P(Y = aX + b) = 1$ para algunas constantes $a$ y $b$. Por lo tanto, usando (3.28) resulta

$$\left[q'(\theta)\right]^2 \leq \mathrm{Var}_\theta(\delta(\mathbf{X})) \cdot \mathrm{Var}_\theta\left(\psi(\mathbf{X},\theta)\right) \tag{3.29}$$

y la igualdad vale si y sólo si

$$\frac{\partial \ln p(\mathbf{X},\theta)}{\partial \theta} = \psi(\mathbf{x},\theta) = a(\theta)\delta(\mathbf{x}) + b(\theta) \text{ con probabilidad 1.} \tag{3.30}$$

Usando (3.26) y (3.29) resulta

$$\mathrm{Var}_\theta(\delta(\mathbf{X}) \geq \frac{q'(\theta)^2}{I(\theta)} \tag{3.31}$$

que es lo que se afirma en (i).

(ii) (3.31) valdrá como igualdad si y sólo si cumple (3.30). Mostraremos que (3.30) se cumple si y sólo si se cumple (3.25).

Integrando respecto de $\theta$ en (3.30), se obtiene

$$\ln p(\mathbf{x},\theta) = \delta(\mathbf{x}) \int a(\theta)d\theta + g(\mathbf{x}) + \int b(\theta)d\theta$$

que se puede escribir como

$$\ln p(\mathbf{x},\theta) = \delta(\mathbf{x})c(\theta) + g(\mathbf{x}) + B(\theta)$$

donde $c(\theta) = \int a(\theta)d\theta$ y $B(\theta) = \int b(\theta)d\theta$. Luego, despejando $p(\mathbf{x}, \theta)$ resulta

$$p(\mathbf{x}, \theta) = e^{B(\theta)} e^{\delta(\mathbf{x})c(\theta)} e^{g(\mathbf{x})}$$

y llamando $A(\theta) = e^{B(\theta)}$ y $h(\mathbf{x}) = e^{g(\mathbf{x})}$; resulta (3.25).

Supongamos ahora que se cumple (3.25), mostraremos que se cumple (3.30).

Si se cumple (3.25), tomando logaritmos se tiene

$$\ln p(\mathbf{x}, \theta) = \ln A(\theta) + c(\theta)\delta(\mathbf{x}) + \ln h(\mathbf{x})$$

y derivando se obtiene

$$\frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta} = \frac{A'(\theta)}{A(\theta)} + c'(\theta)\delta(\mathbf{x})$$

y por lo tanto se cumple (3.30). Esto prueba el punto (ii).

**Observación 1:** Si $\delta(\mathbf{X})$ es un estimador insesgado de $q(\theta)$ y $\mathrm{Var}_\theta(\delta(\mathbf{x})) = [q'(\theta)]^2/I(\theta)$ para todo $\theta \in \Theta$. Entonces del punto (i) del Teorema 1 resulta que $\delta(\mathbf{X})$ es IMVU. Por lo tanto esto da otro criterio para verificar si un estimador insesgado dado es IMVU.

**Observación 2:** Si $p(\mathbf{x}, \theta) = A(\theta)e^{c(\theta)\delta(\mathbf{x})}h(\mathbf{x})$, y si $\delta(\mathbf{X})$ es un estimador insesgado de $q(\theta)$, entonces $\delta(\mathbf{X})$ es un estimador IMVU de $q(\theta)$. Esto resulta de (i) y (ii).

**Observación 3:** Si $\delta(\mathbf{X})$ es un estimador de $\theta$, su varianza debe ser mayor o igual que $1/I(\theta)$. Luego se puede esperar que cuanto mayor sea $I(\theta)$ (como $1/I(\theta)$ será menor) existe la posibilidad de encontrar estimadores con menor varianza y por lo tanto más precisos. De ahí el nombre de "número de información" que se le da a $I(\theta)$. Es decir cuanto mayor es $I(\theta)$, mejores estimadores de $\theta$ se pueden encontrar, y por lo tanto se puede decir que más información da el vector $\mathbf{X}$ sobre $\theta$. El hecho de que se pueden encontrar estimadores con varianza aproximadamente igual a $1/I(\theta)$ será cierto para $n$ grande. Para esto consultar sección 3.13 y el apéndice (B) de este capítulo. Para una generalización del Teorema de Rao–Cramer el caso en que $\theta$ es un vector puede consultarse el Teorema 4.3.1 de Zacks [7] y el Teorema 7.3 de Lehmann [5].

El siguiente teorema nos indica que una muestra aleatoria de tamaño $n$ $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ de una familia con densidad $p(\mathbf{X}, \boldsymbol{\theta})$ nos da $n$ veces más información que una sola observación.

