Notas de Clase :

Elementos de Inferencia Estadística

Felipe Osorio

Departamento de Matemática, Universidad Técnica Federico Santa María

Índice general

Capítulo 1

Preliminares

1.1. Vectores Aleatorios

El propósito de esta sección es introducir algunas propiedades elementales de vectores aleatorios útiles a lo largo de este curso. Se asume que el lector es familiar con el concepto de variable aleatoria unidimensional.

Un vector aleatorio n-dimensional $\boldsymbol{X}$ es una función (medible) desde el espacio de probabilidad Ω a $\mathbb{R}^n$, esto es

$$\boldsymbol{X} : \Omega \to \mathbb{R}^n.$$

Por convención asumiremos que el vector aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ es un vector columna.

Definición 1.1 (Función de distribución). Para $\boldsymbol{X}$ distribuído en $\mathbb{R}^n$, la *función de distribución* de $\boldsymbol{X}$ es una función $F : \mathbb{R}^n \to [0, 1]$, tal que

$$F(\boldsymbol{x}) = \mathsf{P}(\boldsymbol{X} \leq \boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^n \tag{1.1}$$

y denotamos $\boldsymbol{X} \sim F$ o $\boldsymbol{X} \sim F_X$.

La función en (1.1) debe ser entendida como

$$F(\boldsymbol{x}) = \mathsf{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n),$$

que corresponde a la probabilidad del evento $\bigcap_{k=1}^{n} \{X_k \leq x_k\}$.

Propiedad 1.2. La función de distribución acumulada tiene las siguientes propiedades:

(a) $F(\boldsymbol{x})$ es función monótona creciente y contínua a la derecha en cada uno de los componentes de $\boldsymbol{X}$,
(b) $0 \leq F(\boldsymbol{x}) \leq 1$,
(c) $F(-\infty, x_2, \ldots, x_n) = \cdots = F(x_1, \ldots, x_{n-1}, -\infty) = 0$,
(d) $F(+\infty, \ldots, +\infty) = 1$.

Sea F la función de distribución del vector aleatorio $\boldsymbol{X}$. Entonces, existe una función no-negativa f tal que

$$F(\boldsymbol{x}) = \int_{-\infty}^{x} f(\boldsymbol{u}) \, \mathsf{d}\boldsymbol{u}, \qquad \boldsymbol{x} \in \mathbb{R}^n,$$

en este caso decimos que $\boldsymbol{X}$ es un vector aleatorio contínuo con *función de densidad* f. Por el teorema fundamental del Cálculo, tenemos que

$$f(\boldsymbol{x}) = \frac{\partial^n F(\boldsymbol{x})}{\partial x_1 \cdots \partial x_n}.$$

Además, considere $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, para $\boldsymbol{x}, \boldsymbol{y}$ vectores en $\bar{\mathbb{R}}^n$, entonces

$$\boldsymbol{x} \leq \boldsymbol{y} \quad \text{esto es,} \quad x_i \leq y_i, \quad \text{para } i = 1, \ldots, n.$$

Esto permite definir un rectángulo n-dimensional en $\mathbb{R}^n$ como

$$I = (\boldsymbol{a}, \boldsymbol{b}] = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a} < \boldsymbol{x} \leq \boldsymbol{b}\}$$

para todo $\boldsymbol{a}, \boldsymbol{b} \in \bar{\mathbb{R}}^n$. Entonces, también por el teorema fundamental del Cálculo, tenemos que si

$$f(\boldsymbol{x}) = \frac{\partial^n F(\boldsymbol{x})}{\partial x_1 \cdots \partial x_n}.$$

existe y es continua (casi en toda parte) sobre un rectángulo I, entonces

$$\mathsf{P}(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{x})\, \mathsf{d}\boldsymbol{x}, \quad \forall A \subset I.$$

Naturalmente la función de densidad debe satisfacer

$$\int_{\mathbb{R}^n} f(\boldsymbol{x})\, \mathsf{d}\boldsymbol{x} = 1.$$

Considere el vector aleatorio n-dimensional $\boldsymbol{X}$ particionado como $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$ donde $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$ son vectores $n_1 \times 1$ y $n_2 \times 1$, respectivamente, con $n = n_1 + n_2$. Tenemos que $\boldsymbol{X}_i \sim F_i$, $i = 1, 2$, de este modo $\boldsymbol{X}$ se denomina la *conjunta* de $\boldsymbol{X}_1, \boldsymbol{X}_2$ mientras que los $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$ son llamados *marginales* de $\boldsymbol{X}$.

Note que, las funciones de distribución marginal pueden ser recuperadas desde la distribución conjunta mediante

$$F_1(\boldsymbol{s}) = F(\boldsymbol{s}, +\boldsymbol{\infty}), \qquad F_2(\boldsymbol{t}) = F(+\boldsymbol{\infty}, \boldsymbol{t}), \qquad \forall \boldsymbol{s} \in \mathbb{R}^{n_1}, \boldsymbol{t} \in \mathbb{R}^{n_2}.$$

Cuando $\boldsymbol{X}$ es absolutamente contínua con función de densidad $f(\boldsymbol{x}) = f(\boldsymbol{x}_1, \boldsymbol{x}_2)$, entonces la función de densidad de $\boldsymbol{X}_i$ también es absolutamente contínua y puede ser obtenida como

$$f_1(\boldsymbol{s}) = \int_{\mathbb{R}^{n_2}} f(\boldsymbol{s}, \boldsymbol{u})\, \mathsf{d}\boldsymbol{u}, \quad f_2(\boldsymbol{t}) = \int_{\mathbb{R}^{n_1}} f(\boldsymbol{u}, \boldsymbol{t})\, \mathsf{d}\boldsymbol{u}, \quad \forall \boldsymbol{s} \in \mathbb{R}^{n_1}, \boldsymbol{t} \in \mathbb{R}^{n_2},$$

el resultado anterior es análogo para el caso de distribuciones discretas. Si $\boldsymbol{X}$ es absolutamente contínuo y $f_1(\boldsymbol{x}_1) > 0$, entonces la *densidad condicional* de $\boldsymbol{X}_2$ dado $\boldsymbol{X}_1 = \boldsymbol{x}_1$ es

$$f_{X_2|X_1=x_1}(\boldsymbol{u}) = \frac{f_X(\boldsymbol{x}_1, \boldsymbol{u})}{f_1(\boldsymbol{x}_1)},$$

con función de distribución de $\boldsymbol{X}_2$ condicional a $\boldsymbol{X}_1 = \boldsymbol{x}_1$ dada por

$$F_{X_2|X_1=x_1}(\boldsymbol{u}) = \int_{-\infty}^{u} f_{X_2|X_1=x_1}(\boldsymbol{t})\, \mathsf{d}\boldsymbol{t},$$

tenemos además que

$$f_{X_2|X_1=x_1}(\boldsymbol{u}) = \frac{f_X(\boldsymbol{x}_1, \boldsymbol{u})}{\int_{\mathbb{R}^{n_2}} f_X(\boldsymbol{x}_1, \boldsymbol{t})\, \mathsf{d}\boldsymbol{t}}.$$

1.2. Operadores de esperanza y covarianza

Considere $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ vector aleatorio n-dimensional con función de densidad f. Entonces la esperanza de cualquier función $\boldsymbol{g}$ de $\boldsymbol{X}$ está dada por

$$\mathsf{E}(\boldsymbol{g}(\boldsymbol{X})) = \int_{\mathbb{R}^n} \boldsymbol{g}(\boldsymbol{t}) f(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t},$$

siempre que la integral (n-dimensional) exista.

Más generalmente, sea $\boldsymbol{Z} = (Z_{ij})$ una función matricial $m \times n$, entonces podemos definir el operador de esperanza de una matriz aleatoria como

$$\mathsf{E}(\boldsymbol{Z}(\boldsymbol{X})) = (\mathsf{E}(Z_{ij})), \qquad Z_{ij} = Z_{ij}(\boldsymbol{X}). \tag{1.2}$$

De la definición en (1.2) se desprenden una serie de resultados útiles con relación al operador de esperanza. Por ejemplo, sea $\boldsymbol{A} = (a_{ij})$ una matriz de constantes, entonces

$$\mathsf{E}(\boldsymbol{A}) = \boldsymbol{A}.$$

Resultado 1.3. *Sea $\boldsymbol{A} = (a_{ij})$, $\boldsymbol{B} = (b_{ij})$ y $\boldsymbol{C} = (c_{ij})$ matrices de constantes $l \times m$, $n \times p$ y $l \times p$, respectivamente. Entonces*

$$\mathsf{E}(\boldsymbol{AZB} + \boldsymbol{C}) = \boldsymbol{A}\,\mathsf{E}(\boldsymbol{Z})\boldsymbol{B} + \boldsymbol{C}.$$

Demostración. Sea $\boldsymbol{Y} = \boldsymbol{AZB} + \boldsymbol{C}$, entonces

$$Y_{ij} = \sum_{r=1}^{m} \sum_{s=1}^{n} a_{ir} Z_{rs} b_{sj} + c_{ij},$$

de este modo

$$\begin{aligned} \mathsf{E}(\boldsymbol{AZB} + \boldsymbol{C}) = (\mathsf{E}(Y_{ij})) &= \left(\sum_{r=1}^{m} \sum_{s=1}^{n} a_{ir}\, \mathsf{E}(Z_{rs}) b_{sj} + c_{ij} \right) \\ &= \boldsymbol{A}\,\mathsf{E}(\boldsymbol{Z})\boldsymbol{B} + \boldsymbol{C}. \end{aligned}$$

□

Un caso particular importante corresponde a la esperanza de una transformación lineal. Considere el vector aleatorio n-dimensional, $\boldsymbol{Y} = \boldsymbol{AX}$, donde $\boldsymbol{X}$ es vector aleatorio $m \times 1$, entonces $\mathsf{E}(\boldsymbol{AX}) = \boldsymbol{A}\,\mathsf{E}(\boldsymbol{X})$. Esta propiedad puede ser extendida para sumas de vectores aleatorios, como

$$\mathsf{E}\left(\sum_i \boldsymbol{A}_i \boldsymbol{X}_i \right) = \sum_i \boldsymbol{A}_i\, \mathsf{E}(\boldsymbol{X}_i),$$

de manera similar tenemos que

$$\mathsf{E}\left(\sum_i \alpha_i \boldsymbol{Z}_i \right) = \sum_i \alpha_i\, \mathsf{E}(\boldsymbol{Z}_i),$$

donde α_i son constantes y los $\boldsymbol{Z}_i$ son matrices aleatorias.

Definición 1.4 (Matriz de covarianza). Sean $\boldsymbol{X}$ e $\boldsymbol{Y}$ vectores aleatorios m y n-dimensionales, respectivamente. Se define la *matriz de covarianza* entre $\boldsymbol{X}$ e $\boldsymbol{Y}$ como la matriz $m \times n$,

$$\mathsf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = (\mathsf{Cov}(X_i, Y_j)).$$

Podemos apreciar, a partir de la definición de covarianza que

$$\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y}) = \mathsf{E}\{(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathsf{E}(\boldsymbol{Y}))^\top\}.$$

En efecto, sean $\boldsymbol{\mu} = \mathsf{E}(\boldsymbol{X})$ y $\boldsymbol{\eta} = \mathsf{E}(\boldsymbol{Y})$. Entonces,

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y}) &= (\mathsf{Cov}(X_i,Y_j)) = (\mathsf{E}(X_i-\mu_i)(Y_j-\eta_j)) \\ &= \mathsf{E}([(X_i-\mu_i)(Y_j-\eta_j)]) = \mathsf{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\eta})^\top].\end{aligned}$$

Tenemos ademas el siguiente resultado

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y}) &= \mathsf{E}\{(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathsf{E}(\boldsymbol{Y}))^\top\} \\ &= \mathsf{E}(\boldsymbol{X}\boldsymbol{Y}^\top - \mathsf{E}(\boldsymbol{X})\boldsymbol{Y}^\top - \boldsymbol{X}\,\mathsf{E}^\top(\boldsymbol{Y}) + \mathsf{E}(\boldsymbol{X})\,\mathsf{E}^\top(\boldsymbol{Y})) \\ &= \mathsf{E}(\boldsymbol{X}\boldsymbol{Y}^\top) - \mathsf{E}(\boldsymbol{X})\,\mathsf{E}^\top(\boldsymbol{Y}).\end{aligned}$$

Se define la *matriz de dispersión (varianza)*, como $\mathsf{Cov}(\boldsymbol{X}) = \mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})$. De este modo, tenemos

$$\mathsf{Cov}(\boldsymbol{X}) = (\mathsf{Cov}(X_i,X_j)) = \mathsf{E}\{(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))^\top\},$$

y, de la misma manera que para el caso de la matriz de covarianza,

$$\mathsf{Cov}(\boldsymbol{X}) = \mathsf{E}(\boldsymbol{X}\boldsymbol{X}^\top) - \mathsf{E}(\boldsymbol{X})\,\mathsf{E}^\top(\boldsymbol{X}).$$

EJEMPLO 1.5. Sea $\boldsymbol{a}$ vector de constantes $n \times 1$, entonces

$$\mathsf{Cov}(\boldsymbol{X}-\boldsymbol{a}) = \mathsf{Cov}(\boldsymbol{X}).$$

En efecto, note que

$$\boldsymbol{X}-\boldsymbol{a}-\mathsf{E}(\boldsymbol{X}-\boldsymbol{a}) = \boldsymbol{X}-\mathsf{E}(\boldsymbol{X}),$$

por tanto, tenemos

$$\mathsf{Cov}(\boldsymbol{X}-\boldsymbol{a},\boldsymbol{X}-\boldsymbol{a}) = \mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})$$

RESULTADO 1.6. *Si $\boldsymbol{X}$ e $\boldsymbol{Y}$ son vectores aleatorios m y n-dimensionales, respectivamente y $\boldsymbol{A} \in \mathbb{R}^{l\times m}$, $\boldsymbol{B} \in \mathbb{R}^{p\times n}$, entonces*

$$\mathsf{Cov}(\boldsymbol{A}\boldsymbol{X},\boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}^\top.$$

DEMOSTRACIÓN. Sean $\boldsymbol{U} = \boldsymbol{A}\boldsymbol{X}$ y $\boldsymbol{V} = \boldsymbol{B}\boldsymbol{Y}$, entonces

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{A}\boldsymbol{X},\boldsymbol{B}\boldsymbol{Y}) &= \mathsf{Cov}(\boldsymbol{U},\boldsymbol{V}) = \mathsf{E}\{(\boldsymbol{U}-\mathsf{E}(\boldsymbol{U}))(\boldsymbol{V}-\mathsf{E}(\boldsymbol{V}))^\top\} \\ &= \mathsf{E}\{(\boldsymbol{A}\boldsymbol{X}-\boldsymbol{A}\,\mathsf{E}(\boldsymbol{X}))(\boldsymbol{B}\boldsymbol{Y}-\boldsymbol{B}\,\mathsf{E}(\boldsymbol{Y}))^\top\} \\ &= \mathsf{E}\{\boldsymbol{A}(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathsf{E}(\boldsymbol{Y}))^\top\boldsymbol{B}^\top\} \\ &= \boldsymbol{A}\,\mathsf{E}\{(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathsf{E}(\boldsymbol{Y}))^\top\}\boldsymbol{B}^\top \\ &= \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}^\top.\end{aligned}$$

□

Tenemos el siguiente caso particular,

$$\mathsf{Cov}(\boldsymbol{A}\boldsymbol{X}) = \mathsf{Cov}(\boldsymbol{A}\boldsymbol{X},\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{A}^\top = \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{A}^\top.$$

EJEMPLO 1.7. Considere $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{U}$ y $\boldsymbol{V}$ vectores aleatorios n-dimensionales y $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ y $\boldsymbol{D}$ matrices de órdenes apropiados, entonces

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{AX}+\boldsymbol{BY},\boldsymbol{CU}+\boldsymbol{DV}) &= \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{U})\boldsymbol{C}^\top + \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{V})\boldsymbol{D}^\top \\ &\quad + \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y},\boldsymbol{U})\boldsymbol{C}^\top + \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y},\boldsymbol{V})\boldsymbol{D}^\top.\end{aligned}$$

tomando $\boldsymbol{U}=\boldsymbol{X}$, $\boldsymbol{V}=\boldsymbol{Y}$, $\boldsymbol{C}=\boldsymbol{A}$ y $\boldsymbol{D}=\boldsymbol{B}$, tenemos

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{AX}+\boldsymbol{BY}) &= \mathsf{Cov}(\boldsymbol{AX}+\boldsymbol{BY},\boldsymbol{AX}+\boldsymbol{BY}) \\ &= \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{A}^\top + \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}^\top \\ &\quad + \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y},\boldsymbol{X})\boldsymbol{A}^\top + \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y})\boldsymbol{B}^\top.\end{aligned}$$

RESULTADO 1.8. *Toda matriz de dispersión es simétrica y semidefinida positiva*

DEMOSTRACIÓN. La simetría de la matriz de dispersión es obvia. Para mostrar que $\mathsf{Cov}(\boldsymbol{X})$ es semidefinida positiva, sea $\boldsymbol{Z}=\boldsymbol{X}-\mathsf{E}(\boldsymbol{X})$, y considere la variable aleatoria $Y=\boldsymbol{a}^\top\boldsymbol{Z}$, para $\boldsymbol{a}\in\mathbb{R}^n$ un vector arbitrareo. Entonces,

$$\begin{aligned}\boldsymbol{a}^\top\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{a} &= \boldsymbol{a}^\top\,\mathsf{E}(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))^\top\boldsymbol{a} \\ &= \mathsf{E}(\boldsymbol{a}^\top(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))^\top\boldsymbol{a}) \\ &= \mathsf{E}(\boldsymbol{a}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{a}) = \mathsf{E}(Y^2)\geq 0\end{aligned}$$

y por tanto, $\mathsf{Cov}(\boldsymbol{X})$ es semidefinida positiva.

