Capítulo 1

Elementos de Inferencia

1.1. Suficiencia

Considere $X_1, \ldots, X_n$ variables aleatorias IID desde $\mathsf{Exp}(\theta)$, de este modo

$$f(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} \theta \exp(-\theta x_i) = \theta^n \exp\Big(- \theta \sum_{i=1}^{n} x_i \Big) = \theta^n \exp(-\theta n \overline{x}).$$

Es decir, para esta densidad conjunta **sólo** necesitamos conocer el tamaño muestral y la media muestral.

IDEA. Hemos reducido la información contenia en las n variables a una **única** estadística $T(X_1, \ldots, X_n)$.

Note que $T : \mathcal{X}^n \to \mathbb{R}$ reduce una colección de n observaciones a un único número y por tanto no puede ser inyectiva. Es decir, en general $T(X_1, \ldots, X_n)$ provee **menos** información sobre θ que $(X_1, \ldots, X_n)$.
Para algunos modelos una estadística T será igualmente informativa sobre θ que la muestra $(X_1, \ldots, X_n)$. Tales estadísticas son llamadas *estadísticas suficientes* (es suficiente usar T en lugar de $(X_1, \ldots, X_n)$).

DEFINICIÓN 1.1 (Suficiencia). Sea $X_1, \ldots, X_n$ variables aleatorias IID desde el modelo $\{\mathsf{P}_\theta : \theta \in \Theta\}$. Una estadística $T : \mathcal{X}^n \to \mathbb{R}$ se dice suficiente para θ, si

$$\mathsf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n | T = t),$$

no depende de θ, para todo $(x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ y todo $t \in \mathbb{R}$.

EJEMPLO 1.2. Suponga $X_1, \ldots, X_n$ variables aleatorias IID desde $\mathsf{Ber}(\theta)$, donde $\theta \in (0, 1)$. Aquí $\mathcal{X} = \{0, 1\}$ mientras que $\Theta = (0, 1)$. Considere

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

sus valores son denotados como $t \in \mathcal{T} = \{0, 1, \ldots, n\}$. Ahora, note que la distribución conjunta de $X_1, \ldots, X_n$ es dada por

$$p(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i} = \theta^{\sum_{i=1}^{n} x_i} (1 - \theta)^{n - \sum_{i=1}^{n} x_i}.$$

Por otro lado, sabesmos que

$$T \sim \mathsf{Bin}(n, \theta),$$

con probabilidad

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

De este modo,

$$\begin{aligned}\mathsf{P}(X_1 = x_1, \dots, X_n = x_n | T = t) &= \frac{\mathsf{P}(X_1 = x_1, \dots, X_n = x_n, T = t)}{\mathsf{P}(T = t)} \\ &= \frac{\mathsf{P}(\{\cap_{i=1}^n X_i = x_i\} \cap \{T = t\})}{\mathsf{P}(T = t)} = \frac{\mathsf{P}(X_1 = x_1, \dots, X_n = x_n)}{\mathsf{P}(T = t)} \\ &= \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}.\end{aligned}$$

Es decir, conocer $(X_1, \dots, X_n)$ además de conocer $T(X_1, \dots, X_n)$ no añade información sobre θ.

TEOREMA 1.3 (Factorización de Fisher-Neyman). *Suponga que $X_1, \dots, X_n$ tiene densidad conjunta $f(\boldsymbol{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Una estadística $T : \mathcal{X}^n \to \mathbb{R}$ es suficiente para $\boldsymbol{\theta}$ si y solo si, existe $g : \mathbb{R} \times \Theta \to \mathbb{R}$ y $h : \mathcal{X} \to \mathbb{R}$ tal que*

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = g(T(x_1, \dots, x_n); \boldsymbol{\theta}) h(\boldsymbol{x}).$$

DEMOSTRACIÓN. En Casella y Berger (2002, pág. 276), se presenta una demostración para el caso discreto. En el caso continuo una prueba usando Teoría de la Medida es dada en Lehmann (1986, pág. 54). □

EJEMPLO 1.4. Sea $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ variables IID desde una distribución $\mathsf{Geo}(\theta)$. De este modo, la densidad conjunta asume la forma:

$$p(\boldsymbol{x}; \theta) = \prod_{i=1}^n \theta(1-\theta)^{x_i} = \theta^n (1-\theta)^{\sum_{i=1}^n x_i},$$

para $x_i \in \{0, 1, \dots\}$. Aplicando el resultado anterior con

$$g(T(\boldsymbol{x}); \theta) = \theta^n (1-\theta)^{T(\boldsymbol{x})}, \qquad h(\boldsymbol{x}) = 1,$$

sigue que $T(\boldsymbol{x}) = \sum_{i=1}^n X_i$ es estadística suficiente.

EJEMPLO 1.5. Sea $X_1, \dots, X_n$ una $\mathsf{m.a.}(n)$ desde $\mathsf{U}(a, b)$ con $\boldsymbol{\theta} = (a, b)^\top$ $(a < b)$. La densidad conjunta es dada por:

$$f(\boldsymbol{x}; a, b) = \prod_{i=1}^n \frac{1}{b-a} I_{[a,b]}(x_i) = \frac{1}{(b-a)^n} \prod_{i=1}^n I_{[a,b]}(x_i)$$

Ahora,

$$\begin{aligned}\prod_{i=1}^n I_{[a,b]}(x_i) = 1 \quad &\iff \quad a \le x_i \le b, \ \forall\, i \\ &\iff \quad a \le x_{(1)} \le x_{(n)} \le b.\end{aligned}$$

Es decir, podemos escribir la densidad conjunta como

$$f(\boldsymbol{x}; a, b) = \frac{1}{(b-a)^n} I_{[a,\infty)}(x_{(1)}) I_{(-\infty,b]}(x_{(n)}).$$

De este modo, $\boldsymbol{T}(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ es suficiente para (a, b).

1.2. Función de verosimilitud

DEFINICIÓN 1.6 (Función de verosimilitud). Para una observación $\boldsymbol{x}$ fijada de un vector aleatorio $\boldsymbol{X}$ con densidad $f(\cdot;\boldsymbol{\theta})$. La función de verosimilitud

$$L(\cdot;\boldsymbol{x}) : \Theta \to \mathbb{R}_+,$$

es definida como

$$L(\boldsymbol{\theta};\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta.$$

OBSERVACIÓN. La verosimilitud corresponde a la **densidad conjunta** de los datos que se desea analizar.