**Teorema 2:** *Sea $\mathbf{X}_1, \ldots, \mathbf{X}_n$ una muestra aleatoria de una distribución con densidad $p(\mathbf{x}, \theta)$ con $\theta \in \Theta \subset I\!R$. Luego, si se denomina $I_n(\theta)$ al número de información de $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ y $I_1(\theta)$ al número de información de $\mathbf{X}_1$, entonces se tiene $I_n(\theta) = nI_1(\theta)$.*

DEMOSTRACIÓN: Se tiene que

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \theta) = \prod_{i=1}^{n} p(\mathbf{x}_i, \theta)$$

y entonces

$$\ln p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \theta) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i, \theta) \; .$$

Por lo tanto,

$$\frac{\partial \ln p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln p(\mathbf{x}_i, \theta)}{\partial \theta}$$

Con lo cual, por ser $\mathbf{X}_1, \ldots, \mathbf{X}_n$ independientes, se tiene

$$I(\theta) = \mathrm{Var}_\theta \left( \frac{\partial \ln p(\mathbf{X}_1, \ldots, \mathbf{X}_n, \theta)}{\partial \theta} \right) = \sum_{i=1}^{n} \mathrm{Var}_\theta \left( \frac{\partial \ln p(\mathbf{X}_i, \theta)}{\partial \theta} \right) = nI_1(\theta) \, .$$

**Ejemplo 1:** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $Bi(\theta, 1)$. Luego se tiene

$$p(x, \theta) = \theta^x (1 - \theta)^{1-x}$$

luego

$$\ln p(x, \theta) = x \ln \theta + (1 - x) \ln(1 - \theta)$$

y por lo tanto

$$\frac{\partial \ln p(x, \theta)}{\partial \theta} = \frac{x}{\theta} - \frac{1 - x}{1 - \theta} = \frac{x - \theta}{\theta(1 - \theta)}$$

luego

$$\begin{aligned} I_1(\theta) &= E\left( \left( \frac{\partial \ln p(X_1, \theta)}{\partial \theta} \right)^2 \right) \\ &= \left[ \frac{1}{\theta(1 - \theta)} \right]^2 E_\theta (X - \theta)^2 \\ &= \left[ \frac{1}{\theta(1 - \theta)} \right]^2 \mathrm{Var}_\theta (X - \theta)^2 \\ &= \frac{1}{\theta(1 - \theta)} \end{aligned}$$

y por lo tanto,

$$I_n(\theta) = \frac{n}{\theta(1-\theta)} \ .$$

Consideremos el estimador insesgado de $\theta$, $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$. Se tiene que

$$\mathrm{Var}_\theta(\overline{X}) = \frac{\theta(1-\theta)}{n} = \frac{1}{I(\theta)}$$

y por lo tanto, de acuerdo con la observación 2 es IMVU. Esto es un ejemplo donde el estimador IMVU satisface la desigualdad de Rao–Cramer como igualdad. Esto podríamos haberlo visto directamente mostrando que $\overline{X}$ es el estadístico suficiente de una familia exponencial.

Veremos ahora un ejemplo donde el estimador IMVU satisface la desigualdad de Rao–Cramer estrictamente.

Sea $q(\theta) = \theta(1-\theta) = \mathrm{Var}_\theta(X_1)$. Conocemos por el ejemplo 2 de la sección 3.3, que

$$\delta(X_1, X_2, \ldots, X_n) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

es un estimador insesgado de $q(\theta)$. Además se tiene

$$\begin{aligned} \delta(X_1, \ldots, X_n) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i - n\overline{X}^2\right) \\ &= \frac{n}{n-1}\overline{X}(1-\overline{X}) \end{aligned}$$

Luego $\delta(X_1, \ldots, X_n)$ *depende* del estadístico suficiente y completo $\overline{X}$ y por lo tanto es IMVU.

Sin embargo se tendrá que

$$\mathrm{Var}_\theta(\delta(X_1, \ldots, X_n)) > \frac{q'(\theta)^2}{nI_1(\theta)} \tag{3.32}$$

ya que $\delta(X_1, \ldots, X_n)$ no es el estadístico suficiente de una familia exponencial.

Para la verificación directa de (3.32) ver Problema 11 de 3.13.

## 3.14 Consistencia de estimadores

La teoría asintótica estudia las propiedades de los procedimientos de inferencia estadística cuando el tamaño de la muestra $n$ que se utiliza es grande, más precisamente, en el límite cuando $n$ tiende a infinito.

Una propiedad deseable para un estimador, es que cuando $n$ es grande la sucesión $\delta_n(X_1, \ldots, X_n)$ se aproxime en algún sentido al valor que queremos estimar. Para precisar estas ideas introduciremos el concepto de consistencia.

Sea $\mathcal{F} = \{F(x, \boldsymbol{\theta}) \text{ con } \boldsymbol{\theta} \in \Theta\}$ una familia de distribuciones y supongamos que para cada $n$ se tiene un estimador $\delta_n(X_1, \ldots, X_n)$ de $q(\boldsymbol{\theta})$ basado en una muestra aleatoria de tamaño $n$. Daremos la siguiente definición:

**Definición 1:** $\delta_n(X_1, \ldots, X_n)$ es una *sucesión fuertemente consistente de estimadores* de $q(\boldsymbol{\theta})$ si

$$\lim_{n\to\infty} \delta_n(X_1, \ldots, X_n) = q(\boldsymbol{\theta}) \qquad c.t.p.$$

o sea si $P_{\boldsymbol{\theta}}(\delta_n(X_1, \ldots, X_n) \to q(\boldsymbol{\theta})) = 1$ para todo $\boldsymbol{\theta} \in \Theta$.

