# MAT-206: Sesión 12, Funciones de inferencia

**Felipe Osorio**

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Sea $\boldsymbol{Y}$ vector aleatorio $N$-dimensional desde un modelo estadístico

$$\mathcal{P} = \{\mathsf{P}_\theta : \boldsymbol{\theta} \in \Theta\}, \qquad \Theta \in \mathbb{R}^k,$$

y $\mathcal{Y}$ es el espacio muestral.

**Definición 1:**

Una función $\boldsymbol{\Psi} : \Theta \times \mathcal{Y} \to \mathbb{R}^k$ tal que $\boldsymbol{\Psi}(\boldsymbol{\theta}; \cdot)$ es medible para todo $\boldsymbol{\theta} \in \Theta$ se dice una función de inferencia.

*Observación:*

Para una función de inferencia $\boldsymbol{\Psi}$ y una muestra $\boldsymbol{Y} \in \mathcal{Y}$ dadas, es posible obtener un estimador $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ como solución de la ecuación

$$\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{Y}) = \mathbf{0}.$$

*Ejemplo (Función score):*

Denote por $\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta})$ a la función de log-verosimilitud para $\boldsymbol{\theta}$. La función score

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

corresponde a una función de inferencia.

Note que, aquél valor $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$ que maximiza $\ell(\boldsymbol{\theta})$ y es una solución de la ecuación

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{0},$$

sobre $\Theta$, es llamado el estimador máximo verosímil para $\boldsymbol{\theta}$.

**Definición 2 (Función de inferencia regular):**

Una función de inferencia $\boldsymbol{\Psi}$ se dice regular para todo $\boldsymbol{\theta} \in \Theta$ si satisface las siguientes condiciones:

(i) Tiene segundo momento finito.

(ii) Es insesgada, esto es, $\mathsf{E}_\theta\{\boldsymbol{\Psi}(\boldsymbol{\theta})\} = \mathbf{0}$.

(iii) La matriz de sensibilidad

$$\boldsymbol{S}_\Psi(\boldsymbol{\theta}) = \mathsf{E}_\theta \Big\{ -\frac{\partial \boldsymbol{\Psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \Big\},$$

es no singular para todo $\boldsymbol{\theta}$.

**Definición 3 (Matriz de variabilidad):**

Considere $\boldsymbol{\Psi}$ función de inferencia regular la matriz de variabilidad, es definida como:

$$\boldsymbol{V}_{\Psi}(\boldsymbol{\theta}) = \mathsf{Cov}_{\theta}(\boldsymbol{\Psi}(\boldsymbol{\theta})) = \mathsf{E}_{\theta}\{\boldsymbol{\Psi}(\boldsymbol{\theta})\boldsymbol{\Psi}^{\top}(\boldsymbol{\theta})\},$$

**Definición 4 (Información de Godambe):**

Para $\boldsymbol{\Psi}$ función de inferencia regular, la matriz de información de Godambe es dada por:

$$\boldsymbol{G}_{\Psi}(\boldsymbol{\theta}) = \boldsymbol{S}_{\Psi}^{\top}(\boldsymbol{\theta})\boldsymbol{V}_{\Psi}^{-1}(\boldsymbol{\theta})\boldsymbol{S}_{\Psi}(\boldsymbol{\theta}).$$

# Función de inferencia

*Observación:*

Note que, para $\boldsymbol{C}(\boldsymbol{\theta})$ matriz no estocástica

$$\boldsymbol{\Phi}(\boldsymbol{\theta}; \boldsymbol{Y}) = \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{Y}),$$

también es una función de inferencia. En cuyo caso anotamos $\boldsymbol{\Phi} \sim \boldsymbol{\Psi}$.

*Observación:*

Asuma las condiciones A1 a A4. De este modo, $\mathsf{E}\{\boldsymbol{U}(\boldsymbol{\theta})\} = \boldsymbol{0}$ y tenemos que

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{Cov}\{\boldsymbol{U}(\boldsymbol{\theta})\},$$

corresponde a la matriz de información de Fisher.

### Definición 5 (Función de estimación óptima):

Sea $\mathbf{\Psi}$ y $\mathbf{\Phi}$ funciones de inferencia regulares para el vector de parámetros $\boldsymbol{\theta}$. Si se satisface que

$$\boldsymbol{G}_{\Psi}^{-1}(\boldsymbol{\theta}) - \boldsymbol{G}_{\Phi}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0},$$

para cualquier $\mathbf{\Psi}$, entonces se dice que $\mathbf{\Phi}$ es función de inferencia óptima.