EJEMPLO 1.7. Sea $X_1, \ldots, X_n$ variables aleatorias IID con distribución $\mathsf{N}(\theta, 1)$. Entonces,

$$\begin{aligned} L(\theta;\boldsymbol{x}) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\big\{ -\tfrac{1}{2}(x_i-\theta)^2 \big\} = (2\pi)^{-n/2} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 \Big\} \\ &\propto \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 \Big\}. \end{aligned}$$

Ahora,

$$\sum_{i=1}^{n}(x_i-\theta)^2 = \sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\theta)^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\theta)^2,$$

pues $\sum_{i=1}^{n}(x_i-\overline{x})(\overline{x}-\theta) = 0$. De este modo,

$$\begin{aligned} L(\theta;\boldsymbol{x}) &\propto \exp\Big\{ -\frac{1}{2}\Big[\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\theta)^2 \Big] \Big\} \\ &\propto \exp\Big\{ -\frac{n}{2}(\overline{x}-\theta)^2 \Big\}. \end{aligned}$$

Es decir, $L(\theta;\boldsymbol{x})$ es proporcional a una densidad $\mathsf{N}_1(\overline{x}, 1/n)$.

Es conveniente usar la función de log-verosimilitud dada por

$$\ell(\boldsymbol{\theta};\boldsymbol{x}) = \log L(\boldsymbol{\theta};\boldsymbol{x}) = \log f(\boldsymbol{x};\boldsymbol{\theta}).$$

Note también que para $\boldsymbol{\theta} \in \Theta$ fijado,

$$L(\boldsymbol{\theta};\boldsymbol{X}), \qquad \text{y} \qquad \ell(\boldsymbol{\theta};\boldsymbol{X}),$$

corresponden a variables aleatorias.

EJEMPLO 1.8. Sea $X \sim \mathsf{N}(\theta, 1)$ y $Y \sim \mathsf{N}(2\theta, 1)$ independientes. Entonces,

$$\ell(\theta; X) = -\frac{1}{2}\log 2\pi - \frac{1}{2}(X-\theta)^2,$$

de este modo, la variable aleatoria

$$-2\ell(\theta; X) - \log 2\pi = (X-\theta)^2 \sim \chi^2(1).$$

Análogamente,

$$\ell(\theta; X, Y) = \log f(X;\theta) + \log f(Y;\theta) = -\log 2\pi - \frac{1}{2}(X-\theta)^2 - \frac{1}{2}(Y-2\theta)^2.$$

Es decir,

$$-2\ell(\theta; X) - 2\log 2\pi \sim \chi^2(2).$$

Observación. Considere dos conjuntos $\boldsymbol{x}, \boldsymbol{y}$ independientes, con densidades $f_1(\boldsymbol{x};\boldsymbol{\theta})$ y $f_2(\boldsymbol{x};\boldsymbol{\theta})$ que comparten un parámetro común $\boldsymbol{\theta}$. Entonces la verosimilitud de los datos combinados es:

$$L(\boldsymbol{\theta}) = f_1(\boldsymbol{x};\boldsymbol{\theta}) f_2(\boldsymbol{x};\boldsymbol{\theta}) = L_1(\boldsymbol{\theta}) L_2(\boldsymbol{\theta}).$$

Además, la función de log-verosimilitud

$$\ell(\boldsymbol{\theta}) = \log L_1(\boldsymbol{\theta}) + \log L_2(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\theta}) + \ell_2(\boldsymbol{\theta}).$$

El caso más simple, es para una muestra de vectores aleatorios IID. En cuyo caso, tenemos

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i;\boldsymbol{\theta}), \qquad \text{y} \qquad \ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}).$$

1.3. Función score e información de Fisher

A continuación definimos cantidades que surgen a partir de la log-verosimilitud. Considere los siguientes *condiciones de regularidad*

Supuesto A1. *Las distribuciones* $\{\mathsf{P}_\theta : \boldsymbol{\theta} \in \Theta\}$ *tienen soporte común, de modo que el conjunto*

$$A = \{\boldsymbol{x} : f(\boldsymbol{x};\boldsymbol{\theta}) \geq 0\},$$

no depende de $\boldsymbol{\theta}$.

Observación. Distribuciones pertenecientes a la FE satisfacen la condición anterior.

Ejemplo 1.9 (Contraejemplos). Considere $X \sim \mathsf{U}(0,\theta)$, con $\theta \in (0,\infty)$ cuya densidad es

$$f(x;\theta) = \frac{1}{\theta} I_{[0,\theta]}(x).$$

También la familia de distribuciones exponencial con dos parámetros $Y \sim \mathsf{Exp}(a,b)$,

$$f(y;a,b) = \frac{1}{b} \exp\Big(-\frac{(x-a)}{b}\Big) I_{[a,\infty)}(y), \qquad a,b>0.$$

Supuesto A2. *El espacio paramétrico* $\Theta \subset \mathbb{R}^p$ *es un conjunto abierto.*

Supuesto A3. *Para todo* $\boldsymbol{x} \in A$ *la función de log-verosimilitud es 3-veces continuamente diferenciable con respecto a* $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^\top$.

Definición 1.10 (Función score). Suponga las condiciones A1 a A3 para todo $\boldsymbol{x} \in A$, se define la *función score* como el vector de derivadas parciales de la log-verosimilitud

$$\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x}) = \frac{\partial \ell(\boldsymbol{\theta};\boldsymbol{x})}{\partial \boldsymbol{\theta}} = \Big(\frac{\partial \ell(\boldsymbol{\theta};\boldsymbol{x})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta};\boldsymbol{x})}{\partial \theta_p}\Big)^\top.$$

Supuesto A4. *Suponga que existen funciones integrables* $F_1(x)$, $F_2(x)$ *y* $H(x)$ *tal que*

$$\int H(\boldsymbol{x}) f(\boldsymbol{x};\boldsymbol{\theta})\, \mathsf{d}\boldsymbol{x} < M,$$

para $M \in \mathbb{R}$ un valor apropiado y que se satisface

$$\Big|\frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_i}\Big| < F_1(\boldsymbol{x}), \qquad \Big|\frac{\partial^2 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\Big| < F_2(\boldsymbol{x}),$$
$$\Big|\frac{\partial^3 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k}\Big| < H(\boldsymbol{x}), \qquad i,j,k = 1,\ldots,p.$$

Esta condición implica que podemos intercambiar las operaciones de integración y diferenciación. Por ejemplo,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_A f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} = \int_A \frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}.$$

RESULTADO 1.11. *Bajo las condiciones A1 a A4, tenemos*

$$\mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{X})\} = \mathbf{0}, \qquad \forall\, \boldsymbol{\theta} \in \Theta$$

DEMOSTRACIÓN. Tenemos

$$\begin{aligned}
\mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{X})\} &= \int_A \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x};\boldsymbol{\theta})\, f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} \\
&= \int_A \frac{1}{f(\boldsymbol{x};\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta})\, f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} \\
&= \int_A \frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} = \frac{\partial}{\partial \boldsymbol{\theta}} \int_A f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} \\
&= \mathbf{0}
\end{aligned}$$

□

EJEMPLO 1.12. Considere $X \sim \mathsf{N}(\theta, 1)$. Entonces,

$$U(\theta; X) = \frac{\partial}{\partial \theta}\Big[-\frac{1}{2}\log 2\pi - \frac{1}{2}(X-\theta)^2\Big] = X - \theta.$$

De este modo, $U(\theta; X) \sim \mathsf{N}(0, 1)$. Así, es directo

$$\mathsf{E}\{U(\theta; X)\} = \mathsf{E}(X - \theta) = 0.$$

DEFINICIÓN 1.13 (Matriz de información de Fisher). Suponga las condiciones A1 a A3. Entonces la *matriz de información de Fisher* se define como:

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{Cov}_\theta\{\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{X})\} = \mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{X})\boldsymbol{U}^\top(\boldsymbol{\theta};\boldsymbol{X})\}.$$

Es decir, $\boldsymbol{\mathcal{F}}(\boldsymbol{\theta})$ tiene elementos

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = \mathsf{E}_\theta\Big\{\frac{\partial}{\partial \theta_i}\ell(\boldsymbol{\theta};\boldsymbol{X})\frac{\partial}{\partial \theta_i}\ell(\boldsymbol{\theta};\boldsymbol{X})\Big\}.$$

OBSERVACIÓN. En ocasiones escribimos $\boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta})$ pero la información **no** es aleatoria.