**Definición 2:** $\delta_n(X_1, \ldots, X_n)$ es una *sucesión débilmente consistente de estimadores* de $q(\boldsymbol{\theta})$ si

$$\lim_{n\to\infty} \delta_n(X_1, \ldots, X_n) = q(\boldsymbol{\theta}) \quad \text{en probabilidad.}$$

Es decir, para todo $\varepsilon > 0$ y $\boldsymbol{\theta} \in \Theta$

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}(|\delta_n(X_1, \ldots, X_n) - q(\boldsymbol{\theta})| > \varepsilon) = 0\,.$$

**Observación 1:** Puesto que convergencia en c.t.p. implica convergencia en probabilidad, entonces toda sucesión fuertemente convergente también lo será débilmente.

**Ejemplo 1:** Sea $X_1, \ldots, X_n$ una muestra aleatoria de una función de distribución $F(x)$ totalmente desconocida, tal que $E_F(X_1)$ existe. Llamemos $q(F)$ a $E_F(X_1)$. Si

$$\delta_n(X_1, \ldots, X_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$

por la ley fuerte de los grandes números este estimador resulta fuertemente consistente para $q(F)$.

Si además $E_F(X^2) < \infty$, entonces

$$\delta_n(X_1, \ldots, X_n) = s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\,\overline{X}_n^2$$

es fuertemente consistente para $q(F) = \mathrm{Var}_F X_1$. En efecto,

$$s_n^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\,\overline{X}_n^2\,.$$

Por la ley fuerte de los grande números

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \to E_F(X_1^2) \quad \text{c.t.p.} \qquad \text{y} \qquad \frac{1}{n}\sum_{i=1}^{n} X_i \to E_F(X_1) \quad \text{c.t.p.}$$

Luego, $\overline{X}_n^2 \to E_F(X_1)^2$ c.t.p. y como $n/(n-1)$ converge a 1 se tiene que

$$\lim_{n\to\infty} s_n^2 = \mathrm{Var}_F(X_1) \quad \text{c.t.p.}$$

**Observación 2:** Si $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución $N(\mu, \sigma^2)$ se tiene que $\overline{X}_n$ es fuertemente consistente para $\mu$ y $s_n^2$ es fuertemente consistente para $\sigma^2$, ya que por lo visto recién

$$\lim_{n\to\infty} \overline{X}_n = E(X_1) \quad \text{c.t.p.}$$

y

$$\lim_{n\to\infty} s_n^2 = \mathrm{Var}\ (X_1) \quad \text{c.t.p.}$$

y sabemos que $E(X_1) = \mu$ y $\mathrm{Var}(X_1) = \sigma^2$.

El siguiente teorema nos da una condición suficiente para que una sucesión de estimadores sea débilmente consistente.

**Teorema 1:** *Sea, para todo $n$, $\delta_n = \delta_n(X_1, \ldots, X_n)$ un estimador de $q(\boldsymbol{\theta})$ basado en una muestra aleatoria de tamaño $n$. Si $\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n) \to 0$ y $E_{\boldsymbol{\theta}}(\delta_n) \to q(\boldsymbol{\theta})$, entonces $\delta_n(X_1, \ldots, X_n)$ es débilmente consistente.*

Demostración: Debemos ver que

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}(|\delta_n(X_1, \ldots, X_n) - q(\boldsymbol{\theta})| > \varepsilon) = 0 \quad \forall \varepsilon > 0.$$

Por la desigualdad de Markov se tiene

$$\begin{aligned} P_{\boldsymbol{\theta}}(|\delta_n(X_1,\ldots,X_n) - q(\boldsymbol{\theta})| \geq \varepsilon) &\leq \frac{E_{\boldsymbol{\theta}}(\delta_n(X_1,\ldots,X_n) - q(\boldsymbol{\theta}))^2}{\varepsilon^2} \\ &\leq \frac{\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n) + \left[E_{\boldsymbol{\theta}}(\delta_n) - q(\boldsymbol{\theta})\right]^2}{\varepsilon^2} \end{aligned}$$

Como por hipótesis $E_{\boldsymbol{\theta}}(\delta_n) - q(\boldsymbol{\theta})) \to 0$ y $(\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n)) \to 0$ se obtiene el resultado.

El siguiente teorema muestra que si $\delta_n(X_1,\ldots,X_n)$ es una sucesión de estimadores IMVU para $q(\boldsymbol{\theta})$ entonces cumple la hipótesis del Teorema 1.

**Teorema 2:** *Sea $\delta_n(X_1,\ldots,X_n)$ una sucesión de estimadores IMVU para $q(\boldsymbol{\theta})$, donde $X_1,\ldots,X_n$ es una muestra aleatoria de una distribución perteneciente a la familia $F(x,\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Luego $\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n(X_1,\ldots,X_n))$ tiende a cero si $n$ tiende a infinito.*

Demostración: Sea

$$\delta_n^*(X_1,\ldots,X_n) = \frac{\sum_{i=1}^n \delta_1(X_i)}{n}$$

luego $E_{\boldsymbol{\theta}}(\delta_n^*) = E_{\boldsymbol{\theta}}(\delta_1) = q(\boldsymbol{\theta})$, es decir $\delta_n^*$ es un estimador insesgado de $q(\boldsymbol{\theta})$.