Ahora, suponga que $\mathsf{Cov}(\boldsymbol{X})$ es semidefinida positiva de rango r ($r\leq n$). Luego $\mathsf{Cov}(\boldsymbol{X})=\boldsymbol{B}\boldsymbol{B}^\top$ donde $\boldsymbol{B}\in\mathbb{R}^{n\times r}$ de rango r. Sea $\boldsymbol{Y}$ vector aleatorio r-dimensional con $\mathsf{E}(\boldsymbol{Y})=\boldsymbol{0}$ y $\mathsf{Cov}(\boldsymbol{Y})=\boldsymbol{I}$. Haciendo $\boldsymbol{X}=\boldsymbol{BY}$, sigue que $\mathsf{E}(\boldsymbol{X})=\boldsymbol{0}$ y

$$\mathsf{Cov}(\boldsymbol{X})=\mathsf{Cov}(\boldsymbol{BY})=\boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y})\boldsymbol{B}^\top=\boldsymbol{B}\boldsymbol{B}^\top.$$

Es decir, corresponde a una matriz de covarianza. □

RESULTADO 1.9. *Sea $\boldsymbol{X}$ vector aleatorio n-dimensional y considere la transformación lineal $\boldsymbol{Y}=\boldsymbol{AX}+\boldsymbol{b}$, donde $\boldsymbol{A}$ es una matriz de constantes $m\times n$ y $\boldsymbol{b}$ es vector de constantes $m\times 1$. Entonces*

$$\mathsf{E}(\boldsymbol{Y})=\boldsymbol{A}\,\mathsf{E}(\boldsymbol{X})+\boldsymbol{b},\qquad \mathsf{Cov}(\boldsymbol{Y})=\boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{A}^\top.$$

EJEMPLO 1.10. Sea $\boldsymbol{X}$ vector aleatorio n-dimensional con media $\mathsf{E}(\boldsymbol{X})=\boldsymbol{\mu}$ y matriz de dispersión $\mathsf{Cov}(\boldsymbol{X})=\boldsymbol{\Sigma}$. Sea

$$\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$$

la descomposición espectral de $\boldsymbol{\Sigma}$, donde $\boldsymbol{U}$ es matriz ortogonal y $\boldsymbol{\Lambda}=\mathrm{diag}(\boldsymbol{\lambda})$, y considere la siguiente transformación

$$\boldsymbol{Z}=\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top(\boldsymbol{X}-\boldsymbol{\mu})$$

de este modo, obtenemos que

$$\mathsf{E}(\boldsymbol{Z})=\boldsymbol{0}\qquad\text{y}\qquad\mathsf{Cov}(\boldsymbol{Z})=\boldsymbol{I}.$$

En efecto, la transformación $\boldsymbol{Z}=\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})$ también satisface que $\mathsf{E}(\boldsymbol{Z})=\boldsymbol{0}$ y $\mathsf{Cov}(\boldsymbol{Z})=\boldsymbol{I}$.

Suponga que $\boldsymbol{Z}$ es una matriz aleatoria $n \times p$ cuyas filas son vectores aleatorios independientes $p \times 1$, cada uno con la misma matriz de covarianza $\boldsymbol{\Sigma}$. Considere la partición

$$\boldsymbol{Z}^\top = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n),$$

donde $\mathsf{Cov}(\boldsymbol{Z}_i) = \boldsymbol{\Sigma}$, para $i = 1, \ldots, n$. Tenemos que

$$\operatorname{vec}(\boldsymbol{Z}^\top) = \begin{pmatrix} \boldsymbol{Z}_1 \\ \vdots \\ \boldsymbol{Z}_n \end{pmatrix},$$

y dado que todos los $\boldsymbol{Z}_i$ son independientes con la misma matriz de covarianza, podemos escribir

$$\mathsf{Cov}(\operatorname{vec}(\boldsymbol{Z}^\top)) = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma} \end{pmatrix} = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}.$$

Ahora suponga que llevamos a cabo la transformación lineal $\boldsymbol{Y} = \boldsymbol{AZB}$, donde $\boldsymbol{A} \in \mathbb{R}^{r \times n}$, $\boldsymbol{B} \in \mathbb{R}^{p \times q}$ son matrices de constantes. Entonces $\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{A}\,\mathsf{E}(\boldsymbol{Z})\boldsymbol{B}$, mientras que

$$\operatorname{vec}(\boldsymbol{Y}^\top) = (\boldsymbol{A} \otimes \boldsymbol{B}^\top) \operatorname{vec}(\boldsymbol{Z}^\top),$$

de modo que

$$\mathsf{E}(\operatorname{vec}(\boldsymbol{Y}^\top)) = (\boldsymbol{A} \otimes \boldsymbol{B}^\top)\, \mathsf{E}(\operatorname{vec}(\boldsymbol{Z})^\top).$$

Lo que lleva a calcular fácilmente la matriz de covarianza

$$\begin{aligned} \mathsf{Cov}(\operatorname{vec}(\boldsymbol{Y}^\top)) &= (\boldsymbol{A} \otimes \boldsymbol{B}^\top)\, \mathsf{Cov}(\operatorname{vec}(\boldsymbol{Z}^\top))(\boldsymbol{A} \otimes \boldsymbol{B}^\top)^\top \\ &= (\boldsymbol{A} \otimes \boldsymbol{B}^\top)(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma})(\boldsymbol{A}^\top \otimes \boldsymbol{B}) \\ &= \boldsymbol{AA}^\top \otimes \boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}. \end{aligned}$$

Definición 1.11 (Matriz de correlación). Sea $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ vector aleatorio con media $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$. Se define la matriz de correlaciones como $\boldsymbol{R} = (\rho_{ij})$, donde

$$\rho_{ij} = \frac{\mathsf{Cov}(X_i, X_j)}{\{\mathsf{var}(X_i)\, \mathsf{var}(X_j)\}^{1/2}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \qquad i, j = 1, \ldots, p.$$

Note que, para $\boldsymbol{\Sigma}$ matriz de covarianza del vector aleatorio $\boldsymbol{X}$ y con $\boldsymbol{D} = \operatorname{diag}(\sigma_{11}, \ldots, \sigma_{pp})$ $(= \operatorname{diag}(\boldsymbol{\Sigma}))$ podemos escribir

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{D}^{-1/2}.$$

Cada elemento de la diagonal de $\boldsymbol{R}$ es igual a 1, mientras que sus elementos fuera de la diagonal están entre -1 y 1. Además se desprende desde la definición que $\boldsymbol{R}$ es una matriz semidefinida positiva.

Resultado 1.12. *Sea $\boldsymbol{X}$ vector aleatorio p-dimensional con $\mathsf{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ y $\mathsf{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Sea $\boldsymbol{A}$ una matriz $p \times p$. Entonces*

$$\mathsf{E}(\boldsymbol{X}^\top \boldsymbol{AX}) = \operatorname{tr}(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{A\mu}.$$

DEMOSTRACIÓN. Tenemos

$$\begin{aligned}\mathsf{E}(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}) &= \mathsf{E}(\operatorname{tr} \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}) = \mathsf{E}(\operatorname{tr} \boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^\top)\\ &= \operatorname{tr} \mathsf{E}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^\top) = \operatorname{tr} \boldsymbol{A}\, \mathsf{E}(\boldsymbol{X}\boldsymbol{X}^\top)\\ &= \operatorname{tr} \boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}.\end{aligned}$$

□

Considere el siguiente caso especial: sea $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{a}$, entonces $\mathsf{Cov}(\boldsymbol{Y}) = \mathsf{Cov}(\boldsymbol{X})$ y tenemos

$$\mathsf{E}[(\boldsymbol{X} - \boldsymbol{a})^\top \boldsymbol{A}(\boldsymbol{X} - \boldsymbol{a})] = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \boldsymbol{a})^\top \boldsymbol{A}(\boldsymbol{\mu} - \boldsymbol{a}).$$

EJEMPLO 1.13. Sea $\mathbf{1}_n = (1, \ldots, 1)^\top$ vector n-dimensional cuyos componentes son todos 1. Note que, $\mathbf{1}_n^\top \mathbf{1}_n = n$. Considere el vector aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, entonces

$$\boldsymbol{X}^\top \boldsymbol{X} = \sum_{i=1}^n X_i^2, \qquad \mathbf{1}^\top \boldsymbol{X} = \sum_{i=1}^n X_i.$$

De este modo, tenemos

$$\begin{aligned}\sum_{i=1}^n (X_i - \overline{X})^2 &= \sum_{i=1}^n X_i^2 - n\overline{X}^2 = \boldsymbol{X}^\top \boldsymbol{X} - n\Big(\frac{1}{n}\mathbf{1}^\top \boldsymbol{X}\Big)^2\\ &= \boldsymbol{X}^\top \boldsymbol{X} - n\Big(\frac{1}{n}\mathbf{1}^\top \boldsymbol{X}\Big)\Big(\frac{1}{n}\mathbf{1}^\top \boldsymbol{X}\Big) = \boldsymbol{X}^\top \boldsymbol{X} - \frac{1}{n}\boldsymbol{X}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{X}\\ &= \boldsymbol{X}^\top \Big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}_n\Big)\boldsymbol{X}, \qquad \boldsymbol{J}_n = \mathbf{1}_n \mathbf{1}_n^\top\end{aligned}$$

Llamaremos a $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{J}_n$ la *matriz de centrado*. Suponga que $X_1, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuídas con media μ y varianza σ^2. Sigue que,

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

pues $\mathsf{Cov}(X_i, X_j) = 0$ $(i \neq j)$. Por tanto, podemos usar el Resultado (1.12) para calcular la esperanza de la variable aleatoria,

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

obteniendo

$$\mathsf{E}(Q) = \sigma^2 \operatorname{tr}(\boldsymbol{C}) + \mu^2 \mathbf{1}^\top \boldsymbol{C}\mathbf{1}.$$

Es fácil verificar que

$$\begin{aligned}\operatorname{tr}(\boldsymbol{C}) &= \operatorname{tr}\Big(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\Big) = \operatorname{tr}(\boldsymbol{I}) - \frac{1}{n}\operatorname{tr}(\mathbf{1}\mathbf{1}^\top) = n - \frac{1}{n}\mathbf{1}^\top \mathbf{1} = n - 1,\\ \boldsymbol{C}\mathbf{1} &= \Big(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\Big)\mathbf{1} = \mathbf{1} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0},\end{aligned}$$

de donde sigue que $\mathsf{E}(Q) = \sigma^2(n-1)$.

RESULTADO 1.14. *Si $\boldsymbol{X}$ es vector aleatorio $n \times 1$. Entonces su distribución está determinada por las distribuciones de las funciones lineales $\boldsymbol{a}^\top \boldsymbol{X}$, para todo $\boldsymbol{a} \in \mathbb{R}^n$.*

DEMOSTRACIÓN. La función característica de $\boldsymbol{a}^\top \boldsymbol{X}$ es

$$\varphi_{a^\top X}(t) = \mathsf{E}\{\exp(it\boldsymbol{a}^\top \boldsymbol{X})\},$$

de modo que

$$\varphi_{a^\top X}(1) = \mathsf{E}\{\exp(i\boldsymbol{a}^\top \boldsymbol{X})\} \quad (= \varphi_X(\boldsymbol{a})).$$

Es considerada como una función de $\boldsymbol{a}$, esto es, la función característica (conjunta) de $\boldsymbol{X}$. El resultado sigue notando que una distribución en $\mathbb{R}^n$ está completamente determinada por su función característica. □

La función característica permite un método bastante operativo para el cálculo del k-ésimo momento de un vector aleatorio $\boldsymbol{X}$. En efecto,

$$\begin{aligned}\boldsymbol{\mu}_k(\boldsymbol{X}) &= \begin{cases} \mathsf{E}(\boldsymbol{X} \otimes \boldsymbol{X}^\top \otimes \cdots \otimes \boldsymbol{X}^\top), & k \text{ par}, \\ \mathsf{E}(\boldsymbol{X} \otimes \boldsymbol{X}^\top \otimes \cdots \otimes \boldsymbol{X}^\top \otimes \boldsymbol{X}), & k \text{ impar}, \end{cases} \\ &= \begin{cases} i^{-k} \dfrac{\partial^k \varphi(\boldsymbol{t})}{\partial \boldsymbol{t} \partial \boldsymbol{t}^\top \cdots \partial \boldsymbol{t}^\top}\Big|_{t=0}, & k \text{ par}, \\ i^{-k} \dfrac{\partial^k \varphi(\boldsymbol{t})}{\partial \boldsymbol{t} \partial \boldsymbol{t}^\top \cdots \partial \boldsymbol{t}^\top \partial \boldsymbol{t}}\Big|_{t=0}, & k \text{ impar}. \end{cases}\end{aligned}$$

1.3. Independencia de vectores aleatorios

Sea $\boldsymbol{Z} = (\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$ con $\boldsymbol{X}$, $\boldsymbol{Y}$ vectores aleatorios n y q-dimensionales, respectivamente. Se dicen independientes si y sólo si

$$F(\boldsymbol{x}, \boldsymbol{y}) = G(\boldsymbol{x})H(\boldsymbol{y}),$$

donde $F(\boldsymbol{z})$, $G(\boldsymbol{x})$ y $H(\boldsymbol{y})$ son las funciones de distribución de $\boldsymbol{Z}$, $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente.

Si $\boldsymbol{Z}$, $\boldsymbol{X}$ e $\boldsymbol{Y}$ tienen densidades $f(\boldsymbol{z})$, $g(\boldsymbol{x})$ y $h(\boldsymbol{y})$, respectivamente. Entonces $\boldsymbol{X}$ e $\boldsymbol{Y}$ son independientes si

$$f(\boldsymbol{z}) = g(\boldsymbol{x})h(\boldsymbol{y}).$$

En cuyo caso, obtenemos como resultado

$$f(\boldsymbol{x}|\boldsymbol{y}) = g(\boldsymbol{x}).$$

RESULTADO 1.15. *Sean $\boldsymbol{X}$ e $\boldsymbol{Y}$ dos vectores aleatorios independientes. Entonces para funciones cualquiera κ y τ, tenemos*

$$\mathsf{E}\{\kappa(\boldsymbol{X})\tau(\boldsymbol{Y})\} = \mathsf{E}\{\kappa(\boldsymbol{X})\}\,\mathsf{E}\{\tau(\boldsymbol{Y})\},$$

si las esperanzas existen.

DEMOSTRACIÓN. En efecto, es fácil notar que

$$\begin{aligned}\mathsf{E}\{\kappa(\boldsymbol{X})\tau(\boldsymbol{Y})\} &= \int\int \kappa(\boldsymbol{x})\tau(\boldsymbol{y})g(\boldsymbol{x})h(\boldsymbol{y})\,\mathsf{d}\boldsymbol{x}\,\mathsf{d}\boldsymbol{y} \\ &= \Big(\int \kappa(\boldsymbol{x})g(\boldsymbol{x})\,\mathsf{d}\boldsymbol{x}\Big)\Big(\int \tau(\boldsymbol{y})h(\boldsymbol{y})\,\mathsf{d}\boldsymbol{y}\Big) \\ &= \mathsf{E}\{\kappa(\boldsymbol{X})\}\,\mathsf{E}\{\tau(\boldsymbol{Y})\}.\end{aligned}$$

□

1.4. Cambios de variable

Considere la función $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$, el *Jacobiano* se define como el valor absoluto del determinante de $\mathsf{D}\boldsymbol{f}(\boldsymbol{x})$ y es denotado por

$$J(\boldsymbol{y} \to \boldsymbol{x}) = |\mathsf{D}\boldsymbol{f}(\boldsymbol{x})|_+ = \operatorname{abs}(\det(\mathsf{D}\boldsymbol{f}(\boldsymbol{x}))),$$

donde $\boldsymbol{y} = \boldsymbol{f}(x)$. Note que si $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{y})$ y $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$, entonces tenemos

$$\begin{aligned} J(\boldsymbol{z} \to \boldsymbol{x}) &= J(\boldsymbol{z} \to \boldsymbol{y}) \cdot J(\boldsymbol{y} \to \boldsymbol{x}) \\ J(\boldsymbol{y} \to \boldsymbol{x}) &= \{J(\boldsymbol{x} \to \boldsymbol{y})\}^{-1} \end{aligned}$$

El siguiente resultado presenta una de aplicación del Jacobiano de una transformación para obtener la función de densidad de una transformación de un vector aleatorio.

Proposición 1.16 (Transformación de vectores aleatorios continuos). *Sea $\boldsymbol{X}$ vector aleatorio n-dimensional con densidad $f_X(\boldsymbol{x})$ y soporte $S = \{\boldsymbol{x} : f_X(\boldsymbol{x}) > 0\}$. Para $\boldsymbol{g} : S \to \mathbb{R}^n$ diferenciable e invertible, sea $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$. Entonces la densidad de $\boldsymbol{Y}$ está dada por*

$$\begin{aligned} f_Y(\boldsymbol{y}) &= |\mathsf{D}\boldsymbol{g}^{-1}(\boldsymbol{y})|_+ \, f_X(\boldsymbol{g}^{-1}(\boldsymbol{y})) \\ &= \{J(\boldsymbol{y} \to \boldsymbol{x})\}^{-1} f_X(\boldsymbol{g}^{-1}(\boldsymbol{y})). \end{aligned}$$

Ejemplo 1.17. Sea $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{B}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$, $\boldsymbol{X} \in \mathbb{R}^{n\times q}$, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ y $\boldsymbol{B} \in \mathbb{R}^{q\times q}$. Entonces

$$\mathsf{d}\,\boldsymbol{Y} = \boldsymbol{A}(\mathsf{d}\,\boldsymbol{X})\boldsymbol{B},$$

vectorizando obtenemos

$$\operatorname{vec}\mathsf{d}\,\boldsymbol{Y} = (\boldsymbol{B}^\top \otimes \boldsymbol{A}) \operatorname{vec}\mathsf{d}\,\boldsymbol{X},$$

esto es, $\mathsf{D}\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{B}^\top \otimes \boldsymbol{A}$, por tanto

$$J(\boldsymbol{Y} \to \boldsymbol{X}) = |\boldsymbol{B}^\top \otimes \boldsymbol{A}|_+ = |\boldsymbol{A}|_+^q |\boldsymbol{B}^\top|_+^n = |\boldsymbol{A}|_+^q |\boldsymbol{B}|_+^n$$

1.5. Modelo estadístico

El punto de partida es considerar que los datos observados $\boldsymbol{y}$ son una realización de una variable aleatoria Y cuya distribución es parcialmente conocida. Asumiremos que la distribución de la variable de interés es un miembro de una clase de medidas de probabilidad definida como

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

que es indexada por $\boldsymbol{\theta} \in \Theta$. El conjunto Θ es denominado *espacio paramétrico*.