EJEMPLO 1.14. Sean $X_1, \ldots, X_n$ variables aleatorias $\mathsf{N}(\mu, \sigma^2)$ con σ^2 conocido. Entonces,

$$L(\mu) \propto \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\Big\},$$

así

$$\ell(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 + c,$$

con c una contante. Además,

$$U(\mu; \boldsymbol{X}) = \dot{\ell}(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu).$$

De este modo,

$$\mathcal{F}_n(\mu) = \mathsf{var}\{U(\mu; \boldsymbol{X})\} = \frac{1}{\sigma^4}\sum_{i=1}^{n}\mathsf{var}(X_i - \mu) = \frac{n}{\sigma^2}.$$

INTERPRETACIÓN. Los datos contienen más información sobre μ si:

(a) σ^2 es pequeño ($\sigma^2 \to 0$).
(b) conforme n crece ($n \to \infty$).

RESULTADO 1.15. *Suponga las condiciones A1 a A4. Entonces,*

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{E}_\theta\{-\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{X})\} = \mathsf{E}_\theta\Big\{-\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big\}.$$

DEMOSTRACIÓN. Tenemos que

$$\begin{aligned}
\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}) &= \frac{\partial \log f(\boldsymbol{x}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = \frac{\partial}{\partial\boldsymbol{\theta}}\Big\{\frac{\partial}{\partial\boldsymbol{\theta}^\top}\log f(\boldsymbol{x}; \boldsymbol{\theta})\Big\} = \frac{\partial}{\partial\boldsymbol{\theta}}\Big\{\frac{1}{f(\boldsymbol{x}; \boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}^\top} f(\boldsymbol{x}; \boldsymbol{\theta})\Big\} \\
&= \frac{1}{f^2(\boldsymbol{x}; \boldsymbol{\theta})}\Big\{\frac{\partial^2 f(\boldsymbol{x}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} f(\boldsymbol{x}; \boldsymbol{\theta}) - \frac{\partial}{\partial\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta})\,\frac{\partial}{\partial\boldsymbol{\theta}^\top} f(\boldsymbol{x}; \boldsymbol{\theta})\Big\} \\
&= \frac{1}{f(\boldsymbol{x}; \boldsymbol{\theta})}\frac{\partial^2 f(\boldsymbol{x}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} - \Big[\frac{1}{f(\boldsymbol{x}; \boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta})\Big]\Big[\frac{1}{f(\boldsymbol{x}; \boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}^\top} f(\boldsymbol{x}; \boldsymbol{\theta})\Big].
\end{aligned}$$

Por otro lado, note que

$$\begin{aligned}
\mathsf{E}_\theta\Big\{\frac{1}{f(\boldsymbol{x}; \boldsymbol{\theta})}\frac{\partial^2 f(\boldsymbol{x}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big\} &= \int_A \frac{1}{f(\boldsymbol{x}; \boldsymbol{\theta})}\frac{\partial^2 f(\boldsymbol{x}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} f(\boldsymbol{x}; \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} \\
&= \int_A \frac{\partial^2 f(\boldsymbol{x}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\,\mathrm{d}\boldsymbol{x} = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\int_A f(\boldsymbol{x}; \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} \\
&= \mathbf{0}.
\end{aligned}$$

De este modo,

$$\begin{aligned}
\mathsf{E}_\theta\{-\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{X})\} &= \mathsf{E}_\theta\Big\{\Big[\frac{1}{f(\boldsymbol{x}; \boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta})\Big]\Big[\frac{1}{f(\boldsymbol{x}; \boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}^\top} f(\boldsymbol{x}; \boldsymbol{\theta})\Big]\Big\} \\
&= \mathsf{E}_\theta\Big\{\Big[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{x}; \boldsymbol{\theta})\Big]\Big[\frac{\partial}{\partial\boldsymbol{\theta}^\top}\log f(\boldsymbol{x}; \boldsymbol{\theta})\Big]\Big\} \\
&= \mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{X})\boldsymbol{U}^\top(\boldsymbol{\theta}; \boldsymbol{X})\} = \mathsf{Cov}(\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{X})).
\end{aligned}$$

□

OBSERVACIÓN. Este resultado permite obtener la matriz de información de Fisher de dos maneras equivalente en **modelos regulares** (esto es, bajo los Supuestos A1 a A4). Es decir,

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{X})\boldsymbol{U}^\top(\boldsymbol{\theta}; \boldsymbol{X})\} = \mathsf{E}_\theta\Big\{-\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big\}.$$

DEFINICIÓN 1.16. La matriz

$$\boldsymbol{J}(\boldsymbol{\theta};\boldsymbol{X}) = -\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top},$$

se denomina *información observada*.

OBSERVACIÓN. En efecto,

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{E}_\theta\{\boldsymbol{J}(\boldsymbol{\theta};\boldsymbol{X})\},$$

que también es llamada *matriz de información esperada*.

EJEMPLO 1.17. Considere $X \sim \mathsf{Bin}(n,\theta)$ con $\theta \in (0,1)$. De este modo,

$$\ell(\theta;x) = \log\binom{n}{x} + (n-x)\log(1-\theta) + x\log\theta,$$

así,

$$U(\theta;x) = -\frac{n-x}{1-\theta} + \frac{x}{\theta},$$

obteniendo la derivada de $U(\theta;x)$,

$$U'(\theta;x) = -\frac{n-x}{(1-\theta)^2} - \frac{x}{\theta^2}.$$

Además, como $\mathsf{E}(X) = n\theta$, sigue que

$$\begin{aligned}\mathcal{F}(\theta) = \mathsf{E}_\theta\{-U'(\theta;X)\} &= \frac{n-\mathsf{E}(X)}{(1-\theta)^2} + \frac{\mathsf{E}(X)}{\theta^2} = \frac{n-n\theta}{(1-\theta)^2} + \frac{n\theta}{\theta} \\ &= n\Big(\frac{1}{1-\theta} + \frac{1}{\theta}\Big) = \frac{n}{\theta(1-\theta)}.\end{aligned}$$

Capítulo 2

Estimación

Suponga que tenemos $X_1, \ldots, X_n$ una muestra aleatoria desde $\mathcal{P} = \{\mathsf{P}_\theta : \boldsymbol{\theta} \in \Theta\}$. Deseamos entender el mecanismo o modelo verdadero (basado en $\boldsymbol{\theta}_0 \in \Theta$) que generó los datos. A continuación presentamos ideas básicas sobre procedimientos para seleccionar $\boldsymbol{\theta}$ desde los datos (estimación) así como discutir las propiedades de tales procedimientos.