Por otro lado, $\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n^*(X_1,\ldots,X_n)) = \mathrm{Var}_{\boldsymbol{\theta}}(\delta_1(X_1))/n$. Por ser $\delta_n(X_1,\ldots,X_n)$ IMVU de $q(\boldsymbol{\theta})$ se cumple

$$\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n(X_1,\ldots,X_n)) \leq \mathrm{Var}_{\boldsymbol{\theta}}(\delta_n^*(X_1,\ldots,X_n)) = \mathrm{Var}_{\boldsymbol{\theta}}(\delta_1(X_1))/n$$

y por lo tanto,

$$\lim_{n\to\infty} \mathrm{Var}_{\boldsymbol{\theta}}(\delta_n(X_1,\ldots,X_n)) = 0.$$

**Corolario 1:** *Si $\delta_n(X_1,\ldots,X_n)$ es una sucesión de estimadores IMVU para $q(\boldsymbol{\theta})$ donde $X_1,\ldots,X_n$ es una muestra aleatoria de una distribución perteneciente a la familia $\mathcal{F} = \{F(x,\boldsymbol{\theta}) \text{ con } \boldsymbol{\theta} \in \Theta\}$ entonces $\delta_n(X_1,\ldots,X_n)$ es una sucesión de estimadores débilmente consistentes.*

Demostración: Resulta inmediatamente de los teoremas 1 y 2.

## 3.15 Consistencia de los estimadores de los momentos

En este párrafo demostraremos la consistencia de los estimadores de los momentos.

**Teorema 3:** *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $\mathcal{F} = \{F(x,\theta)$ con $\theta \in \Theta \subset I\!R\}$, $h(x)$ una función continua con valores en $I\!R$ y supongamos que $E_\theta(h(X_1)) = g(\theta)$ es, como función de $\theta$, continua y estrictamente monótona. Sea el estimador de momentos $\widehat{\theta}_n$ definido como la solución de*

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) = E_\theta(h(X_1)) = g(\theta).$$

*Luego con probabilidad 1 existe $n_0$ tal que para todo $n \geq n_0$ la ecuación que define $\widehat{\theta}_n$ tiene solución y es fuertemente consistente para $\theta$.*

DEMOSTRACIÓN: Sea $\varepsilon > 0$. Hay que demostrar que, con probabilidad 1, existe $n_0$ tal que

$$|\widehat{\theta}_n - \theta| < \varepsilon \quad \text{para } n \geq n_0.$$

Supongamos que $g(\theta)$ es estrictamente creciente. El caso contrario se demuestra en forma análoga. Luego, se tiene,

$$g(\theta - \varepsilon) < g(\theta) < g(\theta + \varepsilon)\,.$$

Sea $\delta = \min(g(\theta + \varepsilon) - g(\theta),\, g(\theta) - g(\theta - \varepsilon))$; luego

$$g(\theta - \varepsilon) \leq g(\theta) - \delta < g(\theta) < g(\theta) + \delta \leq g(\theta + \varepsilon)\,.$$

Por otro lado, por la ley fuerte de los grandes números

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = g(\theta) \quad \text{c.t.p.}$$

Luego, con probabilidad 1, dado $\delta > 0$ existe $n_0$ tal que para todo $n \geq n_0$ se tiene

$$g(\theta) - \delta \leq \frac{1}{n}\sum_{i=1}^{n} h(X_i) \leq g(\theta) + \delta\,.$$

De esta desigualdad se infiere que

$$g(\theta-\varepsilon) \leq \frac{1}{n}\sum h(X_i) \leq g(\theta+\varepsilon) \quad \text{para } n \geq n_0$$

y como $g(\theta)$ es continua y estrictamente creciente, para $n \geq n_0$ existe un único valor $\widehat{\theta}_n$ que satisface

$$\frac{1}{n}\sum h(X_i) = E_{\widehat{\theta}_n}(h(X_1)) = g(\widehat{\theta}_n)$$

Además dicho valor debe estar entre $\theta-\varepsilon$ y $\theta+\varepsilon$, es decir que $\theta-\varepsilon \leq \widehat{\theta}_n \leq \theta+\varepsilon$ para $n \geq n_0$ que es lo que queríamos demostrar.

## 3.16 Consistencia de los estimadores de máxima verosimilitud

En esta sección enunciaremos un teorema que establece la consistencia de los estimadores de máxima verosimilitud para el caso de un solo parámetro. La demostración se dará en el Apéndice A.

$$\max_{\boldsymbol{\theta}\in\Theta} \prod_{i=1}^{n} p(x_i, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i, \widehat{\boldsymbol{\theta}}_n) \tag{3.33}$$

Se puede demostrar que bajo condiciones muy generales $\widehat{\boldsymbol{\theta}}_n$ definido por (3.33) es fuertemente consistente.

**Teorema 1:** *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución discreta o continua con densidad en la familia $p(x,\theta)$ con $\theta \in \Theta$, donde $\Theta$ es un intervalo abierto de $\mathbb{R}$. Supongamos que $p(x,\theta)$ es derivable respecto de $\theta$ y que el conjunto $S = \{x : p(x,\theta) \neq 0\}$ es independiente de $\theta$ para todo $\theta \in \Theta$. Sea $\widehat{\theta}_n$ el estimador de máxima verosimilitud de $\theta$, que satisface*

$$\sum_{i=1}^{n} \frac{\partial \ln p(x_i, \widehat{\theta}_n)}{\partial \theta} = 0 \tag{3.34}$$

*Supongamos finalmente que la ecuación (3.34) tiene a lo sumo una solución y que $\theta \neq \theta'$ implica que $p(x,\theta) \neq p(x,\theta')$. Entonces $\lim_{n\to\infty} \widehat{\theta}_n = \theta$ c.t.p., es decir, $\widehat{\theta}_n$ es una sucesión de estimadores fuertemente consistente.*

Con el objetivo de simplificar la demostración, la condiciones utilizadas en el Teorema 1 son más fuertes que las estrictamente necesarias para que el teorema sea válido. El teorema también vale en el caso de que haya más de un parámetro. Para una demostración más general se puede consultar el Teorema 5.3.1 de Zacks [7] y en Wald [6].