Idea. Se desea conocer algunas propiedades de $\mathcal{P}$ basado en una (única) muestra aleatoria

Definición 1.18. Un *modelo estadístico* es un par $(\mathcal{Y}, \mathcal{P})$, donde $\mathcal{Y}$ es el conjunto de todos los resultados posibles, conocido como *espacio muestral*.

Definición 1.19. Una *muestra aleatoria*[1] $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ es una colección de variables aleatorias independientes, donde cada Y_i es distribuída como P_θ. En este contexto, n corresponde al número de variables aleatorias y es llamado *tamaño muestral*.

[1] Usualmente utilizaremos la notación $\mathsf{m.a.}(n)$.

En la práctica, se dispondrá de observaciones que son realizaciones de variables aleatorias independientes $Y_1, \ldots, Y_n$. En este caso $\mathcal{P}$ es una familia de medidas producto

$$\mathsf{P}_\theta = \mathsf{P}_{1,\theta} \otimes \cdots \otimes \mathsf{P}_{n,\theta}\,.$$

Una muestra aleatoria IID, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, es una colección de variables aleatorias independientes, donde cada Y_i tiene la misma distribución. Es decir, $\mathsf{P}_{i,\theta} = \mathsf{P}_{1,\theta}$ para todo i.

Evidentemente, cada Y_i tendrá función de distribución acumulada *común*, digamos F. Si F es *conocido*, podemos usar cálculo de probabilidades para deducir y estudiar sus propiedades. Sin embargo, en la práctica, F es *desconocido*, y el objetivo es tratar de inferir sus propiedades *desde los datos*. Frecuentemente, el interés es una *función no aleatoria* de F, tal como la media o su q-ésimo cuantil,

$$\mathsf{E}(Y) = \int y \,\mathsf{d}F(y), \qquad y_q = F^{-1}(q) = \text{ínf}\{y : F(y) \geq q\}.$$

Las cantidades $\mathsf{E}(Y)$, $\mathsf{var}(Y)$ y $F^{-1}(q)$ son llamadas *parámetros*, dependen de F y son generalmente desconocidos.

Observación. La familia $\mathcal{P} = \{\mathsf{P}_\theta : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k\}$ es parametrizada y decimos que el modelo es *paramétrico*.[2]

Observación. Recuerde que, para $A \subset \mathcal{Y}$, la función de probabilidad para el caso continuo es dada por

$$\mathsf{P}_\theta(A) = \int_A f(y; \boldsymbol{\theta}) \,\mathsf{d}\, y,$$

donde $f(y; \boldsymbol{\theta})$ es la función de densidad de Y. Mientras que para el caso discreto, tenemos

$$\mathsf{P}_\theta(A) = \sum_{y \in A} \mathsf{P}_\theta(\{y\}).$$

En efecto, usaremos la notación $\mathsf{P}_\theta(\boldsymbol{y})$ o $\mathsf{P}_\theta(\boldsymbol{Y} = \boldsymbol{y})$ en lugar de $\mathsf{P}_\theta(\{\boldsymbol{y}\})$.

Ejemplo 1.20. Considere $X_i \sim \mathsf{Bin}(n, \theta)$ es decir

$$p(x_i; \theta) = \binom{n}{x_i} \theta^{x_i} (1-\theta)^{n-x_i},$$

con $n \in \mathbb{N}$, $\theta \in (0,1)$ y $x_i \in \{0, 1, \ldots, n\}$. Es decir $\mathcal{X} = \{0, 1, \ldots, n\}$ y $\Theta = (0,1)$ (estamos considerando n *fijo*). De este modo,

$$\mathcal{P} = \{\mathsf{Bin}(n, \theta) : \theta \in (0,1)\}.$$

Ejemplo 1.21. Suponga que $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ son IID tal que $Y_i \sim \mathsf{Poi}(\theta)$, para $i = 1, \ldots, n$. Aquí

$$p(y_i; \theta) = \frac{e^{-\theta} \theta^{y_i}}{y_i!}, \qquad \theta \in (0, \infty),$$

y $y_i \in \{0, 1, \ldots\}$. Esto lleva al modelo estadístico

$$\mathcal{P} = \{\mathsf{Poi}(\theta)^{\otimes n} : \theta \in (0, \infty)\}.$$

[2] Si k no es entero, decimos que el modelo es *no paramétrico*.

Note que la *densidad conjunta* (asociada a $\mathsf{Poi}(\theta)^{\otimes n}$) es dada por

$$p(\boldsymbol{y};\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{y_i}}{y_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}.$$

EJEMPLO 1.22. Suponga $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\Big\{ -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\Big\}.$$

En este caso tenemos, $\mathcal{Y} = \mathbb{R}^p$, mientras que el modelo estadístico es definido como

$$\mathcal{P} = \{\mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} > 0\}.$$

EJEMPLO 1.23. Se ha sugerido que el crecimiento (durante siete semanas) de plantas de soya puede ser descrito por el modelo

$$Y_t = a + bt + \epsilon_t, \qquad t = 1, \ldots, 7.$$

Además, asumiremos que los errores satisfacen

$$\mathsf{E}(\epsilon_t) = 0, \qquad \mathsf{var}(\epsilon_t) = \sigma^2,$$

y $\mathsf{E}(\epsilon_t\epsilon_r) = 0$ para $t \neq r$. De esta manera el modelo para $\boldsymbol{Y} = (Y_1, \ldots, Y_7)^\top$ toma valores en $\mathcal{Y} = \mathbb{R}^7$ y puede ser escrito como:

$$\mathcal{P} = \{\mathsf{P}_\theta : \mathsf{E}(Y_t) = a + bt, \mathsf{var}(Y_t) = \sigma^2, \quad 1 \leq t \leq 7;$$
$$\boldsymbol{\theta} = (a, b, \sigma^2) \in \Theta \subset \mathbb{R}^2 \times \mathbb{R}_+\},$$

Usualmente en este tipo de problemas se asume que la distribución que caracteriza los primeros momentos de $\{\epsilon_t\}$ es conocida.

OBSERVACIÓN (Parametrización). La parametrización usada no es única. En lugar de $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$, podemos escoger

$$\mathcal{P} = \{\mathsf{P}_\phi : \phi \in \Phi\},$$

donde $\theta = h^{-1}(\phi)$ para h función 1 a 1.

EJEMPLO 1.24. Considere $X \sim \mathsf{Exp}(\lambda)$ con densidad

$$f(x;\lambda) = \lambda \exp(-\lambda x), \qquad x \geq 0, \lambda > 0.$$

Tenemos la parametrización alternativa, considerando $\mathsf{E}(X) = 1/\lambda = \mu$. De este modo,

$$f(x;\mu) = \frac{1}{\mu}\exp(-x/\mu), \qquad x \geq 0, \mu > 0.$$

SUPUESTO A0 (Identificabilidad). *Para $\theta_1, \theta_2 \in \Theta$ con $\theta_1 \neq \theta_2$. Si las distribuciones P_{θ_1} y P_{θ_2} son diferentes, entonces se dice que el modelo es* ***identificable****.*

La elección del modelo estadístico es esencial para realizar inferencia. Aunque todos los modelos están errados, algunos resultarán útiles y se deberá mantener el modelo tan sencillo como sea posible. Por tanto, una tarea crucial será evaluar la validez del modelo propuesto.

DEFINICIÓN 1.25. Una *estadística* T es una función de la muestra. Es decir,

$$T : \mathcal{Y} \to \mathcal{T},$$

tal que $T(\boldsymbol{y}) = t$. Note además que $T(\boldsymbol{Y})$ es una variable aleatoria con función de densidad $f(t;\theta)$.

EJEMPLO 1.26. Sean $Y_1, \ldots, Y_n$ variables aleatorias IID desde $\mathsf{N}(\mu, \sigma^2)$. En este caso,

$$\mathcal{P} = \{\mathsf{P}_\theta : \boldsymbol{\theta} = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+\}.$$

Ahora, considere

$$T_1(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad T_2(\boldsymbol{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Tenemos $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^2$, aquí $\boldsymbol{T}(Y_1, \ldots, Y_n) \to (T_1, T_2)$.

EJEMPLO 1.27. Considere la función de distribución empírica

$$\widehat{F}(X_1, \ldots, X_n)(x) = \widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x),$$

donde $(X_1, \ldots, X_n)^\top$ es una muestra aleatoria desde la distribución F (asociado al modelo $\mathcal{P}$) y $I(A)$ representa la función indicadora del evento A. Es decir, tenemos que $\widehat{F} : \mathbb{R}^n \to \mathbb{R}$ es una estadística.

OBSERVACIÓN. Un estadístico depende *sólo* de la muestra. **No** puede depender de cantidades desconocidas. Por ejemplo, para poblaciones normales

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ (= T_1), \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \ (= T_2),$$

son estadísticos para μ y σ^2, respectivamente. Mientras que

$$T_3 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2,$$

no es un estadístico para σ^2 pues depende de μ.

1.6. Familia exponencial

DEFINICIÓN 1.28. La clase de modelos $\{\mathsf{P}_\theta : \theta \in \Theta\}$ se dice una *familia exponencial* 1-paramétrica si existen funciones $\eta(\theta)$, $b(\theta)$ y funciones real valuadas T y h tal que su densidad adopta la forma

$$f(x) = \exp[\eta(\theta)T(x) - b(\theta)]h(x), \tag{1.3}$$

donde $x \in \mathcal{X} \subset \mathbb{R}^q$. Cuando una variable aletoria tiene la densidad en Ecuación (1.3), anotamos $X \sim \mathsf{FE}(\theta)$.

OBSERVACIÓN. El parámetro $\eta := \eta(\theta)$ es llamado *parámetro natural*, y

$$b(\eta) = \log \int \exp\big(\eta T(x)\big)h(x)\,\mathsf{d}\,x,$$

corresponde a un factor de normalización. Además, el modelo estadístico puede ser escrito como:

$$\mathcal{P} = \{\mathsf{P}_\eta : \eta \in \Gamma\}, \qquad \Gamma := \eta(\Theta), \tag{1.4}$$

donde el conjunto

$$\Gamma = \{\eta : b(\eta) < \infty\},$$

se denomina *espacio paramétrico natural*. Cuando escribimos la familia de densidades $\{\mathsf{P}_\eta : \eta \in \Gamma\}$ como en (1.4), decimos que la familia está escrita en *forma canónica*.

EJEMPLO 1.29. Sea $\mathsf{P}_\theta \overset{\mathsf{d}}{=} \mathsf{Poi}(\theta)$ con parámetro de media desconocido. Entonces $x \in \{0, 1, 2, \dots\}$, y

$$p(x;\theta) = \frac{\theta^x e^{-\theta}}{x!} = \frac{1}{x!}\exp(x\log\theta - \theta), \quad \theta > 0.$$

Es decir, P_θ corresponde a una FE (1-paramétrica) con $q = 1$, $\eta(\theta) = \log\theta$, $T(x) = x$, $b(\theta) = \theta$ y $h(x) = 1/x!$.

EJEMPLO 1.30. Sea $\mathsf{P}_\theta \overset{\mathsf{d}}{=} \mathsf{Bin}(n, \theta)$, $\theta \in (0, 1)$ y $x \in \{0, 1, \dots, n\}$. De este modo,

$$\begin{aligned} p(x;\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}\exp[x\log\theta + (n-x)\log(1-\theta)] \\ &= \binom{n}{x}\exp\left[x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right], \end{aligned}$$

y $\mathsf{Bin}(n, \theta)$ está en FE (1-paramétrica) con $q = 1$, $\eta(\theta) = \log(\theta/(1-\theta))$, $T(x) = x$, $b(\theta) = -n\log(1-\theta)$ y $h(x) = \binom{n}{x}$.

OBSERVACIÓN. Note que las funciones η, b y T *no son únicas*. En efecto, una manera alternativa de escribir la densidad de la famila exponencial es, por ejemplo,

$$f(x) = a(\theta)\exp[\eta(\theta)T(x)]h(x),$$

donde $h(x) \geq 0$ y $a(\theta) > 0$, representa un factor de normalización.

EJEMPLO 1.31 (contraejemplo). Sea $\mathcal{P}$ la clase de distribuciones exponencial de dos parámetros, digamos $\mathsf{Exp}(1, \theta)$, esto es con densidad

$$f(x) = \exp[-(x-\theta)]I_{[\theta,\infty)}(x), \qquad \theta \in \mathbb{R}.$$

Esta familia no pertenece a la clase FE.[3]

Suponga ahora que $X_1, \dots, X_n$ son variables aleatorias IID con distribución común P_θ en la FE 1-paramétrica y $\theta \in \Theta$. De este modo,

$$\begin{aligned} f(\boldsymbol{x}) &= \prod_{i=1}^n \exp[\eta(\theta)T(x_i) - b(\theta)]h(x_i) \\ &= \exp\left[\eta(\theta)\sum_{i=1}^n T(x_i) - nb(\theta)\right]\prod_{i=1}^n h(x_i) \\ &= \exp[\eta(\theta)T^{(n)}(\boldsymbol{x}) - b^{(n)}(\theta)]\widetilde{h}(\boldsymbol{x}) \end{aligned}$$

con

$$T^{(n)}(\boldsymbol{x}) = \sum_{i=1}^n T(x_i), \qquad b^{(n)}(\theta) = nb(\theta), \qquad \widetilde{h}(\boldsymbol{x}) = \prod_{i=1}^n h(x_i). \tag{1.5}$$

Es decir, la distribución conjunta de $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ también pertenece a la FE 1-paramétrica.

[3] El factor $I_{[\theta,\infty)}(x)$ no puede ser escrito en la forma exponencial.

RESULTADO 1.32. *Si X está en la FE 1-paramétrica. La función generadora de momentos de $T(X)$ es dada por*

$$M_T(s) = \exp[b(s+\eta) - b(\eta)],$$

para s en una vecindad de cero.

DEMOSTRACIÓN. Para obtener la MGF de $T(X)$, debemos calcular

$$\begin{aligned} M_T(s) = \mathsf{E}\{\exp(sT(x))\} &= \int_{\mathcal{X}} \exp(sT(x)) \exp(\eta T(x) - b(\eta)) h(x)\, \mathsf{d}\, x \\ &= \int_{\mathcal{X}} \exp[(s+\eta)T(x)) - b(\eta)] h(x)\, \mathsf{d}\, x \\ &= \exp[b(s+\eta) - b(\eta)] \int_{\mathcal{X}} \exp[(s+\eta)T(x)) - b(s+\eta)] h(x)\, \mathsf{d}\, x \\ &= \exp[b(s+\eta) - b(\eta)]. \end{aligned}$$

□

La *función generadora de cumulantes* está definida como

$$K_T(s) = \log M_T(s),$$

y permite obtener los cumulantes de T,

$$\kappa_r = \frac{\mathsf{d}^r}{\mathsf{d}\, s^r} K_T(s)\Big|_{s=0}.$$

En particular las primeras dos derivadas de $K_T(s)$ asumen la forma,

$$K_T' = M_T'/M_T, \qquad K_T'' = [M_T M_T'' - (M_T')^2]/M_T^2,$$

y evaluando en $s = 0$, lleva a los primeros cumulantes:

$$\kappa_1 = \mathsf{E}(T), \qquad \kappa_2 = \mathsf{E}(T^2) - (\mathsf{E}(T))^2 = \mathsf{var}(T).$$

Para la familia exponencial, sigue que la función generadora de cumulantes para T es dada por:

$$K_T(s) = b(s+\eta) - b(\eta).$$

RESULTADO 1.33. *Sea X una variable aleatoria perteneciente a la FE 1-paramétrica. De este modo,*

$$\mathsf{E}(T) = b'(\eta), \qquad \mathsf{var}(T) = b''(\eta).$$

DEMOSTRACIÓN. Basta notar que

$$\begin{aligned} \mathsf{E}(T) &= \frac{\mathsf{d}}{\mathsf{d}\, s} K_T(s)\Big|_{s=0} = b'(s+\eta)\big|_{s=0} = b'(\eta), \\ \mathsf{var}(T) &= \frac{\mathsf{d}^2}{\mathsf{d}\, s^2} K_T(s)\Big|_{s=0} = b''(s+\eta)\big|_{s=0} = b''(\eta). \end{aligned}$$

□

Para una muestra IID, $X_1, \ldots, X_n$, y usando el Resultado 1.33 sigue inmediatamente que (ver Ecuación (1.5))

$$\mathsf{E}(T^{(n)}(\boldsymbol{X})) = \mathsf{E}\Big(\sum_{i=1}^{n} T(X_i)\Big) = \sum_{i=1}^{n} \mathsf{E}(T(X_i)) = nb'(\theta)$$
$$\mathsf{var}(T^{(n)}(\boldsymbol{X})) = \mathsf{var}\Big(\sum_{i=1}^{n} T(X_i)\Big) = \sum_{i=1}^{n} \mathsf{var}(T(X_i)) = nb''(\theta).$$

Sea $\mu = \mathsf{E}(Y)$, es decir, tenemos una transformación[4] 1-1 entre θ y el parámetro de media

$$\mu = \mu(\theta) = b'(\theta),$$

pues $\eta = \eta(\theta)$. Además, conforme θ varia en Θ, tenemos que $\mu(\theta)$ es conocida como *función (o superficie) de esperanza*.

Cuando $Y \sim \mathsf{FE}(\theta)$ con función generadora de cumulantes $b(\cdot)$, tenemos $\mu(\theta) = b'(\theta)$. Podemos parametrizar la densidad $f(y)$ en términos de μ. De este modo, haciendo $\theta = \theta(\mu)$, sigue que

$$\mathsf{var}(Y) = b''(\theta) = \frac{\mathsf{d}\,\mu}{\mathsf{d}\,\theta}\Big|_{\theta=\theta(\mu)} = V(\mu),$$

donde $V(\mu)$ es llamada *función de varianza* de Y.

Definición 1.34. La familia de distribuciones $\{\mathsf{P}_\theta : \boldsymbol{\theta} \in \Theta\}$ con $\Theta \subset \mathbb{R}^k$, se dice una *familia exponencial* k-paramétrica si existen funciones $\eta_1(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta})$ y $b(\boldsymbol{\theta})$ y funciones real-valuadas $T_1, \ldots, T_k$, h tal que la densidad puede ser escrita como

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \exp[\boldsymbol{\eta}^\top(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{x}) - b(\boldsymbol{\theta})]h(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^p,$$

donde $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta}))^\top$ y $\boldsymbol{T}(\boldsymbol{x}) = (T_1(\boldsymbol{x}), \ldots, T_k(\boldsymbol{x}))^\top$.