Definición 2.1. Una función $\boldsymbol{T} : \mathcal{X}^n \to \Gamma$ es llamado un estimador (puntual). Este es usado para estimar $\boldsymbol{\gamma} = \boldsymbol{g}(\boldsymbol{\theta})$.

Un estimador es una regla que provee un valor plausible sobre el verdadero $\boldsymbol{\gamma}$ (equivalentemente $\boldsymbol{\theta}$) que generó los datos. El valor $\boldsymbol{T}(\boldsymbol{x})$ es llamado estimación de $\boldsymbol{g}(\boldsymbol{\theta})$ y corresponde a una realización de la variable aleatoria $\boldsymbol{T}(\boldsymbol{X})$.

Observación. Usualmente anotamos un estimador (y una estimación) como $\widehat{\boldsymbol{\gamma}} = \boldsymbol{T}(X_1, \ldots, X_n)$ y distinguimos el método usado como $\widehat{\boldsymbol{\gamma}}_{\mathsf{ML}}$, $\widehat{\boldsymbol{\gamma}}_{\mathsf{MM}}$ o $\widehat{\boldsymbol{\gamma}}_{\mathsf{LS}}$.

2.1. Métodos de estimación

2.1.1. Método de los momentos. Este procedimiento **no** requiere conocer la distribución subyacente de la variable aleatoria de interés X ($\sim \mathsf{P}_\theta \in \mathcal{P}$), sino que requiere **asumir** formas específicas para sus momentos. El objetivo es substituir estos momentos por sus contrapartes empíricas.

Para formalizar el procedimiento, considere $X_1, \ldots, X_n$ una $\mathsf{m.a.}(n)$ desde P_θ (unidimensional). Es decir,

$$\mathcal{P} = \{\mathsf{P}_\theta^{\otimes n} : \boldsymbol{\theta} \in \Theta\},$$

y sea

$$\mu_k = \mu_k(\mathsf{P}_\theta) = \mathsf{E}(X^k) = \int x^k \, \mathsf{d}\, \mathsf{P}_\theta = \int x^k f_X(x; \boldsymbol{\theta}) \, \mathrm{d}x.$$

Además, suponga que P_θ tiene momentos finitos $\mu_1, \ldots, \mu_r$ para algún r.

Asumiremos también que el parámetro de interés γ depende de $\boldsymbol{\theta}$ a través de los momentos μ_k como:

$$\gamma = h(\mu_1(\mathsf{P}_\theta), \ldots, \mu_r(\mathsf{P}_\theta)),$$

donde h es una función conocida. Esto lleva a la siguiente definición

Definición 2.2 (Estimador de momentos). Suponga $X_1, \ldots, X_n$ que sigue el modelo estadístico $\{\mathsf{P}_\theta^{\otimes n} : \boldsymbol{\theta} \in \Theta\}$. El estimador de momentos es definido como

$$\widehat{\gamma}_{\mathsf{MM}} = h(m_1, \ldots, m_r),$$

donde $m_k = \widehat{\mu}_k$ es el momento empírico (o muestral) de orden k, dado por

$$m_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k.$$

EJEMPLO 2.3. Considere $X_1, \ldots, X_n$ una muestra aleatoria desde P_θ y considere

$$\gamma = \int x f(x;\theta)\,\mathsf{d}x.$$

Usando el método de momentos, tenemos que:

$$\widehat{\gamma}_{\mathsf{MM}} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

EJEMPLO 2.4. Considere que el parámetro de interés es la desviación estándar σ. Note que

$$\sigma^2 = \int (x-\mu_1)^2 f(x;\boldsymbol{\theta})\,\mathsf{d}x,$$

de este modo,

$$\sigma = \sqrt{\mu_2 - \mu_1^2} = h(\mu_1, \mu_2),$$

cuyo estimador usando el método de momentos adopta la forma:

$$\widehat{\sigma}_{\mathsf{MM}} = \sqrt{m_2 - m_1^2}.$$

Note que

$$\begin{aligned} m_2 - m_1^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big)^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(n\overline{x})^2\Big) \\ &= \frac{1}{n}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \end{aligned}$$

De este modo,

$$\widehat{\sigma}_{\mathsf{MM}} = \Big\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\Big\}^{1/2},$$

es decir, $\widehat{\sigma}_{\mathsf{MM}} \neq s$ (desviación estándar muestral).

EJEMPLO 2.5. El sesgo de una variable aleatoria X con distribución F es definida como

$$\gamma = \frac{\mathsf{E}_F(X - \mathsf{E}_F(X))^3}{(\mathsf{var}_F(X))^{3/2}} = \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{(\mu_2 - \mu_1^2)^{3/2}},$$

así el estimador de momentos asume la forma

$$\widehat{\gamma}_{\mathsf{MM}} = \frac{m_3 - 3m_2m_1 + 2m_1^3}{(m_2 - m_1^2)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\}^{3/2}}$$

EJEMPLO 2.6. La función de distribución acumulada es definida como

$$F(t) = \mathsf{P}_F((-\infty, t]),$$

para t fijo. Un estimador natural para la probabilidad del conjunto $(-\infty, t]$ es la frecuencia relativa,

$$\widehat{F}_n(t) = \widehat{F}_n(t;\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(x_i).$$

$\widehat{F}_n$ se denomina la *función de distribución empírica*. Note que

$$m_k = \mu_k(\widehat{F}_n) = \int x^k\,\mathsf{d}\widehat{F}_n,$$

es decir, $\widehat{F}_n$ es un estimador de momentos de F.

OBSERVACIÓN. En los ejemplos anteriores **no** hemos asumido alguna distribución específica para P_θ. Es efecto, el método de momentos se puede entender como un procedimiento libre de distribución y por tanto corresponde a un método **no paramétrico**.

EJEMPLO 2.7. Sea $X_1, \ldots, X_n$ una muestra aleatoria desde una distribución log-normal con vector de parámetros $\boldsymbol{\theta} = (\mu, \sigma^2)^\top \in \mathbb{R}_+ \times \mathbb{R}_+$. Es decir, cada X_i tiene densidad

$$f(z; \mu, \sigma^2) = \frac{1}{\sigma z \sqrt{2\pi}} \exp\Big\{ - \frac{(\log z - \mu)^2}{2\sigma^2} \Big\}, \quad z > 0.$$

En este caso

$$\mu_1 = \exp(\mu + \sigma^2/2), \qquad \mu_2 = \exp(\sigma^2)\big(\exp(\mu + \sigma^2/2)\big)^2,$$

y portanto los estimadores de momentos son dados por:

$$\widehat{\mu}_{\mathsf{MM}} = 2\log m_1 - \frac{1}{2}\log m_2, \qquad \widehat{\sigma}^2_{\mathsf{MM}} = \log m_2 - 2\log m_1.$$

OBSERVACIÓN. Una pregunta de interés es: ¿El estimador de momentos es **único**? Considere el siguiente ejemplo.