## 3.17 Estimadores asintóticamente normales y eficientes

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con densidad perteneciente a la familia $p(x, \theta)$ con $\theta \in \Theta$, donde $\Theta$ es un intervalo abierto de $\mathbb{R}$, y sea $\delta_n(X_1, \ldots, X_n)$ un estimador insesgado de $q(\theta)$. Luego suponiendo las condiciones A, B, C y D del Teorema 1 de la sección 3.13 se tiene

$$E_\theta[\delta_n(X_1, \ldots, X_n)] = q(\theta) \tag{3.35}$$

$$\mathrm{Var}_\theta(\delta_n(X_1, \ldots, X_n)) \geq \frac{[q'(\theta)]^2}{nI_1(\theta)} \, . \tag{3.36}$$

(3.35) y (3.36) son equivalentes a:

$$E_\theta[\sqrt{n}(\delta_n(X_1, \ldots, X_n) - q(\theta))] = 0 \tag{3.37}$$

$$\mathrm{Var}_\theta[\sqrt{n}(\delta_n(X_1, \ldots, X_n) - q(\theta))] \geq \frac{[q'(\theta)]^2}{I_1(\theta)} \, . \tag{3.38}$$

El mismo Teorema 1 de 3.13, establece que sólo excepcionalmente habrá estimadores que satisfagan simultaneamente (3.37), y la igualdad en (3.38) para $n$ finito. En efecto, esto sucede unicamente si se cumplen

$$q(\theta) = E_\theta(\delta_n(X_1, \ldots, X_n)) \quad \text{y} \quad p(\mathbf{x}, \theta) = A(\theta) e^{c(\theta)\delta_n(x_1, \ldots, x_n)} h(x_1, \ldots, x_n)$$

Sin embargo, bajo condiciones muy generales, existen estimadores (por ejemplo, los de máxima verosimilitud), que para $n$ grande satisfacen aproximadamente (3.37) y la igualdad en (3.38). Para precisar estas propiedades daremos la siguiente definición:

**Definición 1:** Se dice que $\delta_n(X_1, \ldots, X_n)$ es una sucesión de *estimadores asintóticamente normal y eficiente* (A.N.E.) si $\sqrt{n}(\delta_n(X_1, \ldots, X_n) - q(\theta))$ converge en distribución a una normal con media cero y varianza $[q'(\theta)]^2 / I_1(\theta)$.

Es decir que si $\delta_n(X_1, \ldots, X_n)$ es A.N.E., para $n$ grande se comporta aproximadamente como si tuviese distribución $N(q(\theta), [q'(\theta)]^2 / nI_1(\theta))$, es decir como si fuera insesgado con varianza $[q'(\theta)]^2 / nI_1(\theta)$, que es la menor varianza posible de acuerdo con el Teorema de Rao–Cramer.

El siguiente Teorema, demostrado en el Apéndice B, establece que bajo condiciones muy generales los estimadores de máxima verosimilitud son A.N.E.

**Teorema 1:** *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución discreta o continua con densidad perteneciente a la familia $p(x, \theta)$ con $\theta \in \Theta$ y $\Theta$ un abierto en $\mathbb{R}$. Supongamos que $p(\mathbf{x}, \theta)$ tiene derivada tercera respecto de $\theta$ continua y que satisface las condiciones A, C y D del Teorema 1 de 3.13. Sea $\psi(x, \theta) = \frac{\partial \ln p(x,\theta)}{\partial \theta}$ y supongamos además que*

$$\Big|\frac{\partial^3 \ln p(x, \theta)}{\partial \theta^3}\Big| = \Big|\frac{\partial^2 \psi(x, \theta)}{\partial \theta^2}\Big| \leq K$$

*para todo $x \in S$ y para todo $\theta \in \Theta$ ($S$ es el mismo que en la condición A). Sea $\widehat{\theta}_n$ un estimador de máxima verosimilitud de $\theta$ consistente y sea $q(\theta)$ derivable con $q'(\theta) \neq 0$ para todo $\theta$. Entonces $q(\widehat{\theta}_n)$ es A.N.E. para estimar $q(\theta)$.*

Las hipótesis que se han supuesto en este teorema son más fuertes que las estrictamente necesarias con el objetivo de simplificar la demostración. También se puede demostrar un teorema similar para el caso de más de un parámetro. Una demostración más general se puede ver en la sección 5.5 de Zacks [7].

## 3.18 Apéndice A: Demostración de la consistencia de los estimadores de máxima verosimilitud

Comenzaremos probando algunas propiedades de funciones convexas.