Ejemplo 1.35. Suponga $\mathsf{P}_\theta \overset{\mathsf{d}}{=} \mathsf{N}_1(\mu, \sigma^2)$,

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

La densidad de P_θ puede ser escrita como

$$\begin{aligned} f(x; \boldsymbol{\theta}) &= (2\pi\sigma^2)^{-1/2} \exp\Big\{ -\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2 \Big\} \\ &= \exp\Big\{ \frac{\mu\, x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\Big(\frac{\mu^2}{\sigma^2} + \log 2\pi\sigma^2\Big)\Big\}. \end{aligned}$$

De este modo,

$$\eta_1(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}, \quad \eta_2(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}, \quad T_1(x) = x, \quad T_2(x) = x^2,$$

y $b(\boldsymbol{\theta}) = \frac{1}{2}(\mu^2/\sigma^2 + \log 2\pi\sigma^2)$, $h(x) = 1$. De este modo, $\mathsf{N}_1(\mu, \sigma)$ corresponde a una FE 2-paramétrica.

Ejemplo 1.36. Considere la función de probabilidad

$$p(x; \boldsymbol{\theta}) = \theta_1^{I_1(x)} \theta_2^{I_2(x)} \theta_3^{I_3(x)},$$

con

$$I_j(x) = \begin{cases} 1, & x = j, \\ 0, & \text{en otro caso}, \end{cases}$$

[4] suave y estrictamente convexa

para $j = 1, 2, 3$. Así podemos escribir

$$p(x;\boldsymbol{\theta}) = \exp[I_1(x)\log\theta_1 + I_2(x)\log\theta_2 + I_3(x)\log\theta_3],$$

es decir X pertenece a la FE. *Sin embargo*,

$$I_1(x) + I_2(x) + I_3(x) = 1.$$

Por tanto, la familia **no** es estrictamente 3-dimensional. Considere

$$\begin{aligned} p(x;\boldsymbol{\theta}) &= \exp[I_1(x)\log\theta_1 + I_2(x)\log\theta_2 + (1 - I_1(x) - I_2(x))\log\theta_3] \\ &= \theta_3 \exp[I_1(x)(\log\theta_1 - \log\theta_3) + I_2(x)(\log\theta_2 - \log\theta_3)]. \end{aligned}$$

Sea $\theta_3 = 1 - \theta_1 - \theta_2$, sigue que

$$p(x;\boldsymbol{\theta}) = \exp\left[I_1(x)\log\left(\frac{\theta_1}{1-\theta_1-\theta_2}\right) + I_2(x)\log\left(\frac{\theta_2}{1-\theta_1-\theta_2}\right) + \log(1-\theta_1-\theta_2)\right].$$

Finalmente, tenemos que X pertenece a la FE (2-paramétrica) con

$$T_1(x) = I_1(x), \quad T_2(x) = I_2(x), \quad \eta_r = \log\left(\frac{\theta_r}{1-\theta_1-\theta_2}\right),\ r = 1, 2,$$

y $b(\boldsymbol{\theta}) = -\log(1 - \theta_1 - \theta_2)$, $h(x) = 1$.

Observación. Evidentemente, resulta fácil extender los Resultados 1.32 y 1.33. En efecto,

$$M_T(\boldsymbol{s}) = \mathsf{E}[\exp(\boldsymbol{s}^\top \boldsymbol{T})] = \exp[b(\boldsymbol{\eta} + \boldsymbol{s}) - b(\boldsymbol{\eta})], \qquad K_T(\boldsymbol{s}) = b(\boldsymbol{\eta} + \boldsymbol{s}) - b(\boldsymbol{\eta}),$$

y

$$\mathsf{E}(\boldsymbol{T}(\boldsymbol{X})) = \dot{b}(\boldsymbol{\eta}) = \frac{\partial b(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}, \qquad \mathsf{Cov}(\boldsymbol{T}(\boldsymbol{X})) = \ddot{b}(\boldsymbol{\eta}) = \frac{\partial^2 b(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top}.$$

En aplicación de regresión,[5] es útil considerar una definición alternativa de la familia exponencial, incluyendo un parámetro de dispersión. Considere la siguiente definición

Definición 1.37. Se dice que $\mathsf{P}_{\theta,\phi}$ sigue una familia exponencial con parámetro de escala $\phi > 0$, si la función de densidad de un vector aleatorio n-dimensional $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ es dada por:

$$f(\boldsymbol{y};\boldsymbol{\theta},\phi) = \exp[\phi\{\boldsymbol{y}^\top\boldsymbol{\theta} - b(\boldsymbol{\theta})\} + c(\boldsymbol{y},\phi)], \tag{1.6}$$

donde $b(\cdot)$ y $c(\cdot,\cdot)$ son funciones apropiadas, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ denota el parámetro natural ($\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n$).

El parámetro de dispersión $\phi > 0$ usualmente es considerado un parámetro molesto y cuando un vector aleatorio tiene la densidad dada en (1.6), anotamos $\boldsymbol{Y} \sim \mathsf{FE}(\boldsymbol{\theta}, \phi)$.

Observación. Cuando ϕ es desconocido, se suele indicar que el modelo está en la *familia de dispersión exponencial* (Jørgensen, 1997).

Ejemplo 1.38. Suponga $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Tenemos que su función de densidad puede ser escrita como

$$\begin{aligned} f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\} \\ &= \exp\{\boldsymbol{y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - \tfrac{1}{2}\log|2\pi\boldsymbol{\Sigma}|\}, \end{aligned}$$

[5] En particular en modelos lineales generalizados (GLM)

de este modo, el parámetro natural es $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ y $b(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta}$, mientras que $c(\boldsymbol{y}, \boldsymbol{\Phi}) = -\frac{1}{2}(\boldsymbol{y}^\top\boldsymbol{\Phi}\boldsymbol{y} - \log|2\pi\boldsymbol{\Phi}^{-1}|)$.

EJEMPLO 1.39. Usando que

$$\boldsymbol{y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\boldsymbol{y}^\top = (\operatorname{vec}\boldsymbol{y}\boldsymbol{y}^\top)^\top\operatorname{vec}\boldsymbol{\Sigma}^{-1},$$

podemos escribir $\boldsymbol{Y}$ como un miembro de la familia exponencial, considerando

$$\boldsymbol{T}_1(\boldsymbol{Y}) = \boldsymbol{Y}, \quad \boldsymbol{T}_2(\boldsymbol{Y}) = \boldsymbol{Y}\boldsymbol{Y}^\top, \quad \boldsymbol{\eta}_1(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \quad \boldsymbol{\eta}_2(\boldsymbol{\theta}) = \operatorname{vec}\boldsymbol{\Sigma}^{-1}.$$

Las siguientes clases de distribuciones pueden ser caracterizadas mediante transformaciones de vectores aleatorios.

1.7. Familia de posición-escala

Considere U una variable aleatoria con función de distribución F. De este modo, la variable

$$X = U + a,$$

tiene función de distribución

$$\mathsf{P}(X \le x) = F(x - a),$$

para F fijo y $a \in \mathbb{R}$ tenemos que X corresponde a la *familia de posición*. Análogamente la *familia de escala* es generada por la transformación

$$X = bU, \qquad b > 0,$$

en cuyo caso,

$$\mathsf{P}(X \le x) = F(x/b), \qquad b > 0.$$

Esto lleva a la siguiente definición

DEFINICIÓN 1.40. Sea U una variable aleatoria con función de distribución acumulada fija F y considere la trasnformación

$$X = a + bU, \qquad a \in \mathbb{R}, b > 0.$$

Tenemos

$$\mathsf{P}(X \le x) = F\Big(\frac{x-a}{b}\Big).$$

De este modo, X es conocida como *familia de posición-escala*.

Usualmente asociado a F tenemos una función de densidad f, dada por

$$f(x; a, b) = \frac{\mathsf{d}}{\mathsf{d}x}F\Big(\frac{x-a}{b}\Big) = \frac{1}{b}F'\Big(\frac{x-a}{b}\Big) = \frac{1}{b}f\Big(\frac{x-a}{b}\Big).$$

EJEMPLO 1.41. Algunas familias de posición-escala corresponden a:

- Normal, $\mathsf{N}(a, b^2)$:

$$f(y; a, b) = \frac{1}{b}(2\pi)^{-1/2}\exp\Big\{-\frac{1}{2}\Big(\frac{y-a}{b}\Big)^2\Big\}.$$

- Laplace (doble exponencial), $\mathsf{Laplace}(a, b)$:

$$f(y; a, b) = \frac{1}{2b}\exp\Big\{-\frac{|y-a|}{b}\Big\}.$$

- Cauchy, $\mathsf{Cauchy}(a, b^2)$:

$$f(y; a, b) = \frac{b}{\pi}\frac{1}{b^2 + (y-a)^2}.$$

- Logística, $\mathsf{Logística}(a, b)$:

$$f(y; a, b) = \frac{1}{b} \frac{e^{-(y-a)/b}}{(1 + e^{-(y-a)/b})^2}$$

1.8. Familia de distribuciones de contornos elípticos

Esta clase de distribuciones es ampliamente utilizada en modelamiento estadístico desde la perspectiva de robustez, y corresponde a una extensión natural de la distribución normal multivariada así como de la familia de posición-escala. Para introducir ideas, primeramente revisaremos algunas definiciones de la normal.

DEFINICIÓN 1.42. Un vector aleatorio p-dimensional, $\boldsymbol{X}$ tiene distribución normal con vector de medias $\boldsymbol{\mu} \in \mathbb{R}^p$ y matriz de covarianza $\mathsf{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} > 0$ sólo si, para todo vector $\boldsymbol{t}$ la variable aleatoria (uni-dimensional) $\boldsymbol{t}^\top \boldsymbol{X}$ es normal y escribimos $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Note que en la definición anterior **no** se ha hecho supuestos respecto de la independencia de los componentes de $\boldsymbol{X}$.

DEFINICIÓN 1.43. La función característica de $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ está dada por

$$\varphi_X(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top \boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}).$$

Podemos derivar la función característica de $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ desde la normal univariada notando que $Y = \boldsymbol{t}^\top \boldsymbol{X}$ tiene media y varianza dadas por, $\lambda = \mathsf{E}(Y) = \boldsymbol{t}^\top \boldsymbol{\mu}$ y $\sigma^2 = \mathsf{var}(Y) = \boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t} \geq 0$ y tomando

$$\varphi_Y(\boldsymbol{t}) = \varphi_{t^\top X}(1) = \exp(i\lambda - \tfrac{1}{2}\sigma^2) = \exp(i\boldsymbol{t}^\top \boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t})$$

Como un caso particular tenemos que la función característica para $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p)$, es

$$\varphi_Z(\boldsymbol{t}) = \exp(-\tfrac{1}{2}\sigma^2 \boldsymbol{t}^\top \boldsymbol{t}) = \prod_{i=1}^{p} \exp(-\tfrac{1}{2}\sigma^2 t_i^2) = \prod_{i=1}^{p} \varphi_{Z_i}(t_i)$$

y de este modo, se tiene que

$$\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p) \quad \Longleftrightarrow \quad Z_1, \ldots, Z_p \ \mathsf{IID}\ \mathsf{N}(0, \sigma^2).$$

RESULTADO 1.44. *Suponga* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *y considere la transformación lineal*

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b},$$

donde $\boldsymbol{A} \in \mathbb{R}^{m \times p}$ *con* $\mathrm{rg}(\boldsymbol{A}) = m$. *Entonces* $\boldsymbol{Y} \sim \mathsf{N}_m(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$.

DEMOSTRACIÓN. Para mostrar el resultado usaremos la función característica de $\boldsymbol{X}$. Note que

$$\begin{aligned} \varphi_Y(\boldsymbol{t}) &= \mathsf{E}\{\exp(i\boldsymbol{t}^\top \boldsymbol{Y})\} = \mathsf{E}\{\exp(i\boldsymbol{t}^\top (\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}))\} \\ &= \exp(i\boldsymbol{t}^\top \boldsymbol{b})\, \mathsf{E}(i\boldsymbol{t}^\top \boldsymbol{A}\boldsymbol{X}) \end{aligned}$$

Sea $\boldsymbol{h} = \boldsymbol{t}^\top \boldsymbol{A}$, entonces

$$\begin{aligned} \varphi_Y(\boldsymbol{t}) &= \exp(i\boldsymbol{t}^\top \boldsymbol{b})\, \mathsf{E}\left(i\boldsymbol{h}^\top \boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}\right) = \exp(i\boldsymbol{t}^\top \boldsymbol{b})\, \mathsf{E}\left(i\boldsymbol{t}^\top \boldsymbol{A}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top \boldsymbol{t}\right) \\ &= \exp\left(i\boldsymbol{t}^\top (\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}) - \tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top \boldsymbol{t}\right), \end{aligned}$$

y el resultado sigue. □

RESULTADO 1.45. *Si* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *y* $\boldsymbol{\Sigma}$ *es definida positiva, entonces la densidad de* $\boldsymbol{X}$ *es*

$$f(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}.$$

DEMOSTRACIÓN. Sea $Z_1, \ldots, Z_p$ variables aleatorias $\mathsf{IID}\ \mathsf{N}(0,1)$. Entonces la densidad conjunta de $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ es

$$f(\boldsymbol{z}) = \prod_{i=1}^{p} (2\pi)^{-1/2} \exp(-z_i^2/2) = (2\pi)^{-p/2} \exp(-\tfrac{1}{2}\|\boldsymbol{z}\|^2).$$

Considere $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Z}$ con $\boldsymbol{\mu} \in \mathbb{R}^p$ y $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$, con $\boldsymbol{B}$ matriz de rango completo. Entonces, tenemos la transformación inversa

$$\boldsymbol{Z} = \boldsymbol{g}^{-1}(\boldsymbol{X}) = \boldsymbol{B}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}),$$

y $\mathsf{d}\,\boldsymbol{Z} = \mathsf{d}\,\boldsymbol{g}^{-1}(\boldsymbol{X}) = \boldsymbol{B}^{-1}\,\mathsf{d}\,\boldsymbol{X}$, con matriz jacobiana $\mathsf{D}\boldsymbol{g}^{-1}(\boldsymbol{X}) = \boldsymbol{B}^{-1}$, como

$$|\mathsf{D}\boldsymbol{g}^{-1}(\boldsymbol{X})|_+ = |\boldsymbol{B}|^{-1} = |\boldsymbol{B}\boldsymbol{B}^\top|^{-1/2},$$

obtenemos

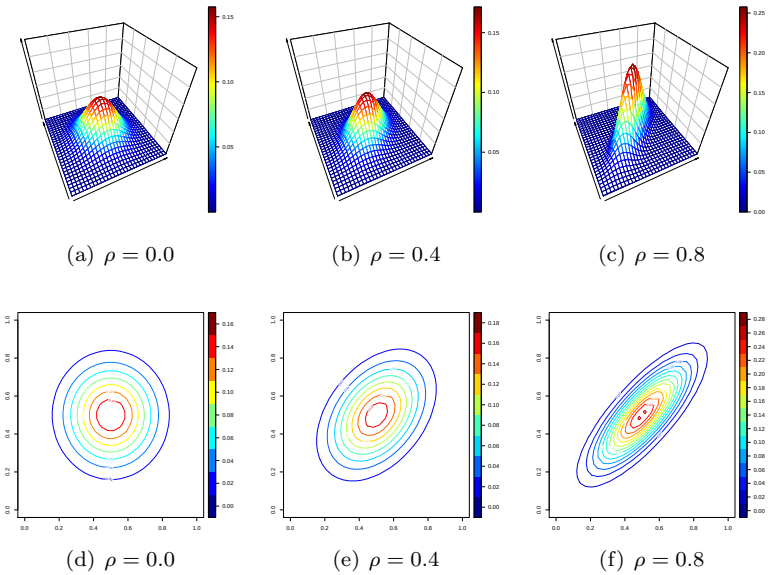
$$\begin{aligned} f(\boldsymbol{x}) &= |\mathsf{D}\boldsymbol{g}^{-1}(\boldsymbol{x})|_+ f_{\boldsymbol{z}}(\boldsymbol{g}^{-1}(\boldsymbol{x})) \\ &= (2\pi)^{-p/2} |\boldsymbol{B}\boldsymbol{B}^\top|^{-1/2} \exp\{\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{B}^{-\top}\boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}, \end{aligned}$$

notando que $\boldsymbol{\Sigma}^{-1} = \boldsymbol{B}^{-\top}\boldsymbol{B}^{-1}$ sigue el resultado deseado. □

EJEMPLO 1.46. Sea $\boldsymbol{X} \sim \mathsf{N}_2(\boldsymbol{0}, \boldsymbol{\Sigma})$ donde

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad -1 < \rho < 1.$$

A continuación se presenta la función de densidad para los casos $\rho = 0.0, 0.4$ y 0.8.



(a) $\rho = 0.0$ (b) $\rho = 0.4$ (c) $\rho = 0.8$

(d) $\rho = 0.0$ (e) $\rho = 0.4$ (f) $\rho = 0.8$

Figura 1. Densidad de $\boldsymbol{X} \sim \mathsf{N}_2(\boldsymbol{0}, \boldsymbol{\Sigma})$ para $\rho = 0.0, 0.4$ y 0.8.

Note que la función de densidad es constante sobre el elipsoide en $\mathbb{R}^p$

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = k,$$

para todo $k > 0$. Este elipsoide tiene centro $\boldsymbol{\mu}$, mientras que $\boldsymbol{\Sigma}$ determina su forma y orientación.

DEFINICIÓN 1.47. Sea $\boldsymbol{U}$ vector aleatorio $p \times 1$ con distribución uniforme sobre el conjunto

$$\mathcal{S}_p = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\| = 1\},$$

la superficie de la esfera unitaria en $\mathbb{R}^p$, y anotamos

$$\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p).$$

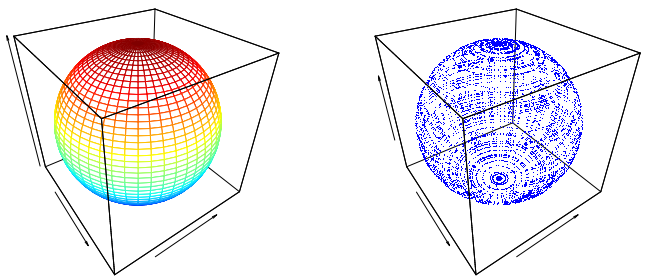
La siguiente definición establece una relación entre la distribución normal y la uniforme sobre la esfera unitaria

PROPIEDAD 1.48. Si $Z_1, \ldots, Z_p$ son variables aleatorias IID con distribución $\mathsf{N}(0,1)$, entonces $\boldsymbol{U} = (U_1, \ldots, U_p)^\top$, definido como

$$\boldsymbol{U} = \frac{\boldsymbol{Z}}{\|\boldsymbol{Z}\|},$$

tiene distribución uniforme sobre la esfera unitaria $\mathcal{S}_p$.