EJEMPLO 2.8. Suponga $X_1, \ldots, X_n$ una muestra aleatoria desde $\mathsf{Poi}(\lambda)$, $\lambda > 0$. Recuerde que

$$\mathsf{E}(X_1) = \mathsf{var}(X_1) = \lambda.$$

De este modo, podemos considerar

$$\widehat{\lambda}_{\mathsf{MM}} = \frac{1}{n}\sum_{i=1}^n x_i \quad \big(= h(m_1) \big)$$

por otro lado, otro estimador puede ser

$$\widetilde{\lambda}_{\mathsf{MM}} = h(m_1, m_2) = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2.$$

Sin embargo, ¿Cuál $\widehat{\lambda}_{\mathsf{MM}}$ o $\widetilde{\lambda}_{\mathsf{MM}}$ es mejor?[1]

En la práctica, tenemos que $\boldsymbol{\theta}$ es vector p-dimensional y usualmente estamos interesados en $\boldsymbol{\gamma} = \boldsymbol{\theta}$. De este modo tenemos que

$$\begin{aligned} \mu_1 &= g_1(\theta_1, \ldots, \theta_p), \\ &\vdots \\ \mu_p &= g_p(\theta_1, \ldots, \theta_p). \end{aligned}$$

Resolviendo para los p-parámetros en función de los momentos, obtenemos

$$\begin{aligned} \theta_1 &= h_1(\mu_1, \ldots, \mu_p), \\ &\vdots \\ \theta_p &= h_p(\mu_1, \ldots, \mu_p). \end{aligned} \tag{2.1}$$

[1] Más adelante estudiaremos como comparar entre dos estimadores.

Finalmente, el estimador $\widehat{\boldsymbol{\theta}}_{\mathsf{MM}}$, puede ser obtenido substituyendo en (2.1) por los momentos muestrales, es decir:

$$\begin{aligned}\widehat{\theta}_1 &= h_1(m_1,\ldots,m_p),\\ &\vdots\\ \widehat{\theta}_p &= h_p(m_1,\ldots,m_p).\end{aligned}$$

Debemos resaltar que el método de momentos requiere resolver el sistema de ecuaciones **no lineal**

$$\begin{aligned}g_1(\theta_1,\ldots,\theta_p)-\mu_1 &= 0,\\ &\vdots\\ g_p(\theta_1,\ldots,\theta_p)-\mu_p &= 0,\end{aligned}$$

que puede ser escrito como:

$$\boldsymbol{\Psi}(\boldsymbol{\theta})=\mathbf{0}, \tag{2.2}$$

donde $\boldsymbol{\Psi}:\mathbb{R}^p\to\mathbb{R}^p$ y por tanto $\widehat{\boldsymbol{\theta}}_{\mathsf{MM}}$ corresponde a una raíz de la **ecuación de estimación** en (2.2). Note que, en general, para obtener $\widehat{\boldsymbol{\theta}}_{\mathsf{MM}}$ se requiere de métodos iterativos, tal como:

$$\boldsymbol{\theta}^{(s+1)}=\boldsymbol{\theta}^{(s)}-\{\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}^{(s)})\}^{-1}\boldsymbol{\Psi}(\boldsymbol{\theta}^{(s)}),\qquad s=0,1,\ldots,$$

donde $\boldsymbol{\theta}^{(s)}$ representa una estimación para $\boldsymbol{\theta}$ en la etapa s-ésima y $\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta})=\partial\boldsymbol{\Psi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^\top$.

Observación. Una dificultad evidente del método de momentos es que $\boldsymbol{\Psi}(\boldsymbol{\theta})=\mathbf{0}$ puede tener múltiples raíces.

2.1.2. Estimación máximo verosímil. Este es uno de los procedimientos de estimación más ampliamente usados. Es motivado por el principio de verosimilitud y los estimadores obtenidos disfrutan de buenas propiedades.

Definición 2.9 (Estimador máximo verosimil). Un estimador $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$ es llamado estimador máximo verosímil (MLE) de $\boldsymbol{\theta}$, si

$$L(\widehat{\boldsymbol{\theta}}_{\mathsf{ML}};\boldsymbol{x})\geq L(\boldsymbol{\theta};\boldsymbol{x}),\qquad \forall\,\boldsymbol{\theta}\in\Theta.$$

Es decir, $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$ debe ser solución del siguiente problema de optimización

$$\max_{\theta\in\Theta} L(\boldsymbol{\theta};\boldsymbol{x})$$

o equivalentemente,

$$\max_{\theta\in\Theta} \ell(\boldsymbol{\theta};\boldsymbol{x}).$$

Además, en ocasiones escribimos

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}:=\operatorname*{argmax}_{\theta\in\Theta}\ell(\boldsymbol{\theta};\boldsymbol{x}).$$

Resultado 2.10 (Invarianza del MLE). *Si $\boldsymbol{\gamma}=\boldsymbol{g}(\boldsymbol{\theta})$ y $\boldsymbol{g}$ es biyectiva. Entonces $\widehat{\boldsymbol{\theta}}$ es el MLE para $\boldsymbol{\theta}$ si y solo si $\widehat{\boldsymbol{\gamma}}=\boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ es el MLE para $\boldsymbol{\gamma}$.*

DEMOSTRACIÓN. Considere $L(\boldsymbol{\theta}; \boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta})$ y como $\boldsymbol{g}$ es biyectiva tenemos que

$$\widetilde{L}(\boldsymbol{\gamma}; \boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{g}^{-1}(\boldsymbol{\gamma})).$$

Además,

$$\begin{aligned} \widetilde{L}(\widehat{\boldsymbol{\gamma}}; \boldsymbol{x}) \geq \widetilde{L}(\boldsymbol{\gamma}; \boldsymbol{x}), \quad \forall\, \boldsymbol{\gamma} \quad &\iff \quad f(\boldsymbol{x}; \boldsymbol{g}^{-1}(\widehat{\boldsymbol{\gamma}})) \geq f(\boldsymbol{x}; \boldsymbol{g}^{-1}(\boldsymbol{\gamma})), \quad \forall\, \boldsymbol{\gamma} \\ \iff f(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}) \geq f(\boldsymbol{x}; \boldsymbol{\theta}), \quad \forall\, \boldsymbol{\theta} \quad &\iff \quad L(\widehat{\boldsymbol{\theta}}; \boldsymbol{x}) \geq L(\boldsymbol{\theta}; \boldsymbol{x}), \quad \forall\, \boldsymbol{\theta}. \end{aligned}$$

□

Para el caso en que $\boldsymbol{g}$ no sea biyectiva, considere la siguiente definición.