### Resultado 1 (Desigualdad de Godambe):

Asuma que $\mathbf{\Psi}$ es función de inferencia regular. Entonces

$$\boldsymbol{G}_{\Psi}^{-1}(\boldsymbol{\theta}) - \boldsymbol{\mathcal{F}}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0},$$

para todo $\boldsymbol{\theta} \in \Theta$, donde la igualdad se cumple sólo si $\mathbf{\Psi} \sim \boldsymbol{U}$, la función score.

*Esbozo de la demostración:*

Considere el caso unidimensional $(k=1)$ y sea $U_n(\theta) = \mathsf{d}\log p(\boldsymbol{y};\theta)/\,\mathsf{d}\,\theta$. Como $\Psi(\theta)$ es regular, tenemos

$$\int \Psi(\theta)\, p(\boldsymbol{y};\theta)\,\mathsf{d}\boldsymbol{y} = 0,$$

de ahí que

$$\begin{aligned} 0 &= \frac{\mathsf{d}}{\mathsf{d}\,\theta}\int \Psi(\theta)p(\boldsymbol{y};\theta)\,\mathsf{d}\boldsymbol{y} = \int \frac{\mathsf{d}}{\mathsf{d}\,\theta}\{\Psi(\theta)p(\boldsymbol{y};\theta)\}\,\mathsf{d}\boldsymbol{y} \\ &= \int \frac{\mathsf{d}}{\mathsf{d}\,\theta}\Psi(\theta)p(\boldsymbol{y};\theta)\,\mathsf{d}\boldsymbol{y} + \int \Psi(\theta)\frac{\mathsf{d}}{\mathsf{d}\,\theta}\,p(\boldsymbol{y};\theta)\,\mathsf{d}\boldsymbol{y} \end{aligned}$$

Por otro lado,

$$\frac{\mathsf{d}}{\mathsf{d}\,\theta}\,p(\boldsymbol{y};\theta) = \frac{p(\boldsymbol{y};\theta)}{p(\boldsymbol{y};\theta)}\,\frac{\mathsf{d}}{\mathsf{d}\,\theta}\,p(\boldsymbol{y};\theta) = p(\boldsymbol{y};\theta)\frac{\mathsf{d}}{\mathsf{d}\,\theta}\log p(\boldsymbol{y};\theta) = p(\boldsymbol{y};\theta)\,U_n(\theta).$$

De este modo,

$$0 = \int \frac{\mathsf{d}}{\mathsf{d}\,\theta}\,\Psi(\theta)\,p(\boldsymbol{y};\theta)\,\mathsf{d}\boldsymbol{y} + \int \Psi(\theta) U_n(\theta) p(\boldsymbol{y};\theta)\,\mathsf{d}\boldsymbol{y}.$$

Notando que

$$\int \Psi(\theta) U_n(\theta) p(\boldsymbol{y};\theta)\,\mathsf{d}\boldsymbol{y} = \mathsf{Cov}(\Psi(\theta), U_n(\theta)).$$

Es decir, tenemos que

$$\mathsf{E}_\theta^2\left(-\frac{\mathsf{d}\,\Psi(\theta)}{\mathsf{d}\,\theta}\right) = \mathsf{Cov}^2\left(\Psi(\theta), U_n(\theta)\right) \leq \mathsf{E}_\theta(\Psi^2(\theta))\,\mathsf{E}_\theta(U_n^2(\theta)).$$

Por lo tanto,

$$\mathsf{E}_\theta(U_n^2(\theta)) \geq \frac{\mathsf{E}_\theta^2(-\,\mathsf{d}\,\Psi(\theta)/\,\mathsf{d}\,\theta)}{\mathsf{E}_\theta(\Psi^2(\theta))}.$$

Finalmente,

$$\mathcal{F}^{-1}(\theta) \leq G_\Psi^{-1}(\theta).$$

**Definición 6 (Algoritmo Newton-scoring):**

Considere la expansión en series de Taylor de $\boldsymbol{\Psi}(\boldsymbol{\theta})$ en torno de $\boldsymbol{\theta}_0$, tenemos

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) \approx \boldsymbol{\Psi}(\boldsymbol{\theta}_0) + \frac{\partial \boldsymbol{\Psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big|_{\theta=\theta_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

como $\boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ y substituyendo $\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0)$ por $\boldsymbol{S}_\Psi(\boldsymbol{\theta}_0)$, sigue que

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \boldsymbol{S}_\Psi^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Psi}(\boldsymbol{\theta}_0),$$

esto sugiere considerar:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{S}_\Psi^{-1}(\boldsymbol{\theta}^{(t)})\boldsymbol{\Psi}(\boldsymbol{\theta}^{(t)}),$$

para llevar a cabo la estimación de parámetros.