El resultado anterior es muy relevante pues permite definir la densidad de un vector aleatorio $\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p)$ y ofrece un procedimiento muy simple para generar observaciones sobre la esfera unitaria. Considere el siguiente gráfico,



(a) Esfera unitaria (b) Datos simulados

Figura 2. Esfera unitaria y datos generados sobre la superficie $\mathcal{S}_p$.

DEFINICIÓN 1.49. Un vector aleatorio $p \times 1$, $\boldsymbol{X}$ se dice que tiene simetría esférica si para cualquier $\boldsymbol{Q} \in \mathcal{O}_p$, sigue que

$$\boldsymbol{Q}\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{X}.$$

EJEMPLO 1.50. Sea $\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p)$, entonces es bastante obvio que $\boldsymbol{Q}\boldsymbol{U} \overset{\mathsf{d}}{=} \boldsymbol{U}$.

DEFINICIÓN 1.51. Un vector aleatorio p-dimensional tiene distribución esférica sólo si su función característica satisface

(a) $\varphi(\boldsymbol{Q}^\top \boldsymbol{t}) = \varphi(\boldsymbol{t})$, para todo $\boldsymbol{Q} \in \mathcal{O}_p$.

(b) Existe una función $\phi(\cdot)$ de una variable escalar tal que $\varphi(\boldsymbol{t}) = \phi(\boldsymbol{t}^\top \boldsymbol{t})$.

En este caso escribimos $\boldsymbol{X} \sim \mathsf{S}_p(\phi)$.

EJEMPLO 1.52. Sea $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$, tenemos que

$$\varphi(\boldsymbol{t}) = \exp\{-\tfrac{1}{2}(t_1^2 + \cdots + t_p^2)\} = \exp(-\tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{t}).$$

RESULTADO 1.53. *Sea $\psi(\boldsymbol{t}^\top \boldsymbol{t})$ la función característica del vector aleatorio $\boldsymbol{X}$. Entonces $\boldsymbol{X}$ tiene representación estocástica*

$$\boldsymbol{X} \overset{\mathsf{d}}{=} R\boldsymbol{U},$$

donde $\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p)$ y $R \sim F(\boldsymbol{X})$ son independientes.

RESULTADO 1.54. *Suponga que $\boldsymbol{X} \overset{\mathsf{d}}{=} R\boldsymbol{U} \sim \mathsf{S}_p(\phi)$ $(\mathsf{P}(\boldsymbol{X} = \boldsymbol{0}) = 0)$, entonces*

$$\|\boldsymbol{X}\| \overset{\mathsf{d}}{=} R, \qquad \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \overset{\mathsf{d}}{=} \boldsymbol{U}.$$

Además $\|\boldsymbol{X}\|$ y $\boldsymbol{X}/\|\boldsymbol{X}\|$ son independientes.

DEFINICIÓN 1.55 (Distribución de contornos elípticos). Un vector aleatorio $p \times 1$, $\boldsymbol{X}$ tiene distribución de *contornos elípticos* con parámetros $\boldsymbol{\mu} \in \mathbb{R}^p$ y $\boldsymbol{\Sigma} \geq 0$, si

$$\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Y}, \qquad \boldsymbol{Y} \sim \mathsf{S}_k(\phi),$$

donde $\boldsymbol{B} \in \mathbb{R}^{p\times k}$ es matriz de rango completo tal que, $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{\Sigma}$ con $\mathrm{rg}(\boldsymbol{\Sigma}) = k$. En cuyo caso escribimos $\boldsymbol{X} \sim \mathsf{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \phi)$.

OBSERVACIÓN. La función característica de $\boldsymbol{X} \sim \mathsf{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \phi)$ es de la forma:

$$\varphi(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top \boldsymbol{\mu})\phi(\boldsymbol{t}^\top \boldsymbol{\Sigma}\boldsymbol{t}).$$

Note además que la representacióin estocástica de $\boldsymbol{X}$ es dada por

$$\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{\mu} + R\,\boldsymbol{B}\boldsymbol{U},$$

donde $R \geq 0$ es independiente de $\boldsymbol{U}$ y $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{\Sigma}$.

DEFINICIÓN 1.56. Se dice que el vector $\boldsymbol{X}$ tiene distribución de contornos elípticos si su función de densidad es de la forma

$$f(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2} g((\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})), \qquad \boldsymbol{x} \in \mathbb{R}^p,$$

donde $g : \mathbb{R} \to [0, \infty)$ es función decreciente, llamada *función generadora de densidad*, tal que:

$$\int_0^\infty u^{p/2-1} g(u)\, \mathsf{d}u < \infty.$$

OBSERVACIÓN. Asuma que $\boldsymbol{X} \sim \mathsf{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \phi)$ con $\mathrm{rg}(\boldsymbol{\Sigma}) = k$. Entonces,

$$Q(\boldsymbol{X}) = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^-(\boldsymbol{X} - \boldsymbol{\mu}) \overset{\mathsf{d}}{=} R^2,$$

donde $\boldsymbol{\Sigma}^-$ es una inversa generalizada de $\boldsymbol{\Sigma}$.

EJEMPLO 1.57 (Distribución t multivariada). La función generadora de densidad de un vector aleatorio con distribución t multivariada asume la forma

$$g(u) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{p/2}} \Big(1 + \frac{u}{\nu}\Big)^{-(\nu+p)/2}, \qquad \nu > 0.$$

En cuyo caso escribimos, $\boldsymbol{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Para esta distribución, tenemos que $R^2/p \sim F_{p,\nu}$.

EJEMPLO 1.58 (Distribución Exponencial Potencia). Para la distribución Exponencial Potencia (Gómez et al., 1988), la función generadora de densidades es dada por

$$g(u) = \frac{p\Gamma(\frac{p}{2})\pi^{-p/2}}{\Gamma(1+\frac{p}{2\lambda})2^{1+\frac{p}{2\lambda}}} \exp(-u^{\lambda}/2), \qquad \lambda > 0.$$

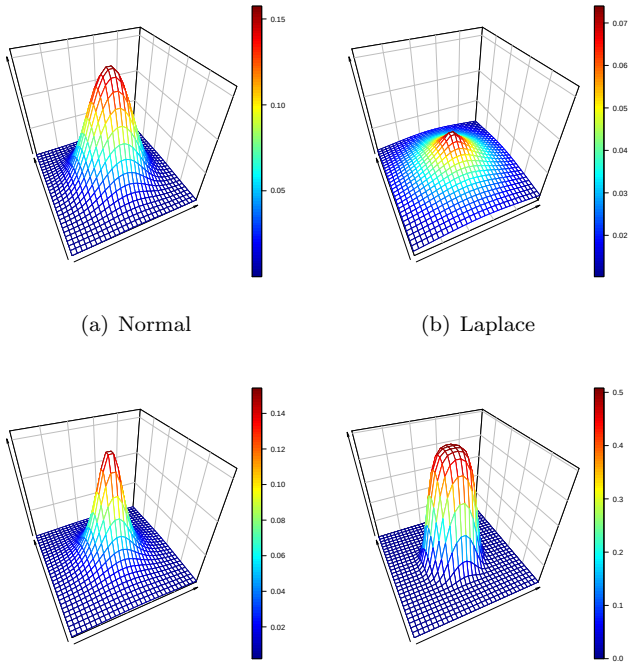
y es usual utilizar la notación $\boldsymbol{X} \sim \mathsf{PE}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda)$. En este caso tenemos que la variable aleatoria positiva R tiene densidad

$$h(r) = \frac{p}{\Gamma(1+\frac{p}{2\lambda})2^{\frac{p}{2\lambda}}} r^{p-1} \exp(-r^{2\lambda}/2), \qquad r > 0.$$

Note también que $R^{2\lambda} \sim \mathsf{Gama}(\frac{1}{2}, \frac{p}{2\lambda})$.

EJEMPLO 1.59. En la siguiente figura se presenta la densidad asociadas a las siguientes funciones g:

- Normal: $g(u) = c_1 \exp(-u/2)$.
- Laplace: $g(u) = c_2 \exp(-\sqrt{u}/2)$.
- Cauchy: $g(u) = c_3(1+u)^{-(p+1)/2}$.
- Exponencial potencia (PE): $g(u) = c_4 \exp(-u^{\lambda}/2)$, $\lambda = 2$.



(a) Normal (b) Laplace

(c) Cauchy (d) PE, $\lambda = 2$

Figura 3. Funciones de densidad del vector $\boldsymbol{X} \sim \mathsf{EC}_2(\boldsymbol{0}, \boldsymbol{I}; g)$ para las distribuciones normal, Laplace, Cauchy y exponencial potencia con $\lambda = 2$.

DEFINICIÓN 1.60 (Distribución de mezcla de escala normal). Sea $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma}$ matriz $p \times p$ definida positiva y H función de distribución de una variable aleatoria positiva, W. Entonces, se dice que el vector aleatorio $\boldsymbol{X}$ sigue una *distribución de mezcla de escala normal* si su función de densidad asume la forma

$$f(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \int_0^\infty w^{p/2} \exp(-wu/2)\, \mathsf{dH}(w),$$

donde $u = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ y anotamos $\boldsymbol{X} \sim \mathsf{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathsf{H})$.

Esta familia de distribuciones ha sido frecuentemente sugerida como una interesante alternativa para producir estimadores robustos, manteniendo la elegancia y simplicidad de la teoría de máxima verosimilitud.

EJEMPLO 1.61 (Distribución Slash). Un vector aleatorio $\boldsymbol{X}$ tiene distribución Slash si su función de densidad es de la forma:

$$f(\boldsymbol{x}) = \nu(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \int_0^1 w^{p/2+\nu-1} \exp(-wu/2)\, \mathsf{d}w.$$

En este caso, tenemos que $h(w) = \nu w^{\nu-1}$, para $w \in (0,1)$ y $\nu > 0$. Es decir, $W \sim \mathsf{Beta}(\nu, 1)$.

OBSERVACIÓN. Un vector aleatorio $\boldsymbol{X} \sim \mathsf{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathsf{H})$ si admite la representación

$$\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{\mu} + W^{-1/2}\boldsymbol{Z},$$

donde $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ y $W \sim \mathsf{H}(\boldsymbol{\delta})$ son independientes. Equivalentemente, podemos utilizar la estructura jerárquica:

$$\boldsymbol{X}|W \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/w), \qquad W \sim \mathsf{H}(\boldsymbol{\delta}).$$

Por ejemplo, un vector aleatorio $\boldsymbol{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ puede ser caracterizado como:

$$\boldsymbol{X}|W \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/w), \qquad W \sim \mathsf{Gamma}(\nu/2, \nu/2).$$

Capítulo 2

Elementos de Inferencia

2.1. 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 2.1 (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 2.2. 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 2.3 (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 2.4. Sea $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ variables IID desde una distribución Geo(θ). 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 2.5. Sea $X_1, \dots, X_n$ una m.a.(n) desde 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).

Observación. Sea $X_1, \dots, X_n$ variables aleatorias IID desde FE(θ). Tenemos que,

$$f(\boldsymbol{x}; \theta) = \exp\Big[\sum_{i=1}^n T(X_i)\eta(\theta) - nb(\theta)\Big]\widetilde{h}(\boldsymbol{x}).$$

Es decir, $\sum_{i=1}^n T(X_i)$ es estadística suficiente para θ.

2.2. Función de verosimilitud

Para motivar ideas, suponga $X_1, \ldots, X_n$ variables aleatorias IID desde una función de distribución desconocida $G(x)$. Asumiremos que $G(x)$ corresponde al modelo estadístico verdadero (o verdadera CDF). Además, sea $F(x)$ un modelo arbitrareo (el modelo asumido). Supondremos también que asociadas a $G(x)$ y $F(x)$ tenemos funciones de densidad $g(x)$ y $f(x)$, respectivamente.

IDEA. Se desea determinar la bondad del modelo asumido $f(x)$ en términos de su cercanía con el modelo verdadero.

La información de Kullback-Leibler (KL) entre las funciones de densidad $g(x)$ y $f(x)$ es dada por:[1]

$$I(g:f) = \int \log\Big(\frac{g(x)}{f(x)}\Big) g(x)\, \mathsf{d}x = \mathsf{E}_G\Big[\log\Big(\frac{g(x)}{f(x)}\Big)\Big].$$

Propiedades de la información KL (o divergencia):

(a) $I(g:f) \geq 0$.
(b) $I(g:f) = 0 \Leftrightarrow g(x) = f(x)$ (casi en toda parte).

EJEMPLO 2.6. Suponga que G y F están dadas, respectivamente por $\mathsf{N}(\theta, \phi^2)$ y $\mathsf{N}(\mu, \sigma^2)$. Entonces,

$$\begin{aligned}\mathsf{E}_G[(X-\mu)^2] &= \mathsf{E}_G[(X-\theta+\theta-\mu)^2] \\ &= \mathsf{E}_G[(X-\theta)^2 + 2(X-\theta)(\theta-\mu) + (\theta-\mu)^2] \\ &= \mathsf{E}_G[(X-\theta)^2] + (\theta-\mu)^2 = \phi^2 + (\theta-\mu)^2\end{aligned}$$

Ahora, para

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\Big\{-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big\},$$

sigue que

$$\begin{aligned}\mathsf{E}_G(\log f(x)) &= \mathsf{E}_G\Big[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(X-\mu)^2\Big] \\ &= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}[\phi^2 + (\theta-\mu)^2]\end{aligned}$$

mientras que (basta notar que $\mathsf{E}_G[(X-\theta)^2] = \phi^2$):

$$\mathsf{E}_G(\log g(x)) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}.$$

De este modo, la información KL del modelo $f(x)$ con respecto a $g(x)$ asume la forma:

$$I(g:f) = \mathsf{E}_G(\log g(x)) - \mathsf{E}_G(\log f(x)) = \frac{1}{2}\Big\{\log\frac{\sigma^2}{\phi^2} + \frac{\phi^2 + (\theta-\mu)^2}{\sigma^2} - 1\Big\}$$

OBSERVACIÓN. Note que el cálculo de la información KL puede no ser factible pues, en general, la distribución g **no** es conocida.

Además, como

$$I(g:f) = \mathsf{E}_G\Big[\log\frac{g(x)}{f(x)}\Big] = \mathsf{E}_G[\log g(x)] - \mathsf{E}_G[\log f(x)],$$

[1] En ocasiones anotamos $I(G:F) = \int \log(g/f)\, \mathsf{d}G$.

para comparar distintos modelos competitivos basta considerar solamente el segundo término, el que es llamado *log-verosimilitud esperada*.

Claramente, mientras mayor sea este valor, mejor será el modelo (será más cercano a g). Además,

$$\mathsf{E}_G[\log f(x)] = \int \log f(x)\, \mathsf{d}G(x),$$

aún depende de la verdadera distribución. Sin embargo, podemos obtener un buen estimador usando en la CDF empírica $\widehat{G}_n$ basada en los datos observados $X_1, \ldots, X_n$. Es decir,

$$\mathsf{E}_{\widehat{G}_n}[\log f(x)] = \int \log f(x)\, \mathsf{d}\widehat{G}_n(x) = \sum_{i=1}^{n} \widehat{g}_n(x_i) \log f(x_i) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i).$$

En efecto, de acuerdo a la Ley de los grandes números,

$$\frac{1}{n}\sum_{i=1}^{n} \log f(x_i) \xrightarrow{n\to\infty} \mathsf{E}_G[\log f(x)].$$

Esto lleva a la definición.

Definición 2.7 (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 2.8. 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 2.9. 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

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

El siguiente lema permite justificar que aquél valor de $\boldsymbol{\theta}$ que se ajusta bien a los datos $\boldsymbol{x}$ debe maximizar $L(\boldsymbol{\theta}; \boldsymbol{x})$.

LEMA 2.10 (Principio de verosimilitud). *Sea $\boldsymbol{\theta}_0$ el parámetro subyacente (o verdadero). Entonces,*

$$\mathsf{E}_{\theta_0}\{\ell(\boldsymbol{\theta}; \boldsymbol{X})\} \leq \mathsf{E}_{\theta_0}\{\ell(\boldsymbol{\theta}_0; \boldsymbol{X})\}, \qquad \forall\, \boldsymbol{\theta}_0, \boldsymbol{\theta} \in \Theta.$$

DEMOSTRACIÓN. En efecto,[2]

$$\begin{aligned}
\mathsf{E}_{\theta_0}\{\log f(\boldsymbol{x}; \boldsymbol{\theta}_0)\} - \mathsf{E}_{\theta_0}\{\log f(\boldsymbol{x}; \boldsymbol{\theta})\} &= -\mathsf{E}_{\theta_0}\left\{\log \frac{f(\boldsymbol{x}; \boldsymbol{\theta})}{f(\boldsymbol{x}; \boldsymbol{\theta}_0)}\right\} \\
&\geq -\log \mathsf{E}_{\theta_0}\left\{\frac{f(\boldsymbol{x}; \boldsymbol{\theta})}{f(\boldsymbol{x}; \boldsymbol{\theta}_0)}\right\} = \log \int \frac{f(\boldsymbol{x}; \boldsymbol{\theta})}{f(\boldsymbol{x}; \boldsymbol{\theta}_0)} f(\boldsymbol{x}; \boldsymbol{\theta}_0)\, \mathsf{d}\boldsymbol{x} \\
&= \log \int f(\boldsymbol{x}; \boldsymbol{\theta})\, \mathsf{d}\boldsymbol{x} = \log 1 = 0.
\end{aligned}$$

□

[2]Recuerde que, para h convexa $\mathsf{E}(h(Y)) \geq h(\mathsf{E}(Y))$.