DEFINICIÓN 2.11. Si $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$ es el MLE de $\boldsymbol{\theta}$ y $\boldsymbol{\gamma} = \boldsymbol{g}(\boldsymbol{\theta})$. Entonces el MLE de $\boldsymbol{\gamma}$ es definido como:

$$\widehat{\boldsymbol{\gamma}}_{\mathsf{ML}} = \boldsymbol{g}(\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}).$$

Si la función de log-verosimilitud $\ell(\boldsymbol{\theta})$ es continuamente diferenciable, el estimador máximo verosímil $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$ es dada como una solución de las ecuaciónes de verosimiltud

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

donde $\boldsymbol{U}(\boldsymbol{\theta}) = \partial \ell(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ corresponde al vector score. En particular, si las ecuaciones de verosimilitud son una ecuación lineal en los parámetros, el estimador máximo verosímil puede ser expresado explícitamente.

EJEMPLO 2.12 (distribución Binomial). Sea $x \in \mathcal{X} = \{0, 1, \ldots, n\}$ una realización desde $\mathsf{Bin}(n, \theta)$. El espacio paramétrico es $\Theta = (0, 1)$ y la función de verosimilitud adopta la forma

$$L(\theta; \boldsymbol{x}) = \binom{n}{x} \theta^x (1 - \theta)^{n-x},$$

mientras que la log-verosimilitud es dada por

$$\ell(\theta; \boldsymbol{x}) = \log \binom{n}{x} + x \log \theta + (n - x) \log(1 - \theta).$$

$\widehat{\theta}_{\mathsf{ML}}$ es solución de la ecuación

$$U(\theta; \boldsymbol{x}) = \dot{\ell}(\theta; \boldsymbol{x}) = \frac{x}{\theta} - \frac{n - x}{1 - \theta} = 0.$$

Si $x \neq 0$ y $x \neq n$ la solución existe, en cuyo caso tenemos

$$\widehat{\theta}_{\mathsf{ML}} = \frac{x}{n}.$$

EJEMPLO 2.13 (distribución Normal). Considere $X_1, \ldots, X_n$ muestra aleatoria desde $\mathsf{N}(\mu, \sigma^2)$. De este modo

$$L(\mu, \sigma^2; \boldsymbol{x}) = (2\pi\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big\}.$$

Lo que permite obtener la función de log-verosimilitud

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2,$$

diferenciando con respecto a μ y σ^2 lleva a las ecuaciones de verosimilitud

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0,$$
$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0,$$

resolviendo estas ecuaciones para μ y σ^2, sigue que

$$\widehat{\mu}_{\mathsf{ML}} = \overline{x}, \qquad \widehat{\sigma}^2_{\mathsf{ML}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\mu}_{\mathsf{ML}})^2.$$

EJEMPLO 2.14 (distribución Uniforme). Sea $X_1, \ldots, X_n$ una muestra de variables aleatorias IID desde $\mathsf{U}[0, \theta]$. Esto es

$$f(x; \theta) = \begin{cases} \dfrac{1}{\theta}, & 0 \leq x \leq \theta, \\ 0, & \text{en otro caso.} \end{cases} = \frac{1}{\theta} I_{[0,\theta]}(x), \qquad \theta > 0.$$

De este modo,[2]

$$L(\theta; \boldsymbol{x}) = \frac{1}{\theta} \prod_{i=1}^{n} I_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[x_i,\infty)}(\theta).$$

Tenemos que

$$\prod_{i=1}^{n} I_{[x_i,\infty)}(\theta) = 1 \iff I_{[x_i,\infty)}(\theta) = 1, \ \forall\, i \iff x_i \leq \theta, \ \forall\, i$$
$$\iff \max_i \{x_i\} \leq \theta \iff I_{[x_{(n)},\infty)}(\theta) = 1,$$

donde $x_{(n)} = \max\{x_1, \ldots, x_n\}$. De este modo, la función

$$L(\theta; \boldsymbol{x}) = \frac{1}{\theta^n} I_{[x_{(n)},\infty)}(\theta),$$

es monótona creciente, de ahí que

$$L(\theta; \boldsymbol{x}) \leq \frac{1}{{x_{(n)}}^n},$$

y por tanto sigue que $\widehat{\theta}_{\mathsf{ML}} = x_{(n)}$.

EJEMPLO 2.15 (distribución Laplace). Suponga $X_1, \ldots, X_n$ variables aleatorias IID con densidad

$$f(x; a, b) = \frac{b}{2} \exp\{-b|x - a|\}, \qquad x \in \mathbb{R},$$

con $a \in \mathbb{R}$ y $b > 0$. De este modo, para b conocido, tenemos

$$\ell(a; \boldsymbol{x}) = \log(b/2) - b \sum_{i=1}^{n} |x_i - a|.$$

Es decir, podemos obtener $\widehat{a}_{\mathsf{ML}}$, equivalentemente, como la solución de

$$\min_a \sum_{i=1}^{n} |x_i - a|,$$

[2]Basta notar que $0 \leq x_i \leq \theta \Rightarrow x_i \leq \theta < \infty$.

y es bien sabido que $\widehat{a}_{\mathsf{ML}} = \mathsf{mediana}\{x_1, \ldots, x_n\}$

En el siguiente ejemplo se presenta la estimación de parámetros mediante máxima verosimilitud para la clase de la familia exponencial. Por simplicidad solamente será considerado el caso de la FE 1-paramétrica.

EJEMPLO 2.16 (Familia Exponencial 1-paramétrica). Sea $X_1, \ldots, X_n$ variables aleatorias IID con distribución común en la FE 1-paramétrica y $\theta \in \Theta$. Considere $\phi = \eta(\theta)$ y sea $\gamma(\phi) = \gamma(\eta(\theta)) = b(\theta)$. Sabemos que la densidad conjunta es dada por

$$L(\phi) = \exp\Big[\phi \sum_{i=1}^{n} T(x_i) - n\gamma(\phi)\Big] \prod_{i=1}^{n} h(x_i).$$

De este modo, la función de log-verosimilitud adopta la forma:

$$\ell(\phi) = \phi \sum_{i=1}^{n} T(x_i) - n\gamma(\phi) + \sum_{i=1}^{n} \log h(x_i),$$

lo que lleva a

$$\frac{\mathsf{d}\ell(\phi)}{\mathsf{d}\phi} = \sum_{i=1}^{n} T(x_i) - n\gamma'(\phi).$$

Resolviendo la condición de primer orden $\mathsf{d}\ell(\phi)/\,\mathsf{d}\phi = 0$, tenemos que $\widehat{\phi}_{\mathsf{ML}}$ es solución de la ecuación:

$$\gamma'(\phi) = \frac{1}{n} \sum_{i=1}^{n} T(x_i).$$

Finalmente por la propiedad de invarianza del MLE sigue que $\widehat{\theta}_{\mathsf{ML}} = \eta^{-1}(\widehat{\phi}_{\mathsf{ML}})$. Por otro lado, es fácil notar que

$$\frac{\mathsf{d}^2\,\ell(\phi)}{\mathsf{d}\phi^2} = -n\gamma''(\phi) = -nb''(\theta) = -\,\mathsf{var}\Big(\sum_{i=1}^{n} T(x_i)\Big) \leq 0.$$

De lo anterior, sigue que $\ell(\phi)$ es cóncava y por tanto su máximo en $\Phi = \eta(\Theta)$ debe ser único.