**Definición 1:** Sea $f(x)$ una función definida sobre un intervalo de $\mathbb{R}$ y que toma valores en $\mathbb{R}$. Diremos que $f(x)$ es *convexa* si:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y) \quad \text{con } 0 \leq \lambda \leq 1$$

y diremos que $f(x)$ es *estrictamente convexa* si:

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) \qquad 0 < \lambda < 1.$$

**Teorema 1:** *Sea $f(x) : I\!R \to I\!R$ una función convexa. Sean $\lambda_1, \ldots, \lambda_n$ tales que $0 \leq \lambda_i \leq 1$ y $\sum_{i=1}^{n} \lambda_i = 1$. Entonces se tiene:*

$$f\Big(\sum_{i=1}^{n} \lambda_i x_i\Big) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

*Además, si $f(x)$ es estrictamente convexa y hay al menos un $\lambda_i$ que cumple $0 < \lambda_i < 1$ (esto es equivalente a que haya por lo menos dos $\lambda_i > 0$), entonces:*

$$f\Big(\sum_{i=1}^{n} \lambda_i x_i\Big) < \sum_{i=1}^{n} \lambda_i f(x_i)$$

DEMOSTRACIÓN: Por inducción (para $n = 2$ se obtiene la definición 1).

**Teorema 2** (Desigualdad de Jensen): *Sea $Y$ una variable aleatoria y $h : I\!R \to I\!R$ una función convexa, luego se tiene*

$$E(h(Y)) \geq h(E(Y))$$

*Además si $h$ es estrictamente convexa y $Y$ no es constante con probabilidad 1 se tiene:*

$$E(h(Y)) > h(E(Y))$$

DEMOSTRACIÓN: Sólo haremos el caso en que $Y$ es discreta y toma un número finito de valores.

Supongamos que $Y$ toma los valores $y_1, y_2, \ldots, y_k$ con probabilidades $p_1, p_2, \ldots, p_k$. Luego aplicando el Teorema 1 se obtiene:

$$h(E(Y)) = h\Big(\sum_{i=1}^{k} y_i p_i\Big) \leq \sum_{i=1}^{n} h(y_i)p_i = E(h(Y))$$

En el caso en que $h$ sea estrictamente convexa y $Y$ no sea constante, hay al menos dos $p_i$ mayores que cero, luego también por el Teorema 1 obtenemos:

$$h(E(Y)) = h\Big(\sum_{i=1}^{k} y_i p_i\Big) < \sum_{i=1}^{k} h(y_i)p_i = E(h(Y))$$

**Teorema 3:** *Sea $f : \mathbb{R} \to \mathbb{R}$ tal que $f"(x) > 0$ para todo $x$; luego $f(x)$ es convexa.*

DEMOSTRACIÓN: Puede verse en cualquier libro de cálculo.

**Teorema 4:** *Sean $p$ y $q$ dos densidades o dos funciones de densidad discretas o continuas distintas. Luego se tiene:*

$$E_p\Big(\ln \frac{q(X)}{p(X)}\Big) < 0$$

*(donde $E_p$ significa que se calcula la esperanza considerando que $X$ tiene una distribución discreta o continua cuya densidad o probabilidad puntual es $p$).*

DEMOSTRACIÓN: Primero veremos que $q(X)/p(X)$ no es constante con probabilidad 1. La demostración se hará suponiendo que $X$ es continua. El caso discreto es totalmente análogo. Supongamos que $q(X)/p(X) = k$ c.t.p., donde $k$ es una constante. Luego $E_p(q(X)/p(X)) = k$. Esto es:

$$\int_{-\infty}^{+\infty} (q(x)/p(x))p(x)dx = k \tag{3.39}$$

pero

$$\int_{-\infty}^{+\infty} (q(x)/p(x))p(x)dx = 1 \tag{3.40}$$

pues $q(x)$ es una densidad. Luego, de (3.39) y (3.40) resulta $k = 1$. Entonces $p(X) = q(X)$ c.t.p. y esto contradice la hipótesis. Por lo tanto $q(X)/p(X)$ no es constante.

Por otro lado $-\ln(x)$ es una función estrictamente convexa ya que:

$$\frac{d^2(-\ln x)}{dx^2} = \frac{1}{x^2} > 0\ .$$

Luego, estamos en condiciones de aplicar la desigualdad de Jensen (Teorema 2), con $Y = q(X)/p(X)$ y $h(x) = -\ln x$. En estas condiciones obtenemos

$$E_p\left[-\ln \frac{q(X)}{p(X)}\right] > -\ln\left[E_p \frac{q(X)}{p(X)}\right] = -\ln \int_{-\infty}^{+\infty} \frac{q(x)}{p(x)} p(x)dx = -\ln 1 = 0.$$

Luego $E_p[-\ln(q(X)/p(X))] > 0$ y $E_p[\ln(q(X)/p(X))] < 0$ con lo que obtenemos la tesis.