2.3. 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 2.11 (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 2.12 (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,

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

RESULTADO 2.13. *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

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

□

EJEMPLO 2.14. 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 2.15 (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 2.16. 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 2.17. *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

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

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 2.18. 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 2.19. 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

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

RESULTADO 2.20. *Sean $\boldsymbol{X}$ e $\boldsymbol{Y}$ vectores aleatorios independientes con informaciones de Fisher $\boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta})$ y $\boldsymbol{\mathcal{F}}_Y(\boldsymbol{\theta})$, respectivamente. Entonces la información de Fisher contenida en $\boldsymbol{Z} = (\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$ es dada por*

$$\boldsymbol{\mathcal{F}}_Z(\boldsymbol{\theta}) = \boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta}) + \boldsymbol{\mathcal{F}}_Y(\boldsymbol{\theta}).$$

DEMOSTRACIÓN. Evidentemente, tenemos

$$\ell(\boldsymbol{\theta}; \boldsymbol{z}) = \log f_1(\boldsymbol{x}; \boldsymbol{\theta}) + \log f_2(\boldsymbol{x}; \boldsymbol{\theta}),$$

mientras que

$$\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{z}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_1(\boldsymbol{x}; \boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}} \log f_2(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}) + \boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{y}),$$

por la independencia entre $\boldsymbol{X}$ e $\boldsymbol{Y}$, sigue que

$$\begin{aligned}\boldsymbol{\mathcal{F}}_Z(\boldsymbol{\theta}) &= \mathsf{Cov}(\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{X}) + \boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{Y})) = \mathsf{Cov}(\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{X})) + \mathsf{Cov}(\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{Y})) \\ &= \boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta}) + \boldsymbol{\mathcal{F}}_Y(\boldsymbol{\theta}).\end{aligned}$$

□

COROLARIO 2.21. *Sea $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ copias IID de una variable aleatoria X con información de Fisher (común) $\boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta})$. Entonces, la información contenida en la muestra es:*

$$\boldsymbol{\mathcal{F}}_n(\boldsymbol{\theta}) = n\boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta}).$$

OBSERVACIÓN. Suponga $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ vectores aleatorios independientes (no necesariamente copias). Entonces,

$$\ell_n(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}; \boldsymbol{X}_i), \quad \boldsymbol{U}_n(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{i=1}^n \boldsymbol{U}_i(\boldsymbol{\theta}; \boldsymbol{X}_i), \quad \boldsymbol{\mathcal{F}}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \boldsymbol{\mathcal{F}}_i(\boldsymbol{\theta}),$$

con $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$.

RESULTADO 2.22 (Dependencia de la parametrización). *Sea X una variable aleatoria con distribución P_θ. Suponga otra parametrización dada por $\theta = h(\phi)$ con h diferenciable. Entonces, la información contenida en X con relación a ϕ es dada por*

$$\mathcal{F}_X^*(\phi) = \mathcal{F}_X(h(\phi))\{h'(\phi)\}^2,$$

donde $\mathcal{F}_X(\theta)$ es la información que $X \sim \mathsf{P}_\theta$ contiene respecto de θ y $\mathcal{F}_X^*(\phi)$ es la información que $X \sim \mathsf{P}_{h(\phi)}$ contiene sobre ϕ.

DEMOSTRACIÓN. La función de log-verosimilitud con relación a $\mathsf{P}_{h(\phi)}$ es:

$$\ell^*(\phi; x) = \log f(x; h(\phi)).$$

De este modo, la función score asume la forma:

$$U^*(\phi; X) = \frac{\partial}{\partial \phi} \log f(x; h(\phi)) = \frac{\partial}{\partial \theta} \log f(x; \theta)\Big|_{\theta = h(\phi)} h'(\phi),$$

luego,

$$\begin{aligned}\mathcal{F}_X^*(\phi) &= \mathsf{var}_\phi\{U^*(\phi;X)\} = \mathsf{var}_\phi\{U(h(\phi);X)\,h'(\phi)\}\\ &= \{h'(\phi)\}^2\,\mathsf{var}_\phi\{U(h(\phi);X)\} = \{h'(\phi)\}^2\mathcal{F}_X(h(\phi)).\end{aligned}$$

□

EJEMPLO 2.23. Considere la distribución $\mathsf{Poi}(\theta)$. Así,

$$f(x;\theta) = \exp(x\log\theta - \theta)/x!$$

Note que el parámetro natural es $\eta = \log\theta$. Luego,

$$f(x;\eta) = \exp(x\eta - e^\eta)/x!$$

En la parametrización original, tenemos que

$$U(\theta;x) = \frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{\partial}{\partial\theta}(x\log\theta - \theta) = \frac{x}{\theta} - 1.$$

De este modo,

$$\mathcal{F}_X(\theta) = \mathsf{var}_\theta\{U(\theta;X)\} = \mathsf{var}_\theta\left(\frac{X}{\theta} - 1\right) = \frac{1}{\theta^2}\,\mathsf{var}_\theta(X) = \frac{1}{\theta}.$$

Por otro lado, $\theta = e^\eta = h(\eta)$. Así,

$$\mathcal{F}_X^*(\eta) = \mathcal{F}_X(h(\eta))\{h'(\eta)\}^2 = \frac{1}{e^\eta}(e^\eta)^2 = e^\eta.$$

OBSERVACIÓN. Considere $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\phi})$ con $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^p$ $(p \leq k)$ y sea $\boldsymbol{H}(\boldsymbol{\phi}) = \partial\boldsymbol{h}(\boldsymbol{\phi})/\partial\boldsymbol{\phi}^\top$ matriz de rango completo. Entonces,

$$\boldsymbol{\mathcal{F}}^*(\boldsymbol{\phi}) = \boldsymbol{H}^\top(\boldsymbol{\phi})\boldsymbol{\mathcal{F}}(\boldsymbol{h}(\boldsymbol{\phi}))\boldsymbol{H}(\boldsymbol{\phi}).$$

Capítulo 3

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 3.1. 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 $\widehat{\boldsymbol{\gamma}}_{\mathsf{LS}}$.

3.1. Métodos de estimación

3.1.1. 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 3.2 (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 3.3. 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 3.4. 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 3.5. 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 3.6. 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 3.7. 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 3.8. 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

$$\begin{aligned} \theta_1 &= h_1(\mu_1, \ldots, \mu_p), \\ &\vdots \\ \theta_p &= h_p(\mu_1, \ldots, \mu_p). \end{aligned} \tag{3.1}$$

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

Finalmente, el estimador $\widehat{\boldsymbol{\theta}}_{\mathsf{MM}}$, puede ser obtenido substituyendo en (3.1) 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:

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \mathbf{0}, \tag{3.2}$$

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 (3.2). 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.

3.1.2. Estimación máximo verosímil. Este es uno de los procedimientos de estimación más ampliamente usados. Es motivado por el principio de verosimilitud (Lema 2.10) y los estimadores obtenidos disfrutan de buenas propiedades.

Definición 3.9 (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 3.10 (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 3.11. 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 3.12 (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 3.13 (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 3.14 (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 3.15 (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 3.16 (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 3.17 (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.

3.1.3. Implementación del método de máxima verosimilitud. Anteriormente hemos revisado ejemplos en los que ha sido posible obtener los estimadores ML de forma explícita. Sin embargo, en general es necesario recurrir a métodos numéricos, los que involucran la elección de una estimación inicial $\boldsymbol{\theta}^{(0)}$ y sucesivamente se construye la secuencia $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$, de manera de alcanzar convergencia a la solución $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$.

Con el objetivo de resolver el problema

$$\max_{\theta \in \Theta} \ell(\boldsymbol{\theta}),$$

suponga la expansión de Taylor en torno de $\boldsymbol{\theta}^*$, como:

$$\ell(\boldsymbol{\theta}^* + \boldsymbol{p}) = \ell(\boldsymbol{\theta}^*) + \Big(\frac{\partial \ell(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}\Big)^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \Big(\frac{\partial^2 \ell(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\Big)\boldsymbol{p} + o(\|\boldsymbol{p}\|^2),$$

donde $o(u)$ es un término de error de orden menor que u, conforme $u \to 0$, es decir,

$$\lim_{u \to 0} \frac{o(u)}{u} = 0.$$

Defina la función cuadrática,

$$q_k(\boldsymbol{p}) = \ell(\boldsymbol{\theta}^{(k)}) + \boldsymbol{U}^\top(\boldsymbol{\theta}^{(k)})\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{H}(\boldsymbol{\theta}^{(k)})\boldsymbol{p},$$

donde $\boldsymbol{U}(\boldsymbol{\theta}) = \partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ y $\boldsymbol{H}(\boldsymbol{\theta}) = \partial^2 \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top$. Minimizando $q_k(\boldsymbol{p})$ con relación a $\boldsymbol{p}$, lleva al sistema de ecuaciones $\partial q_k(\boldsymbol{p})/\partial \boldsymbol{p} = \boldsymbol{0}$. Es decir, obtenemos:

$$\boldsymbol{H}(\boldsymbol{\theta}^{(k)})\boldsymbol{p} = -\boldsymbol{U}(\boldsymbol{\theta}^{(k)}). \tag{3.3}$$

Métodos *tipo-Newton* adoptan la forma:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda_k \boldsymbol{p}_k, \qquad k = 0, 1, \ldots,$$

donde $\boldsymbol{p}_k$ es la dirección de búsqueda dada por la solución del sistema dado en (3.3), mientras que λ_k es un largo de paso que debe ser escogido para garantizar que

$$\ell(\boldsymbol{\theta}^{(k)} + \lambda_k \boldsymbol{p}_k) \geq \ell(\boldsymbol{\theta}^{(k)}).$$

Es fácil notar que la dirección dada por (3.3), satisface

$$\boldsymbol{U}^\top(\boldsymbol{\theta}^{(k)})\boldsymbol{p}_k = \boldsymbol{U}^\top(\boldsymbol{\theta}^{(k)})\{-\boldsymbol{H}(\boldsymbol{\theta}^{(k)})\}^{-1}\boldsymbol{U}(\boldsymbol{\theta}^{(k)}) > 0,$$

es decir, corresponde a una dirección de ascenso. De esta manera el procedimiento para determinar una aproximación de la solución asume la forma:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda_k\{-\boldsymbol{H}(\boldsymbol{\theta}^{(k)})\}^{-1}\boldsymbol{U}(\boldsymbol{\theta}^{(k)}), \qquad k = 0, 1, \ldots,$$

que es conocido como *método de Newton-Raphson*. Es conocido que este procedimiento converge rápidamente siempre que un valor inicial cercano al óptimo haya sido escogido. En casos en que sea difícil llevar a cabo el cálculo de la matriz Hessiana podemos utilizar un *método quasi-Newton*, el cual no involucra información de segundo orden sino que utiliza la dirección de búsqueda:

$$\boldsymbol{p}_k = \{-\boldsymbol{B}_k\}^{-1}\boldsymbol{U}(\boldsymbol{\theta}^{(k)}),$$

donde $\boldsymbol{B}_k$ es una aproximación de la matriz Hessiana $\boldsymbol{H}(\boldsymbol{\theta}^{(k)})$. Sea $\boldsymbol{g}_k = \boldsymbol{U}(\boldsymbol{\theta}^{(k+1)}) - \boldsymbol{U}(\boldsymbol{\theta}^{(k)})$, esta clase de métodos actualiza una estimación de $\{\boldsymbol{H}(\boldsymbol{\theta}^{(k)})\}^{-1}$ usando un método secante, tal como el algoritmo de Davidon-Fletcher-Powell (DFP)

$$\boldsymbol{B}_{k+1}^{-1} = \boldsymbol{B}_k^{-1} + \frac{\boldsymbol{s}_k \boldsymbol{s}_k^\top}{\boldsymbol{s}_k^\top \boldsymbol{g}_k} - \frac{\boldsymbol{B}_k^{-1} \boldsymbol{g}_k \boldsymbol{g}_k^\top \boldsymbol{B}_k^{-1}}{\boldsymbol{g}_k^\top \boldsymbol{B}_k^{-1} \boldsymbol{g}_k},$$

o bien, el método de Broyden-Fletcher-Goldfarb-Shanno (BFGS)

$$\boldsymbol{B}_{k+1}^{-1} = \boldsymbol{B}_k^{-1} + \frac{\boldsymbol{s}_k \boldsymbol{g}_k^\top \boldsymbol{B}_k^{-1}}{\boldsymbol{s}_k^\top \boldsymbol{g}_k} - \frac{\boldsymbol{B}_k^{-1} \boldsymbol{g}_k \boldsymbol{s}_k^\top}{\boldsymbol{s}_k^\top \boldsymbol{g}_k} + \left\{1 + \frac{\boldsymbol{g}_k^\top \boldsymbol{B}_k^{-1} \boldsymbol{g}_k}{\boldsymbol{s}_k^\top \boldsymbol{g}_k}\right\} \frac{\boldsymbol{s}_k \boldsymbol{s}_k^\top}{\boldsymbol{s}_k^\top \boldsymbol{g}_k},$$

donde $\boldsymbol{s}_k = \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}$. Es frecuente usar como valor inicial para $\boldsymbol{B}_0^{-1}$ la matriz identidad. En situaciones en que también es difícil calcular el vector score $\boldsymbol{U}(\boldsymbol{\theta})$ es posible determinar la información de primer orden usando diferenciación numérica.

Un procedimiento popular en estadística corresponde al *método Fisher-scoring*, que corresponde a un algoritmo en la clase quasi-Newton donde $-\boldsymbol{H}(\boldsymbol{\theta})$ es aproximada mediante la matriz de información de Fisher. De este modo, obtenemos el siguiente esquema iterativo:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda_k \boldsymbol{\mathcal{F}}^{-1}(\boldsymbol{\theta}^{(k)}) \boldsymbol{U}(\boldsymbol{\theta}^{(k)}), \qquad k = 0, 1, \ldots,$$

donde $\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{E}\{-\boldsymbol{H}(\boldsymbol{\theta})\}$.

Cuando tenemos $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ observaciones (vectores) independientes, tenemos que

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

De este modo, $\boldsymbol{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \boldsymbol{U}_i(\boldsymbol{\theta})$, con $\boldsymbol{U}_i(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}_i; \boldsymbol{\theta})$. Notando que

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}) \boldsymbol{U}_i^\top(\boldsymbol{\theta}) \xrightarrow{\mathsf{P}} \mathsf{E}\{\boldsymbol{U}_n(\boldsymbol{\theta}) \boldsymbol{U}_n^\top(\boldsymbol{\theta})\},$$

conforme $n \to \infty$, lleva al *Algoritmo BHHH* (Berndt et al., 1974), definido como:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda_k \left\{ \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}^{(k)}) \boldsymbol{U}_i^\top(\boldsymbol{\theta}^{(k)}) \right\}^{-1} \boldsymbol{U}_n(\boldsymbol{\theta}^{(k)}), \qquad k = 0, 1, \ldots.$$

Usualmente, llevamos a cabo la iteración del procedimiento de estimación hasta que

$$\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\| < \tau,$$

donde τ, conocido como tolerancia, sea pequeño (por ejemplo, $\tau = 10^{-6}$). Otra alternativa es considerar el criterio

$$(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)})^\top \boldsymbol{\mathcal{F}}(\boldsymbol{\theta}^{(k)}) (\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}) < \delta,$$

donde δ representa un valor de tolerancia. Por ejemplo, basados en la idea de un *elipsoide de confianza*, Bates y Watts (1981) propusieron un criterio de convergencia basado en un *offset relativo*. Tal idea puede ser aplicada a diversos modelos estadísticos. En efecto, ha sido implementada para los modelos disponibles en las bibliotecas para R `nlme` (Pinheiro et al., 2019) y `heavy` (Osorio, 2019).

EJEMPLO 3.18 (distribución Cauchy). Suponga $X_1, \ldots, X_n$ variables aleatorias desde la distribución $\mathsf{Cauchy}(\theta, 1)$, con densidad

$$f(x;\theta) = \frac{1}{\phi\{1+(x-\theta)^2\}}, \qquad x \in \mathbb{R}, \theta \in \mathbb{R}.$$

Así, la función de log-verosimilitud adopta la forma:

$$\begin{aligned}\ell(\theta;\boldsymbol{x}) &= -n\log\pi - \log\prod_{i=1}^{n}\{1+(x_i-\theta)^2\} \\ &= -n\log\pi - \sum_{i=1}^{n}\log(1+(x_i-\theta)^2).\end{aligned}$$

Calculando la primera derivada, obtenemos

$$U(\theta;\boldsymbol{x}) = \frac{\partial\ell(\theta;\boldsymbol{x})}{\partial\theta} = \sum_{i=1}^{n}\frac{2(x_i-\theta)}{1+(x_i-\theta)^2}.$$

Por tanto, el estimador máximo verosímil debe satifacer la condición de primer-orden:

$$U(\theta;\boldsymbol{x}) = \sum_{i=1}^{n}\frac{2}{1+(x_i-\theta)^2}(x_i-\theta) = \sum_{i=1}^{n}\omega_i(\theta)(x_i-\theta) = 0, \tag{3.4}$$

donde $\omega_i(\theta) = 2/(1+(x_i-\theta)^2)$. Es fácil notar que la Ecuación (3.4) no tiene solución explícita. Para aplicar el algoritmo Newton-Raphson, calculamos

$$\frac{\partial}{\partial\theta}U(\theta;\boldsymbol{x}) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\omega_i(\theta)(x_i-\theta) + \sum_{i=1}^{n}\omega_i(\theta)\frac{\partial}{\partial\theta}(x_i-\theta),$$

como

$$\frac{\partial}{\partial\theta}\omega_i(\theta) = \frac{4}{\{1+(x_i-\theta)^2\}^2}(x_i-\theta) = \omega_i^2(\theta)(x_i-\theta),$$

esto lleva a

$$\frac{\partial}{\partial\theta}U(\theta;\boldsymbol{x}) = \sum_{i=1}^{n}\omega_i^2(\theta)(x_i-\theta)^2 - \sum_{i=1}^{n}\omega_i(\theta).$$

De este modo,

$$-H(\theta;\boldsymbol{x}) = -\frac{\partial}{\partial\theta}U(\theta;\boldsymbol{x}) = \sum_{i=1}^{n}\omega_i(\theta)\{1-\omega_i(\theta)(x_i-\theta)^2\}.$$

Finalmente el método Newton-Raphson, adopta la forma:

$$\begin{aligned}\theta^{(k+1)} &= \theta^{(k)} - \frac{U(\theta^{(k)};\boldsymbol{x})}{H(\theta^{(k)};\boldsymbol{x})} \\ &= \theta^{(k)} + \frac{\sum_{i=1}^{n}\omega_i(\theta^{(k)})(x_i-\theta^{(k)})}{\sum_{i=1}^{n}\omega_i(\theta^{(k)})\{1-\omega_i(\theta^{(k)})(x_i-\theta^{(k)})^2\}}.\end{aligned} \tag{3.5}$$

OBSERVACIÓN. En la biblioteca heavy se encuentra una alternativa al esquema iterativo en (3.5) que utiliza un *Algoritmo EM.*[3]

[3]El Algoritmo EM (o de Esperanza-Maximización) permite obtener los estimadores ML en presencia de *datos perdidos.*

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 3.19 (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 $\widehat{\lambda}_{\mathsf{ML}} = e^{\widehat{\phi}_{\mathsf{ML}}}$.