EJEMPLO 2.17 (distribución Weibull). Suponga $X_1, \ldots, X_n$ muestra aleatoria con distribución Weibull, en cuyo caso,

$$f(x; \boldsymbol{\theta}) = \frac{\alpha}{\beta}\Big(\frac{x}{\beta}\Big)^{\alpha-1} \exp\Big\{-\Big(\frac{x}{\beta}\Big)^{\alpha}\Big\}, \qquad x > 0,$$

con $\boldsymbol{\theta} = (\alpha, \beta)^\top \in \mathbb{R}_+ \times \mathbb{R}_+$. La función de log-verosimilitud es dada por

$$\ell(\boldsymbol{\theta}) = n(\log\alpha - \log\beta) + (\alpha - 1)\sum_{i=1}^{n} \log\Big(\frac{x_i}{\beta}\Big) - \sum_{i=1}^{n}\Big(\frac{x_i}{\beta}\Big)^{\alpha}.$$

Diferenciando obtenemos las ecuaciones:

$$\frac{n}{\alpha} + \sum_{i=1}^{n} \log\Big(\frac{x_i}{\beta}\Big) - \sum_{i=1}^{n}\Big(\frac{x_i}{\beta}\Big)^{\alpha} \log\Big(\frac{x_i}{\beta}\Big) = 0$$

$$-\frac{n\alpha}{\beta} + \frac{\alpha}{\beta}\sum_{i=1}^{n}\Big(\frac{x_i}{\beta}\Big)^{\alpha} = 0,$$

que corresponde a un sistema de ecuaciones no lineales y por tanto métodos iterativos son necesarios.

Debemos resaltar que los procedimientos tipo-Newton son apropiados para resolver problemas *no restringidos*. Mientras que, en general, la estimación máximo verosímil corresponde a un problema de optimización restringida. En efecto, debemos tener que $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} \in \Theta$. Además, los estimadores que surgen de utilizar procedimientos de estimación restringida, suelen tener propiedades ligeramente más complejas que sus contrapartes no restringidas. El siguiente ejemplo, permite notar una forma de contornar esta dificultad mediante reparametrizar el modelo estadístico.

EJEMPLO 2.18 (distribución Poisson). Suponga $X_1, \ldots, X_n$ muestra aleatoria desde $\mathsf{Poi}(\lambda)$. En este caso,

$$\ell(\lambda; \boldsymbol{x}) = \sum_{i=1}^{n} (x_i \log \lambda - \lambda - \log x_i!), \qquad \lambda > 0,$$

despreciando aquellos términos que no dependen de λ, tenemos

$$\ell(\lambda; \boldsymbol{x}) = \sum_{i=1}^{n} (x_i \log \lambda - \lambda).$$

Considere $\phi = \log \lambda$, es decir $\lambda = e^{\phi}$ y note que $\phi \in \mathbb{R}$. Así,

$$\ell(\phi; \boldsymbol{x}) = \sum_{i=1}^{n} (x_i \phi - e^{\phi}).$$

Luego, estimamos ϕ y hacemos $\widehat{\lambda}_{\mathsf{ML}} = e^{\widehat{\phi}_{\mathsf{ML}}}$.

2.2. Propiedades de estimadores puntuales

Es frecuente contar con más de un estimador para un parámetro de interés. De este modo es requerido disponer de algún criterio que permita la comparación de diferentes estimadores. Considere las siguientes definiciones.

DEFINICIÓN 2.19 (Error Cuadrático Medio). Sea $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ un modelo estadístico para la variable X y sea $\boldsymbol{T}$ un estimador para $\gamma = g(\theta)$. El error cuadrático medio (MSE) de T es dado por

$$\mathsf{MSE}(T, \theta) = \mathsf{E}_\theta\{(T - g(\theta))^2\}.$$

OBSERVACIÓN. Es fácil notar que

$$\mathsf{MSE}(T, \theta) = \{\mathsf{E}_\theta(T) - g(\theta)\}^2 + \mathsf{var}_\theta(T).$$

DEFINICIÓN 2.20 (Sesgo). El sesgo de un estimador T es definido como:

$$\mathsf{bias}(T, \theta) = \mathsf{E}_\theta(T) - g(\theta).$$

De este modo, usando la definición anterior, tenemos que:

$$\mathsf{MSE}(T, \theta) = \{\mathsf{bias}(T, \theta)\}^2 + \mathsf{var}_\theta(T).$$

DEFINICIÓN 2.21 (Insesgamiento). Un estimador T para $\gamma = g(\theta)$ se dice insesgado, si

$$\mathsf{E}_\theta(T) = g(\theta), \qquad \forall\, \theta \in \Theta,$$

o equivalentemente,

$$\mathsf{bias}(T, \theta) = 0, \qquad \forall\, \theta \in \Theta.$$

Estimadores que "en promedio" están alejados de $g(\theta)$ son indeseables. Aunque en algunos casos es tolerable un sesgo pequeño. En ocasiones tenemos estimadores en que su sesgo tiende a cero conforme $n \to \infty$.

EJEMPLO 2.22. Sea $X_1, \ldots, X_n$ variables aleatorias IID con varianza finita. Suponga que $\gamma = \sigma^2$ es el parámetro de interés. Sabemos que el estimador MM es dado por:

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

Además,

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = \boldsymbol{X}^\top \Big(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\boldsymbol{X} = \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X},$$

donde $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Como $X_1, \ldots, X_n$ son IID considere

$$\mathsf{E}(\boldsymbol{X}) = \mu\mathbf{1}_n, \qquad \mathsf{Cov}(\boldsymbol{X}) = \sigma^2\boldsymbol{I}_n.$$

De este modo,[3]

$$\mathsf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right\} = \frac{1}{n}\,\mathsf{E}(\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}) = \frac{1}{n}\{\sigma^2 \operatorname{tr}\boldsymbol{C} + \mu^2\mathbf{1}^\top\boldsymbol{C}\mathbf{1}\}.$$

Como $\operatorname{tr}\boldsymbol{C} = \operatorname{tr}\boldsymbol{I} - \frac{1}{n}\operatorname{tr}\mathbf{1}\mathbf{1}^\top = n-1$ y $\boldsymbol{C}\mathbf{1} = \mathbf{0}$, sigue que

$$\mathsf{E}(\widehat{\sigma}^2_{\mathsf{MM}}) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \Big(\frac{n-1}{n}\Big)\sigma^2.$$

Es decir, $\widehat{\sigma}^2_{\mathsf{MM}}$ es un estimador sesgado, y

$$\mathsf{bias}(\widehat{\sigma}^2_{\mathsf{MM}}, \sigma^2) = \mathsf{E}(\widehat{\sigma}^2_{\mathsf{MM}}) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

Aunque $\lim_{n\to\infty}\mathsf{bias}(\widehat{\sigma}^2_{\mathsf{MM}}, \sigma^2) = 0$. El "factor de corrección" $\frac{n}{n-1}$, lleva al estimador

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

que es insesgado.