**Demostración del Teorema 1 de Sección 3.16**

Sea $L_n(X_1,\ldots,X_n,\theta) = (1/n)\sum_{i=1}^n \ln p(X_i,\theta)$. Luego $\widehat{\theta}_n$ satisface $L_n(X_1,\ldots,X_n,\widehat{\theta}_n) = \max_{\theta\in\Theta} L_n(X_1,\ldots,X_n,\theta)$ y

$$\frac{\partial L_n(X_1,\ldots,X_n,\widehat{\theta}_n)}{\partial\theta} = 0\ .$$

Además se tiene

$$L_n(X_1,\ldots,X_n,\theta+\delta) - L_n(X_1,\ldots,X_n,\theta) = \frac{1}{n}\sum_{i=1}^n \ln\Big(\frac{p(X_i,\theta+\delta)}{p(X_i,\theta)}\Big) \quad (3.41)$$

$$L_n(X_1,\ldots,X_n,\theta-\delta) - L_n(X_1,\ldots,X_n,\theta) = \frac{1}{n}\sum_{i=1}^n \ln\Big(\frac{p(X_i,\theta-\delta)}{p(X_i,\theta)}\Big) \quad (3.42)$$

Como $\theta \neq \theta'$ implica $p(X_1,\theta) \neq p(X_1,\theta')$, aplicando el Teorema 4 resulta que

$$E_\theta\left(\ln\left[\frac{p(X_1,\theta+\delta)}{p(X_1,\theta)}\right]\right) < 0 \quad (3.43)$$

$$E_\theta\left(\ln\left[\frac{p(X_1,\theta-\delta)}{p(X_1,\theta)}\right]\right) < 0 \quad (3.44)$$

Entonces, de (3.41), (3.42), (3.43) y (3.44) y de la ley fuerte de los grandes números resulta que con probabilidad igual a 1 existe un $n_0$ tal que $n > n_0$ implica:

$$L_n(X_1,\ldots,X_n,\theta-\delta) < L_n(X_1,\ldots,X_n,\theta)$$

y

$$L_n(X_1,\ldots,X_n,\theta+\delta) < L_n(X_1,\ldots,X_n,\theta)\ .$$

Luego, para $n > n_0$ en el intervalo $(\theta-\delta,\theta+\delta)$ existe un máximo relativo, digamos $\theta_n^*$, que satisface

$$\frac{\partial L_n(X_1,\ldots,X_n,\theta_n^*)}{\partial\theta} = 0\ ,$$

pero hemos supuesto que $\widehat{\theta}_n$ era el único que satisfacía esta igualdad. Luego, $\widehat{\theta}_n = \theta_n^*$ y por lo tanto $\widehat{\theta}_n \in (\theta-\delta,\theta+\delta)$.

## 3.19 Apéndice B: Demostración de la normalidad y eficiencia asintótica de los estimadores de máxima verosimilitud

Demostraremos previamente un lema.

**Lema 1:** *Sea $X_1, \ldots, X_n$ una sucesión de variables aleatorias tales que $\sqrt{n}(X_n - \mu)$ converge en distribución a $N(0, \sigma^2)$. Sea $g(x)$ una función definida en $\mathbb{R}$ tal que $g'(\mu) \neq 0$ y $g'(x)$ es continua en $x = \mu$. Luego se tiene que $\sqrt{n}(g(X_n) - g(\mu))$ converge en distribución a una distribución $N(0, \sigma^2(g'(\mu))^2)$.*

DEMOSTRACIÓN: Primero demostraremos que $X_n \to \mu$ en probabilidad. Sean $\varepsilon > 0$ y $\delta > 0$ arbitrarios y $X$ una variable aleatoria con distribución $N(0, \sigma^2)$. Luego existe $K$ suficientemente grande tal que $P(|X| > K) < \delta$. Por otro lado, $P(|X_n - \mu| > \varepsilon) = P(\sqrt{n}|X_n - \mu\| \geq \sqrt{n}\varepsilon)$. Sea $n_0$ tal que $\sqrt{n_0}\varepsilon \geq K$. Luego si $n \geq n_0$:

$$P(|X_n - \mu| \geq \varepsilon) \leq P(\sqrt{n}|X_n - \mu| \geq K) \; .$$

Como $\sqrt{n}(X_n - \mu)$ converge en distribución a una variable con distribución $N(0, \sigma^2)$, se tiene

$$\lim_{n\to\infty} P(|X_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} P(\sqrt{n}|X_n - \mu| \geq K) = P(|X| \geq K) < \delta \; .$$

Luego

$$\lim_{n\to\infty} (P|X_n - \mu| \geq \varepsilon) < \delta \quad \text{para todo } \delta > 0 \; ,$$

por lo tanto, $\lim_{n\to\infty} P(|X_n - \mu| \geq \varepsilon) = 0$ y resulta $X_n \to \mu$ en probabilidad.

Por otra parte, el teorema del valor medio implica que

$$\sqrt{n}(g(X_n) - g(\mu)) = \sqrt{n}g'(\xi_n)(X_n - \mu) \tag{3.45}$$

con $\xi_n$ un punto intermedio entre $X_n$ y $\mu$. Luego, $\xi_n \to \mu$ en probabilidad y como $g'(x)$ es continua en $\mu$, $g'(\xi_n) \to g'(\mu)$ en probabilidad.

Por lo tanto, como por hipótesis $\sqrt{n}(X_n - \mu)$ converge en distribución a una $N(0, \sigma^2)$ y $g'(\xi_n) \to g'(\mu)$ en probabilidad, aplicando la propiedad 5 de 1.8, resulta que $\sqrt{n}(g(X_n) - g(\mu))$ converge en distribución a una $N(0, \sigma^2(g'(\mu))^2)$.