3.2. 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 3.20 (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 3.21 (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 3.22 (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 3.23. 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,[4]

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

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

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

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

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

[4]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 3.24. 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

$$\operatorname{tr} \mathsf{MSE}(\boldsymbol{T}_*, \boldsymbol{\theta}) \leq \operatorname{tr} \mathsf{MSE}(\boldsymbol{T}, \boldsymbol{\theta}), \tag{3.6}$$

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

$$\det \mathsf{MSE}(\boldsymbol{T}_*, \boldsymbol{\theta}) \leq \det \mathsf{MSE}(\boldsymbol{T}, \boldsymbol{\theta}). \tag{3.7}$$

El criterio dado en la Definición 3.24 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.

Una alternativa es restringirse a alguna subclase de estimadores. De este modo, nos concentraremos en la clase de estimadores insesgados.

DEFINICIÓN 3.25 (Mejor estimador insesgado). Para el modelo estadístico $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$, un estimador insesgado T_* para $\gamma = g(\theta) \in \mathbb{R}$ se dice el mejor estimador insesgado (BUE), si para cualquier otro estimador insesgado

$$\mathsf{var}_\theta(T_*) \leq \mathsf{var}_\theta(T), \qquad \forall\, \theta \in \Theta.$$

RESULTADO 3.26 (Cota de Cramér-Rao). *Suponga que se satisfacen las condiciones A1-A4, que la información de Fisher es tal que* $0 < \mathcal{F}_X(\theta) < \infty$ *y sea* $\gamma = g(\theta)$, *donde* g *es continua y diferenciable con* $g' \neq 0$. *Si* T *es estimador insesgado para* γ, *entonces*

$$\mathsf{var}_\theta(T) \geq \frac{\{g'(\theta)\}^2}{\mathcal{F}_X(\theta)}, \qquad \forall\, \theta \in \Theta.$$

DEMOSTRACIÓN. Considere

$$\begin{aligned}\mathsf{Cov}_\theta(T(\boldsymbol{X}), U(\theta;\boldsymbol{X})) &= \mathsf{E}_\theta(T(\boldsymbol{X})\, U(\theta;\boldsymbol{X})) = \int_A T(\boldsymbol{x}) \frac{\partial}{\partial\theta} \log f(\boldsymbol{x};\theta)\, f(\boldsymbol{x};\theta)\, \mathsf{d}\boldsymbol{x} \\ &= \int_A T(\boldsymbol{x}) \frac{\partial}{\partial\theta} f(\boldsymbol{x};\theta)\, \mathsf{d}\boldsymbol{x}.\end{aligned}$$

Como T es regular[5] y es insesgada, sigue que

$$\mathsf{Cov}_\theta(T(\boldsymbol{X}), U(\theta;\boldsymbol{X})) = \frac{\partial}{\partial\theta} \int_A T(\boldsymbol{x})\, f(\boldsymbol{x};\theta)\, \mathsf{d}\boldsymbol{x} = \frac{\partial}{\partial\theta}\, \mathsf{E}_\theta(T(\boldsymbol{X})) = g'(\theta).$$

Haciendo $h(\theta) = g'(\theta)/\mathcal{F}_X(\theta)$, tenemos

$$\begin{aligned}0 &\leq \mathsf{var}_\theta(T(\boldsymbol{X}) - h(\theta)U(\theta;\boldsymbol{X})) \\ &= \mathsf{var}_\theta(T(\boldsymbol{X})) + h^2(\theta)\, \mathsf{var}_\theta(U(\theta;\boldsymbol{X})) - 2h(\theta)\, \mathsf{Cov}_\theta(T(\boldsymbol{X}), U(\theta;\boldsymbol{X})) \\ &= \mathsf{var}_\theta(T(\boldsymbol{X})) + h^2(\theta)\mathcal{F}_X(\theta) - 2h(\theta)g'(\theta),\end{aligned}$$

es decir,

$$0 \leq \mathsf{var}_\theta(T(\boldsymbol{X})) - \{g'(\theta)\}^2/\mathcal{F}_X(\theta).$$

□

DEFINICIÓN 3.27 (Eficiencia). La eficiencia de un estimador insesgado T es definida como la razón de su varianza y la cota de Cramér-Rao. Esto es,

$$\mathsf{EFF}(T,\theta) = \frac{\{g'(\theta)\}^2/\mathcal{F}_X(\theta)}{\mathsf{var}_\theta(T)}.$$

Un estimador que alcanza la cota de Cramér-Rao se dice un *estimador eficiente*. Más aún, un estimador eficiente es BUE.

EJEMPLO 3.28. Considere $X_1, \ldots, X_n$ variables aleatorias IID desde $\mathsf{Exp}(\lambda)$ con $\lambda > 0$. Sea $\gamma = 1/\lambda$ el parámetro de interés. Un estimador insesgado para λ es $\overline{X}$ con varianza $\frac{1}{n\lambda^2}$. Como la información de Fisher es n/λ^2 y $g'(\lambda) = -1/\lambda^2$, tenemos

$$\frac{\{-1/\lambda^2\}^2}{n/\lambda^2} = \frac{1}{n\lambda^2},$$

es decir, $\overline{X}$ es eficiente.

[5] Es decir, podemos intercambiar las operaciones de integración y diferenciación.

Considere el caso en que $\Theta \subset \mathbb{R}^k$ y que el parámetro de interés es $\boldsymbol{\gamma} \in \mathbb{R}^m$ con $\boldsymbol{\gamma} = \boldsymbol{g}(\boldsymbol{\theta})$. Sea $\boldsymbol{T}$ y $\boldsymbol{T}_*$ dos estimadores insesgados de $\boldsymbol{\gamma}$. Decimos que $\boldsymbol{T}_*$ tiene covarianza más pequeña que $\boldsymbol{T}$, si

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

en cuyo caso escribimos

$$\mathsf{Cov}_\theta(\boldsymbol{T}_*) \leq \mathsf{Cov}_\theta(\boldsymbol{T}). \tag{3.8}$$

Evidentemente, Ecuación (3.8) corresponde a un caso particular de la Definición 3.24 para el caso de estimadores insesgados. Esto permite extender la cota de Cramér-Rao para el caso multiparamétrico.

Suponga que las condiciones A1 a A4 son satisfechas y que la matriz de información de Fisher es no singular. Entonces la cota de Cramér-Rao asume la forma:

$$\mathsf{Cov}_\theta(\boldsymbol{T}) \geq \Big(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big)\boldsymbol{\mathcal{F}}_X^{-1}(\boldsymbol{\theta})\Big(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big)^\top, \qquad \forall\, \boldsymbol{\theta} \in \Theta.$$

Ejemplo 3.29. Suponga que tenemos $X_1, \ldots, X_n$ variables aleatorias independientes desde $\mathsf{N}(\mu, \sigma^2)$ con parámetro de interés $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. La matrix de información de Fisher está dada por

$$\boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

cuya matriz inversa corresponde a la cota de Cramér-Rao. Sabemos que los estimadores insesgados $\overline{X}$ y S^2 son independientes y que

$$\mathsf{var}_\theta(\overline{X}) = \frac{\sigma^2}{n}, \qquad \mathsf{var}_\theta(S^2) = \frac{2\sigma^4}{n-1}.$$

Por tanto no se alcanza la cota inferior para la varianza.

Resultado 3.30. *Bajo las condiciones A1 a A4 el estimador de máxima verosimilitud satisface las siguientes propiedades:*

i) El estimador ML depende de los datos via la estadística suficiente.

ii) Si existe un estimador insesgado y eficiente $\widetilde{\boldsymbol{\theta}}$, entonces $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$.

Demostración. *i)* sigue notando que, por el Teorema de factorización de Fisher-Neyman

$$L(\boldsymbol{\theta}; \boldsymbol{x}) \propto g(\boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\theta}),$$

para $\boldsymbol{T}$ suficiente.

Por simplicidad para la prueba de *ii)* sólo consideraremos el caso en que θ es real-valuado. Sabemos que $\widetilde{\theta}$ alcanza la cota de Cramér-Rao pues es un estimador eficiente. Así, tenemos que

$$\widetilde{\theta}(\boldsymbol{x}) - \theta = \frac{U(\theta; \boldsymbol{x})}{\mathcal{F}_X(\theta)}, \qquad \forall\, \theta \in \Theta,$$

que es válido en particular para $\theta = \widehat{\theta}_{\mathsf{ML}}$. Es decir,

$$\widetilde{\theta}(\boldsymbol{x}) - \widehat{\theta}_{\mathsf{ML}} = \frac{U(\widehat{\theta}_{\mathsf{ML}}; \boldsymbol{x})}{\mathcal{F}_X(\widehat{\theta}_{\mathsf{ML}})}.$$

Como $\widehat{\theta}_{\mathsf{ML}}$ maximiza $\ell(\theta; \boldsymbol{x})$ y por tanto, $U(\widehat{\theta}_{\mathsf{ML}}; \boldsymbol{x}) = 0$. De este modo,

$$\widetilde{\theta}(\boldsymbol{x}) - \widehat{\theta}_{\mathsf{ML}} = 0.$$

□

3.3. Propiedades Asintóticas

Parece razonable que conforme el tamaño muestral crece mayor confianza tendremos en nuestras inferencias, debido a que la muestra contendrá más información con respecto a la distribución subyacente. Para caracterizar las propiedades asintóticas de los estimadores se introducirá algunas nociones de convergencia de variableas aleatorias.

DEFINICIÓN 3.31 (Consistencia). Decimos que una secuencia de estimadores $\{T_n\}$ para el parámetro $\gamma = g(\theta)$ es *débilmente consistente*, si T_n converge en probabilidad a γ, esto es, si para cualquier $\epsilon > 0$ y para todo $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathsf{P}_\theta(|T_n - g(\theta)| > \epsilon) = 0,$$

y escribimos $T_n \xrightarrow{\mathsf{P}} \gamma$. Mientras que, se dice que $\{T_n\}$ converge con probabilidad 1 o casi seguramente (a.s.) a γ, para todo $\theta \in \Theta$, esto es,

$$\mathsf{P}_\theta\left(\lim_{n\to\infty} T_n = g(\theta)\right) = 1,$$

es decir T_n es *consistente fuerte*, en cuyo caso anotamos $T_n \xrightarrow{\text{a.s.}} \gamma$.

Consistencia muy frecuentemente es una consecuencia de la ley de los grandes números. Es tipo más simple es la convergencia de la media muestral. Por el teorema de mapeo continuo es posible obtener la consistencia de otros estimadores.

TEOREMA 3.32 (Teorema del mapeo continuo). *Sea $\{S_n\}$ una secuencia de variables aleatorias, S_0 una variable aletoria y h una función continua.*

i) Si $S_n \xrightarrow{\mathsf{P}} S_0$, entonces

$$h(S_n) \xrightarrow{\mathsf{P}} h(S_0).$$

ii) Si $S_n \xrightarrow{\text{a.s.}} S_0$, entonces

$$h(S_n) \xrightarrow{\text{a.s.}} h(S_0).$$

Una herramienta importante para la verificación de consistencia es la desigualdad de Chebyshev. Para una variable aleatoria Z con media finita, tenemos

$$\mathsf{P}(|Z - \mathsf{E}(Z)| > \tau) \leq \frac{\mathsf{var}(Z)}{\tau^2}.$$

EJEMPLO 3.33. Sea $\{X_n\}$ una secuencia de variables aleatorias IID con función de distribución F. Entonces se tiene:

1. La media aritmética converge a $\mathsf{E}_F(X) = \mu$, es decir, $\overline{X}_n \xrightarrow{\mathsf{P}} \mu$. En efecto, usando la desigualdad de Chebyshev y asumiendo $\sigma < \infty$, tenemos
$$\mathsf{P}(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\mathsf{var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0,$$
para $n \to \infty$.
2. La varianza empírica y la desviación estándar convergen a $\mathsf{var}_F(X) = \sigma^2$ y σ, respectivamente:
$$S_n^2 \xrightarrow{\mathsf{P}} \sigma^2, \qquad S_n \xrightarrow{\mathsf{P}} \sigma.$$

3. La frecuencia relativa de un evento A converge a su probabilidad. Sea

$$Q_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A),$$

donde $I(\cdot)$ denota la función indicadora. Entonces $Q_n(A) \xrightarrow{\mathsf{P}} \mathsf{P}(A)$.

4. Si el j-ésimo momento $\mu_j = \mathsf{E}_F(X^j)$ existe, entonces los momentos empíricos

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j,$$

son consistentes. Es decir, $m_j \xrightarrow{\mathsf{P}} \mu_j$. De ahí que, el estimador de momentos para $\gamma = h(\mu_1, \ldots, \mu_r)$ es un estimador consistente si la función h es contínua.

DEFINICIÓN 3.34 (Convergencia en distribución). Sea $\{X_n\}$ una secuencia de variables aleatorias y sea X otra variable aleatoria. Además, considere F_n la CDF de X_n y F la CDF de X. Se dice que X_n *converge en distribución* a X, en cuyo caso escribimos $X_n \xrightarrow{\mathsf{D}} X$ si

$$\lim_{n\to\infty} F_n(t) = F(t),$$

para todo t donde F es continua.

TEOREMA 3.35. *Sea $\{Z_n\}$ una secuencia de variables aleatorias y sea $M_n(t)$ la MGF de Z_n. Sea Z una variable aleatoria con MGF dada por $M(t)$. Si*

$$M_n(t) \to M(t), \qquad \text{para todo } |t| < h,\ h > 0.$$

Entonces, $Z_n \xrightarrow{\mathsf{D}} Z$.

TEOREMA 3.36 (Teorema de Slutsky). *Considere dos secuencias de variables aleatorias $\{X_n\}$, $\{Y_n\}$, una variable aleatoria X y una constante fija c. Suponga que $X_n \xrightarrow{\mathsf{D}} X$, y $Y_n \xrightarrow{\mathsf{P}} c$. Entonces:*

i) $X_n \pm Y_n \xrightarrow{\mathsf{D}} X \pm c$.
ii) $X_n Y_n \xrightarrow{\mathsf{D}} cX$.
iii) $X_n Y_n^{-1} \xrightarrow{\mathsf{D}} c^{-1}X$ *siempre que* $\mathsf{P}(Y_n = 0) = 0$ *para todo* n *y* $c \neq 0$.

DEFINICIÓN 3.37 (Normalidad asintótica). Una secuencia de estimadores $\{\boldsymbol{T}_n\}$ para el parámetro m-dimensional $\boldsymbol{\gamma} = \boldsymbol{g}(\boldsymbol{\theta})$ es *asintóticamente normal* si para todo $\boldsymbol{\theta} \in \Theta$ la distribución de $\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{g}(\boldsymbol{\theta}))$ converge a una distribución normal con media cero y matriz de covarianza $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Es decir,

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{\mathsf{D}} \mathsf{N}_m(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

La normalidad asintótica permite establecer si un estimador es *asintóticamente eficiente*. En efecto, diremos que un estimador es asintóticamente eficiente si este es asintóticamente normal, con

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \Big(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big)\boldsymbol{\mathcal{F}}_X^{-1}(\boldsymbol{\theta})\Big(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big)^\top,$$

donde $\boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta})$ es la matriz de información de la distribución subyacente.

RESULTADO 3.38 (Método Delta). *Suponga $\boldsymbol{T}_n$ un estimador de la forma $\boldsymbol{T}_n = \boldsymbol{h}(\boldsymbol{S}_n)$ donde la secuencia $\{\boldsymbol{S}_n\}$ es asintóticamente normal, esto es,*

$$\sqrt{n}(\boldsymbol{S}_n - \boldsymbol{\mu}) \xrightarrow{\mathsf{D}} \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

para $\boldsymbol{\mu} \in \mathbb{R}^k$ y $\boldsymbol{\Sigma} > 0$. Si $\partial \boldsymbol{h}(\boldsymbol{\mu})/\partial \boldsymbol{\mu}^\top$ es matriz de rango completo, entonces

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{h}(\boldsymbol{\mu})) \xrightarrow{\mathsf{D}} \mathsf{N}\Big(\mathbf{0}, \Big(\frac{\partial \boldsymbol{h}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^\top}\Big)\boldsymbol{\Sigma}\Big(\frac{\partial \boldsymbol{h}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^\top}\Big)^\top\Big).$$

EJEMPLO 3.39. Sean $X_1, \ldots, X_n$ variables aleatorias IID con $\mathsf{E}(X_i) = \mu \neq 0$ y $\mathsf{var}(X_i) = \sigma^2 < \infty$. El parámetro $\gamma = \log \mu$ es estimado por $\widehat{\gamma}_n = \log \overline{X}_n$. Este estimador es consistente y asintóticamente normal. En efecto, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \sigma^2)$. Como $h(s) = \log s$ y $h'(s) = 1/s$, sigue que

$$\sqrt{n}(\log \overline{X}_n - \log \mu) \xrightarrow{\mathsf{D}} \mathsf{N}\Big(0, \frac{\sigma^2}{\mu^2}\Big).$$

3.3.1. Distribución asintótica de los estimadores ML. Considere una secuencia de estimadores $\{\widehat{\boldsymbol{\theta}}_n\}$ que converge en probabilidad a $\boldsymbol{\theta}_0$ perteneciendo al interior de Θ, y suponga que

SUPUESTO A5. *La matriz*

$$\boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta}_0) = \mathsf{E}_{\theta_0}\Big\{ - \frac{\partial^2 \log f(\boldsymbol{y}_1; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\Big\},$$

existe y es no singular.