Suponga adicionalmente que $X_i \sim \mathsf{N}(\mu, \sigma^2)$, para $i = 1, \ldots, n$. Entonces,

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

y sabemos que

$$\mathsf{E}(U) = n-1, \qquad \text{y} \qquad \mathsf{var}(U) = 2(n-1).$$

Podemos escribir

$$U = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \quad \Longrightarrow \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{\sigma^2}{n}\,U,$$

luego, tenemos

$$\mathsf{var}(\widehat{\sigma}^2_{\mathsf{MM}}) = \frac{\sigma^4}{n^2}\,\mathsf{var}(U).$$

[3]Para $\boldsymbol{X}$ vector aleatorio con $\mathsf{E}(\boldsymbol{X}) = \boldsymbol{\theta}$ y $\mathsf{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, tenemos que $\mathsf{E}(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}) = \operatorname{tr}\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{\theta}^\top\boldsymbol{A}\boldsymbol{\theta}$.

De este modo,

$$\mathsf{MSE}(\widehat{\sigma}^2_{\mathsf{MM}}, \sigma^2) = \Big(- \frac{\sigma^2}{n} \Big)^2 + \frac{\sigma^4}{n^2} \, \mathsf{var}(U) = \frac{\sigma^4}{n^2} + \frac{\sigma^4}{n^2} 2(n-1) = \sigma^4 \Big(\frac{2n-1}{n^2} \Big).$$

Mientras que

$$U = \frac{(n-1)S^2}{\sigma^2} \quad \Longrightarrow \quad S^2 = \frac{\sigma^2}{n-1} \, U.$$

Así,

$$\mathsf{MSE}(S^2, \sigma^2) = 0 + \mathsf{var}(S^2),$$

es decir,

$$\mathsf{MSE}(S^2, \sigma^2) = 0 + \frac{\sigma^4}{(n-1)^2} \, \mathsf{var}(U) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}.$$

Finalmente,

$$\mathsf{MSE}(S^2, \sigma^2) > \mathsf{MSE}(\widehat{\sigma}^2_{\mathsf{MM}}, \sigma^2), \qquad n > 1.$$

Es decir, aunque $\widehat{\sigma}^2_{\mathsf{MM}}$ es un estimador sesgado, este es *mejor* usando el error cuadrático médio. Note además que, bajo normalidad, $\widehat{\sigma}^2_{\mathsf{MM}} = \widehat{\sigma}^2_{\mathsf{ML}}$.

Suponga $\Theta \subseteq \mathbb{R}^k$ y que el parámetro de interés $\boldsymbol{\gamma}$ es k-dimensional, esto es, $\boldsymbol{g} : \Theta \to \Gamma \subseteq \mathbb{R}^m$. Entonces, $\boldsymbol{T}$ se dice un estimador insesgado, si

$$\mathsf{E}(\boldsymbol{T}) = \boldsymbol{g}(\boldsymbol{\theta}), \qquad \forall \, \boldsymbol{\theta} \in \Theta.$$

La extensión del error cuadrático medio para el caso multiparamétrico adopta la forma

$$\begin{aligned} \mathsf{MSE}(\boldsymbol{T}, \boldsymbol{\theta}) &= \mathsf{E}_\theta \{ (\boldsymbol{T} - \boldsymbol{g}(\boldsymbol{\theta}))(\boldsymbol{T} - \boldsymbol{g}(\boldsymbol{\theta}))^\top \} \\ &= \mathsf{Cov}_\theta(\boldsymbol{T}) + \{ \mathsf{E}_\theta(\boldsymbol{T}) - \boldsymbol{g}(\boldsymbol{\theta}) \} \{ \mathsf{E}_\theta(\boldsymbol{T}) - \boldsymbol{g}(\boldsymbol{\theta}) \}^\top. \end{aligned}$$

DEFINICIÓN 2.23. Sean $\boldsymbol{T}$ y $\boldsymbol{T}_*$ dos estimadores para $\boldsymbol{\gamma}$. Decimos que $\boldsymbol{T}_*$ tiene error cuadrático medio más pequeño que $\boldsymbol{T}$ si

$$\boldsymbol{u}^\top \big(\mathsf{MSE}(\boldsymbol{T}_*, \boldsymbol{\theta}) - \mathsf{MSE}(\boldsymbol{T}, \boldsymbol{\theta}) \big) \boldsymbol{u} \leq 0, \qquad \forall \, \boldsymbol{u} \in \mathbb{R}^m,$$

y escribimos

$$\mathsf{MSE}(\boldsymbol{T}_*, \boldsymbol{\theta}) \leq \mathsf{MSE}(\boldsymbol{T}, \boldsymbol{\theta}).$$

En general, evaluar la condición dada por la definición anterior puede ser difícil. Esto ha motivado la introducción de algunos criterios más simples para comparar entre diferentes estimadores. En efecto, decimos que $\boldsymbol{T}_*$ es *T-óptimo*, si

$$\operatorname{tr} \mathsf{MSE}(\boldsymbol{T}_*, \boldsymbol{\theta}) \leq \operatorname{tr} \mathsf{MSE}(\boldsymbol{T}, \boldsymbol{\theta}), \tag{2.3}$$

mientras que $\boldsymbol{T}_*$ se dice *D-óptimo*, si satisface

$$\det \mathsf{MSE}(\boldsymbol{T}_*, \boldsymbol{\theta}) \leq \det \mathsf{MSE}(\boldsymbol{T}, \boldsymbol{\theta}). \tag{2.4}$$

El criterio dado en la Definición 2.23 también es conocido como *M-optimalidad*. Debemos destacar que en ocasiones el error cuadrático medio es definido como

$$\begin{aligned} \mathsf{E}_\theta[\|\boldsymbol{T} - \boldsymbol{g}(\boldsymbol{\theta})\|^2] &= \| \, \mathsf{E}_\theta(\boldsymbol{T}) - \boldsymbol{g}(\boldsymbol{\theta}) \|^2 + \sum_{j=1}^m \mathsf{var}(T_j) \\ &= \| \, \mathsf{bias}(\boldsymbol{T}, \boldsymbol{\theta}) \|^2 + \operatorname{tr} \mathsf{Cov}_\theta(\boldsymbol{T}), \end{aligned}$$

que corresponde al criterio de *T-optimalidad*. Lamentablemente, es muy poco frecuente encontrar un estimador que *siempre* (es decir, para todo $\boldsymbol{\theta} \in \Theta$) sea mejor.