**Demostración del Teorema 1 de la sección 3.17.** Indiquemos por

$$\psi'(x,\theta) = \frac{\partial \psi(x,\theta)}{\partial \theta} \quad \text{y} \quad \psi''(x,\theta) = \frac{\partial^2 \psi(x,\theta)}{\partial \theta^2} \ .$$

El estimador de máxima verosimilitud satisface:

$$\sum_{i=1}^{n} \psi(X_i, \widehat{\theta}_n) = 0 \ .$$

Desarrollando en serie de Taylor alrededor de $\theta$ se obtiene:

$$\sum_{i=1}^{n} \psi(X_i,\theta) + \left(\sum_{i=1}^{n} \psi'(X_i,\theta)\right)(\widehat{\theta}_n - \theta) + \frac{1}{2}\left(\sum_{i=1}^{n} \psi''(X_i,\xi_n)\right)(\widehat{\theta}_n - \theta)^2 = 0 \ ,$$

donde $\xi_n$ es un punto intermedio entre $\widehat{\theta}_n$ y $\theta$. Despejando $(\widehat{\theta}_n - \theta)$ y multiplicando ambos miembros por $\sqrt{n}$ se obtiene:

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \frac{-\sum_{i=1}^{n} \psi(X_i,\theta)/\sqrt{n}}{(1/n)\sum_{i=1}^{n} \psi'(X_i,\theta) + (1/2n)\Big(\sum_{i=1}^{n} \psi''(X_i,\xi_n)\Big)(\widehat{\theta}_n - \theta)}$$

Sea $D(X_1,\ldots,X_n,\theta)$ el denominador de esta última expresión. Vamos a demostrar que:

(a)

$$D(X_1,\ldots,X_n,\theta) \to -I_1(\theta) = -E_\theta\left[\psi(X,\theta)\right]^2 \quad \text{en probabilidad.}$$

(b) $\sum_{i=1}^{n} \psi(X_i,\theta)/\sqrt{n}$ converge en distribución a una distribución $N(0, I_1(\theta))$

Probemos (a). Como $|\psi''(X_i,\theta)| \le K$ para todo $\theta$, se tiene que

$$\left|\frac{1}{2n}\sum_{i=1}^{n} \psi''(X_i,\xi_n)(\widehat{\theta}_n - \theta)\right| \le \frac{K}{2}|(\widehat{\theta}_n - \theta)|$$

y luego como $\widehat{\theta}_n$ es consistente se deduce que:

$$\frac{1}{n}\sum_{i=1}^{n} \psi''(X_i,\xi_n)(\widehat{\theta}_n - \theta) \to 0 \quad \text{en probabilidad.} \tag{3.46}$$

Por otro lado, como $\psi'(X_i,\theta)$ son $n$ variables aleatorias, independientes igualmente distribuidas, por la ley de los grandes números implica que

$$\frac{1}{n}\sum_{i=1}^{n} \psi'(X_i,\theta) \to E(\psi'(X_1,\theta) \quad \text{en probabilidad.} \tag{3.47}$$

Pero de acuerdo con el Lema 1 de la sección 3.13

$$E_\theta(\psi'(X_1, \theta)) = -I_1(\theta) \ .$$

Luego, usando (3.46) y (3.47) se obtiene:

$$D(X_1, \ldots, X_n, \theta) \to -I_1(\theta) \quad \text{en probabilidad,}$$

con lo que queda probado (a). Para probar (b) observemos que, como las variables aleatorias

$$\psi(X_i, \theta) = \frac{\partial \ln p(X_i, \theta)}{\partial \theta}$$

son independientes e igualmente distribuidas con esperanza 0 y varianza $I_1(\theta)$ (ver Lema 1 de la sección 3.13), por el Teorema Central del límite

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, \theta)$$

converge en distribución a $\mathrm{N}(0, I_1(\theta))$.

Luego $\sqrt{n}(\widehat{\theta}_n - \theta)$ converge en distribución a una ley $N(0, I_1(\theta)/(I_1(\theta)^2)$ o sea $N(0, 1/I_1(\theta))$.

Consideremos ahora el estimador de máxima verosimilitud $q(\theta)$ dado por $q(\widehat{\theta}_n)$.

Por el Lema 1 se tendrá que $\sqrt{n}(q(\widehat{\theta}_n) - q(\theta))$ converge en distribución a una $N(0, (q'(\theta))^2/I_1(\theta))$.

## REFERENCIAS DEL CAPITULO 3

[1 ] Bahadur, R.R. (1954). Sufficiency and Statistical Decision Functions. *Annals of Mathematical Statistics* **25**, 423–462.

[2 ] Draper, N. and Smith, H. (1966). *Applied Regression Analysis.* J. Wiley & Sons.

[3 ] Dynkin, E.B. (1961). Necessary and Sufficient Statistics for Families of Distributions. *Selected Translations of Mathematical Statistics and Probability* **1**, 23–41.

[4 ] Lehmann, E.L. (1994). *Testing Statistical Hypothesis.* Chapman & Hall.

[5 ] Lehmann, E.L. (1983). *Theory of Point Estimation.* J. Wiley & Sons.

[6 ] Wald, A.N. (1949). Note on the Consistency of the Maximum Likelihood Estimates. *Annals of Mathematical Statistics* **20**, 595–601.

[7 ] Zacks, S. (1971). *The Theory of Statistical Inference.* J. Wiley & Sons.