RESULTADO 3.40. *Bajo las condiciones A1-A5 una secuencia de máximos locales de $\ell_n(\boldsymbol{\theta})$ tiene distribución asintótica*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_k(\mathbf{0}, \boldsymbol{\mathcal{F}}_1^{-1}(\boldsymbol{\theta}_0)).$$

DEMOSTRACIÓN. Note que la secuencia $\{\widehat{\boldsymbol{\theta}}_n\}$ satisface las ecuaciones de verosimilitud, $\partial \ell_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$. Usando una expansión de Taylor del vector score en torno de $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, resulta[6]

$$\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o_{\mathsf{P}}(\mathbf{1}),$$

haciendo $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$, sigue que

$$\mathbf{0} = \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_{\mathsf{P}}(\mathbf{1}),$$

es decir

$$-\frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + o_{\mathsf{P}}(\mathbf{1}),$$

o equivalentemente,

$$\Big(-\frac{1}{n}\frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\Big)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + o_{\mathsf{P}}(\mathbf{1}).$$

[6] Si $Z_n \xrightarrow{\mathsf{P}} 0$ entonces anotamos $Z_n = o_{\mathsf{P}}(1)$.

Por otro lado, tenemos que

$$-\frac{1}{n}\frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2 \log f(\boldsymbol{Y}_i;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}$$

converge casi seguramente a

$$\boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta}_0) = \mathsf{E}_{\theta_0}\Big(-\frac{\partial^2 \log f(\boldsymbol{Y}_i;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\Big),$$

debido a la ley fuerte de los grandes números. Lo anterior lleva a,

$$\boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + o_{\mathsf{P}}(\mathbf{1}),$$

Note también que,

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(\boldsymbol{Y}_i;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{\partial \log f(\boldsymbol{Y}_i;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} - \mathsf{E}_{\theta_0}\Big(\frac{\partial \log f(\boldsymbol{Y}_i;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\Big)\Big\},\end{aligned}$$

converge en distribución a

$$\mathsf{N}_k(\mathbf{0}, \mathsf{Cov}_{\theta_0}(\boldsymbol{U}_i(\boldsymbol{\theta}_0))) \overset{\mathsf{d}}{=} \mathsf{N}_k(\mathbf{0}, \boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta}_0)).$$

Premultiplicando por $\boldsymbol{\mathcal{F}}_1^{-1}(\boldsymbol{\theta}_0)$ sigue que

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\boldsymbol{\mathcal{F}}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{U}_n(\boldsymbol{\theta}_0) + o_{\mathsf{P}}(\mathbf{1}).$$

Usando que

$$\frac{1}{\sqrt{n}}\boldsymbol{\mathcal{F}}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{U}_n(\boldsymbol{\theta}_0) \overset{\mathsf{D}}{\to} \mathsf{N}_k(\mathbf{0}, \boldsymbol{\mathcal{F}}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{F}}_1^{-1}(\boldsymbol{\theta}_0)),$$

sigue el resultado deseado. □

El resultado anterior es de gran importancia, pues bajo condiciones de regularidad el MLE es consistente, asintóticamente normal y eficiente, es decir es BUE. Además, esto permite el desarrollo de intervalos de confianza y test de hipótesis asintóticos.

Capítulo 4

Intervalos y Regiones de Confianza

El objetivo de esta sección es abordar el problema $\boldsymbol{\theta} \in C$, donde $C \subseteq \Theta$, $C = C(\boldsymbol{X})$ es un conjunto determinado por los datos observados $\boldsymbol{X} = \boldsymbol{x}$.

Observación. Si θ es real-valuado, entonces C corresponde a un intervalo.

Definición 4.1. Una estimación intervalar de un parámetro real-valuado θ es cualquier par de funciones $L(x_1, \ldots, x_n)$ y $U(x_1, \ldots, x_n)$ que satisfacen

$$L(\boldsymbol{x}) \leq U(\boldsymbol{x}), \qquad \forall\, \boldsymbol{x} \in \mathcal{X}.$$

Para $\boldsymbol{X} = \boldsymbol{x}$ tenemos $L(\boldsymbol{x}) \leq \theta \leq U(\boldsymbol{x})$, mientras que $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ es un intervalo aleatorio.

Aunque estamos interesados en intervalos de la forma $[L(\boldsymbol{x}), U(\boldsymbol{x})]$ también podemos tener $(-\infty, U(\boldsymbol{x})]$ o bien $[L(\boldsymbol{x}), \infty)$ para indicar $\theta \leq U(\boldsymbol{x})$ o $L(\boldsymbol{x}) \leq \theta$, respectivamente.

Ejemplo 4.2. Considere X_1, X_2, X_3, X_4 una muestra aleatoria desde $\mathsf{N}(\mu, 1)$. Un estimador intervalar de μ es $[\overline{X} - 1, \overline{X} + 1]$, es decir

$$\mu \in [\overline{X} - 1, \overline{X} + 1]$$

Note que $\overline{X} \sim \mathsf{N}(\mu, 1/4)$, pero

$$\mathsf{P}(\overline{X} = \mu) = 0$$

De ahí que con un estimador intervalar una probabilidad no nula de estar en lo correcto, en efecto,

$$\begin{aligned}
\mathsf{P}(\mu \in [\overline{X} - 1, \overline{X} + 1]) &= \mathsf{P}(\overline{X} - 1 \leq \mu \leq \overline{X} + 1) = \mathsf{P}(-1 \leq \mu - \overline{X} \leq 1) \\
&= \mathsf{P}(-1 \leq \overline{X} - \mu \leq 1) = \mathsf{P}\Big(- \frac{1}{\sqrt{1/4}} \leq \frac{\overline{X} - \mu}{\sqrt{1/4}} \leq \frac{1}{\sqrt{1/4}} \Big) \\
&= \mathsf{P}\Big(-2 \leq \frac{\overline{X} - \mu}{\sqrt{1/4}} \leq 2 \Big) = \mathsf{P}(-2 \leq Z \leq 2) = 0.9544
\end{aligned}$$

en este caso $Z = (\overline{X} - \mu)/\sqrt{1/4} \sim \mathsf{N}(0, 1)$.

Interpretación. De este modo, tenemos un 95 % de chances de cubrir el parámetro verdadero (desconocido) con nuestro estimador intervalar.

Observación. En este contexto $\mathsf{P}_\theta(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$ se denomina *probabilidad de cobertura*

Definición 4.3. El *coeficiente de confianza* de $[L(\boldsymbol{x}), U(\boldsymbol{x})]$ es el ínfimo de las probabilidades de cobertura

$$\inf_{\theta}\ \mathsf{P}_\theta(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$$

Observación. Estimadores intervalares en conjunto con una medida de confianza (coeficiente de confianza) son conocidos como *intervalos de confianza*.

A continuación consideraremos dos de los procedimientos más utilizados para construir intervalos de confianza.

4.1. Método de la Cantidad Pivotal

Definición 4.4. Una variable aleatoria $Q(\boldsymbol{X};\boldsymbol{\theta}) = Q(X_1,\ldots,X_n;\theta)$ es una *cantidad pivotal* o pivote si la distribución de $Q(\boldsymbol{X};\theta)$ **no** depende de θ. Esto es, si $\boldsymbol{X} \sim F(\boldsymbol{x};\theta)$, entonces $Q(\boldsymbol{X};\theta)$ tiene la misma distribución para todo valor de θ.

Observación. La técnica confía en la habilidad de hallar un pivote y un conjunto A tal que el conjunto $\{\theta : Q(\boldsymbol{X};\theta) \in A\}$ sea una estimación intervalar para θ.

Ejemplo 4.5. Si $X_1,\ldots,X_n$ es una muestra aleatoria de tamaño n desde $\mathsf{N}(\mu,\sigma^2)$, entonces

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1),$$

y por tanto es un pivote para μ (siempre que σ^2 sea conocido). Para cualquier constante a sigue que:

$$\begin{aligned}\mathsf{P}\Big(-a \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le a\Big) &= \mathsf{P}\Big(-a\frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le a\frac{\sigma}{\sqrt{n}}\Big)\\ &= \mathsf{P}\Big(\overline{X}-a\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+a\frac{\sigma}{\sqrt{n}}\Big),\end{aligned}$$

es decir obtenemos el intervalo de confianza

$$\Big[\overline{X}-a\frac{\sigma}{\sqrt{n}};\overline{X}+a\frac{\sigma}{\sqrt{n}}\Big],$$

o bien

$$\Big\{\mu : \overline{X}-a\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+a\frac{\sigma}{\sqrt{n}}\Big\}.$$

Además suponga que $a = z_{1-\alpha/2}$ para un valor de α dado. Entonces, es fácil notar que

$$\mathsf{P}\Big(\mu \in \Big[\overline{X}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}};\overline{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big]\Big) = 1-\alpha,$$

corresponde a un intervalo de confianza del $100(1-\alpha)\,\%$ para μ.

Para el caso en que σ^2 sea desconocido podemos usar el pivote

$$T = \frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t(n-1),$$

es decir,

$$\mathsf{P}(-a \le T \le a) = \mathsf{P}\Big(-a \le \frac{\overline{X}-\mu}{s/\sqrt{n}} \le a\Big)$$

que lleva al intervalo de confianza

$$\Big\{\mu : \overline{X}-t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \le \mu \le \overline{X}+t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\Big\}.$$

Observación. Note que los intervalos anteriores son simétricos.

Considere ahora,

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

que es cantidad pivotal y elija a y b, satisfaciendo que

$$\mathsf{P}(a \leq \chi^2 \leq b) = \mathsf{P}\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = 1-\alpha,$$

desde donde obtenemos

$$\left\{\sigma^2 : \frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right\}.$$

Las elecciones de a y b que producen el intervalo con el coeficiente de confianza requerido son $a = \chi^2_{1-\alpha/2}(n-1)$ y $b = \chi^2_{\alpha/2}(n-1)$.

4.2. Intervalos de Confianza Asintóticos

DEFINICIÓN 4.6. Considere $\mathsf{SE} = \sqrt{\mathsf{var}(\widehat{\theta}_n)}$. Entonces $\widehat{\mathsf{SE}} = \sqrt{1/\mathcal{F}_n(\widehat{\theta}_n)}$, luego un intervalo del confianza del $100(1-\alpha)\,\%$ para θ es dado por[1]

$$IC_n(\theta) = [\widehat{\theta}_n - z_{1-\alpha/2}\widehat{\mathsf{SE}}, \widehat{\theta}_n + z_{1-\alpha/2}\widehat{SE}].$$

EJEMPLO 4.7. Sea $X_1, \ldots, X_n$ muestra aleatoria desde $\mathsf{Ber}(p)$. Sabemos que el MLE de p es $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^n x_i$, y

$$\log f(x;p) = x \log p + (1-x)\log(1-p),$$

así

$$U(x;p) = \frac{x}{p} - \frac{1-x}{1-p}, \qquad U'(x;p) = \frac{x}{p^2} + \frac{1-x}{(1-p)^2}.$$

De este modo,

$$\mathcal{F}_1(p) = \mathsf{E}\{-U'(X;p)\} = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)},$$

de ahí que

$$\widehat{\mathsf{SE}} = \frac{1}{\sqrt{\mathcal{F}_n(\widehat{p}_n)}} = \frac{1}{\sqrt{n\mathcal{F}_1(\widehat{p}_n)}} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}},$$

luego, un intervalo de confianza del $100(1-\alpha)\,\%$ para p es dado por

$$\widehat{p}_n \mp z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

OBSERVACIÓN. Considere $\lambda = g(\theta)$. Sabemos que el estimador ML de λ es dado por $\widehat{\lambda}_n = g(\widehat{\theta}_n)$. Además, usando el método Delta, sigue que

$$\frac{(\widehat{\lambda}_n - \lambda)}{\widehat{\mathsf{SE}}(\widehat{\lambda}_n)} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1),$$

donde

$$\widehat{\mathsf{SE}}(\widehat{\lambda}_n) = |g'(\widehat{\theta}_n)|\,\widehat{\mathsf{SE}}(\widehat{\theta}_n).$$

Lo que lleva al intervalo de confianza asintótico

$$IC_n(\lambda) = [\widehat{\lambda}_n - z_{1-\alpha/2}\,\widehat{\mathsf{SE}}(\widehat{\lambda}_n), \widehat{\lambda}_n + z_{1-\alpha/2}\,\widehat{\mathsf{SE}}(\widehat{\lambda}_n)].$$

[1] En efecto, $\mathsf{P}_\theta(\theta \in IC_n(\theta)) \to 1-\alpha$ para $n \to \infty$.

EJEMPLO 4.8. Considere $X_1, \ldots, X_n$ variables aleatorias IID desde una FE 1-paramétrica, y sea $\phi = \eta(\theta)$, $\gamma(\phi) = b(\theta)$. De este modo,

$$f(x;\phi) = \exp[\phi T(x) - \gamma(\phi)]h(x),$$

es decir, la log-verosimilitud para una única observación es dada por

$$\log f(x;\phi) = \phi T(x) - \gamma(\phi) + \log h(x).$$

Lo que lleva a,

$$U(x;\phi) = T(x) - \gamma'(\phi), \qquad U'(x;\phi) = -\gamma''(\phi),$$

de ahí que, la información de Fisher es dada por

$$\mathcal{F}_1(\phi) = \mathsf{E}\{-U'(X;\phi)\} = \gamma''(\phi) = \mathsf{var}(T(X)).$$

Es decir, el error estándar adopta la forma

$$\widehat{\mathsf{SE}} = \frac{1}{\sqrt{\mathcal{F}_n(\widehat{\phi}_n)}} = \frac{1}{\sqrt{n\mathcal{F}_1(\widehat{\phi}_n)}} = \frac{1}{\sqrt{n\gamma''(\widehat{\phi}_n)}},$$

donde $\widehat{\phi}_n$ denota el MLE de ϕ (ver Ejemplo 3.16). Finalmente, un intervalo de confiaza del $100(1-\alpha)\,\%$ para ϕ es dado por

$$IC_n(\phi) = \left[\widehat{\phi}_n \mp z_{1-\alpha/2}\,\frac{1}{\sqrt{n\gamma''(\widehat{\phi}_n)}}\right].$$

Evidentemente, también podemos considerar un intervalo de confianza asintótico para $\theta = g(\phi) = \eta^{-1}(\phi)$ usando el método Delta.

4.3. Regiones de Confianza Asintóticas

Cuando $\boldsymbol{\theta}$ es un vector k-dimensional, podemos definir una región de confianza (asintótica), mediante

$$\lim_{n\to\infty} \mathsf{P}_\theta(\boldsymbol{\theta} \in RC_n(\boldsymbol{\theta})) = 1 - \alpha,$$

cuando $\boldsymbol{\theta}$ es el verdadero vector de parámetros. El mecanismo usado para construir una región de confianza asintótica está basada en el siguiente resultado.

RESULTADO 4.9. *Sea $\{\boldsymbol{T}_n\}$ una secuencia de vectores aleatorios k-dimensionales tal que $\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathsf{D}} \mathsf{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$ y sea $\{\boldsymbol{A}_n\}$ una secuencia de matrices aleatorias tal que $\boldsymbol{A}_n \xrightarrow{\mathsf{P}} \boldsymbol{A}$, donde $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{A}$. Entonces*

$$Q_n = n(\boldsymbol{T}_n - \boldsymbol{\theta})^\top \boldsymbol{A}_n(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathsf{D}} \chi^2(k).$$

DEMOSTRACIÓN. Disponible en Sen y Singer (1993), página 137. □

Basado en el Resultado 3.40 y usando que $\boldsymbol{\mathcal{F}}_1(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{\mathsf{P}} \boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta})$, tenemos

$$n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^\top \boldsymbol{\mathcal{F}}_1(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathsf{D}} \chi^2(k).$$

Es decir, podemos construir una región de confianza del $100(1-\alpha)\,\%$ para $\boldsymbol{\theta}$ como:

$$RC_n(\boldsymbol{\theta}) = \{\boldsymbol{\theta} : n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^\top \boldsymbol{\mathcal{F}}_1(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \leq \chi^2_{1-\alpha}(k)\},$$

donde $\chi^2_{1-\alpha}(k)$ denota un valor cuantil $1-\alpha$ de la distribución chi-cuadrado con k grados de libertad.

Ejemplo 4.10. Considere $X_1, \ldots, X_n$ una muestra aleatoria desde $\mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Es fácil notar que $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ y

$$\boldsymbol{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^\top \xrightarrow{\mathsf{P}} \boldsymbol{\Sigma}.$$

De este modo, la estadística T^2 de Hotelling, satisface que

$$T_n^2 = n(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})^\top \boldsymbol{S}_n^{-1} (\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathsf{D}} \chi^2(p).$$

Luego, una región de confianza asintótica del $100(1-\alpha)\,\%$ para $\boldsymbol{\mu}$ es dada por:

$$RC_n(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : n(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})^\top \boldsymbol{S}_n^{-1} (\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \leq \chi^2_{1-\alpha}(p)\}.$$

Bibliografía

Bates, D.M., Watts, D.G. (1981). A relative offset orthogonality convergence criterion for nonlinear least squares. *Technometrics* **23**, 179-183.

Berndt, E.K., Hall, B.H., Hall, R.E., Hausman, J.A. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurements* **3**, 653-665.

Crudu, F., Osorio, F. (2020). Bilinear form test statistics for extremum estimation. *Economics Letters* **187**, 108885.

Casella, G., Berger, R.L. (2002). *Statistical Inference (2nd Ed.)*. Duxbury, Pacific Grove.

Gómez, E., Gómez-Villegas, M.A., Marín, J.M. (1988). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics - Theory and Methods* **27**, 589-600.

Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall, London.

Jørgensen, B., Labouriau, R. (1994). *Exponential Families and Theoretical Inference*. Lecture Notes, Department of Statistics, University of British Columbia.

Lange, K., Sinsheimer, J.S. (1993). Normal/Independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* **2**, 175-198.

Lange, K., Little, R.J.A., Taylor, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881-896.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. Wiley, New York.

Mood, A.M., Graybill, F.A., Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd Edition. McGraw-Hill, New York.

Osorio, F. (2019). heavy: Robust estimation using heavy-tailed distributions. *R package version 0.38.196*, URL: https://CRAN.R-project.org/package=heavy

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford University Press.

Pinheiro, J., Bates, D.M, DebRoy, S., Sarkar, D. (2019). nlme: Linear and nonlinear mixed effects models. *R package version 3.1-143*, URL: https://CRAN.R-project.org/package=nlme

Rohde, C.A. (2014). *Introductory Statistical Inference with the Likelihood Function*. Springer, New York.

Sen, P.K., Singer, J.M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall, London.

Wasserman, L. (2003). *All of Statistics: A concise course in statistical inference*. Springer, New York.