*Observación:*

Este procedimiento fue propuesto por Jørgensen y Knudsen (2004)[1] quienes lo denominaron algoritmo Newton-scoring.

[1] Scandinavian Journal of Statistics **31**, 93-114.

**Resultado 2:**

Sea $\{\widehat{\boldsymbol{\theta}}_n\}_{n\geq 1}$ una secuencia de raíces de las ecuaciones de estimación

$$\boldsymbol{\Psi}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{\theta}; \boldsymbol{Y}_i) = \mathbf{0},$$

desde la condición de insesgamiento se tiene la consistencia $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_0$.

**Definición 7 (Optimalidad de Crowder):**

Considere la función de inferencia regular (aditiva):

$$\boldsymbol{\Psi}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{C}_i(\boldsymbol{\theta})\boldsymbol{\Psi}_i(\boldsymbol{\theta}; \boldsymbol{Y}_i), \qquad \boldsymbol{\theta} \in \Theta,$$

donde $\boldsymbol{C}_i(\boldsymbol{\theta})$ es una matriz no aleatoria de $\boldsymbol{\theta}$, tal que la secuencia $\{\widehat{\boldsymbol{\theta}}_n\}_{n\geq 1}$ es consistente. Entonces la función de inferencia optimal está definida por

$$\boldsymbol{C}_i(\boldsymbol{\theta}) = \mathsf{E}_\theta^\top\{-\dot{\boldsymbol{\Psi}}_i(\boldsymbol{\theta})\}\,\mathsf{Cov}_\theta^{-1}\{\boldsymbol{\Psi}_i(\boldsymbol{\theta})\}.$$

## Supuestos

**C1:** $\boldsymbol{\Psi}_n(\boldsymbol{\theta}) \to \mathbf{0}$ con probabilidad 1.

**C2:** Existe una vecindad de $\boldsymbol{\theta}_0$ tal que, con prob. 1, $\boldsymbol{\Psi}_n(\boldsymbol{\theta})$ es diferenciable y $\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\theta})$ converge a un límite no estocástico que es no singular en $\boldsymbol{\theta}_0$.

**C3:** $\frac{1}{\sqrt{n}}\boldsymbol{\Psi}_n(\boldsymbol{\theta}) \xrightarrow{\mathsf{D}} \mathsf{N}_k(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta}_0))$.

## Resultado 3 (Normalidad asintótica):

Si $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_0$ y $\boldsymbol{\Psi}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$ con probabilidad 1. Entonces, bajo los supuestos C1 a C3, tenemos que

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p\big(\mathbf{0}, \boldsymbol{G}_{\Psi}^{-1}(\boldsymbol{\theta}_0)\big).$$

*Demostración:*

Detalles en Yuan y Jennrich (1998)[2]

[2] Journal of Multivariate Analysis **65**, 245-260.

*Ejemplo (Ecuaciones de estimación generalizadas, GEE):*

Liang y Zeger (1986)[3] propusieron GEE como un método para análisis de datos con estructura longitudinal.

Considere $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{im_i})^\top$ vector de respuestas $m_i$-dimensional asociada al $i$-ésimo individuo. Es asumido que

$$\mathsf{E}(Y_{ij}) = \mu_{ij}(\boldsymbol{\beta}), \qquad \mathsf{var}(Y_{ij}) = \phi^{-1} V(\mu_{ij}),$$

donde $V(\mu)$ es función de varianza y $\phi > 0$. La estimación de $\boldsymbol{\beta} \in \mathbb{R}^p$ se lleva a cabo mediante resolver el sistema de ecuaciones:

$$\boldsymbol{\Psi}(\boldsymbol{\beta}) := \sum_{i=1}^{n} \Big( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^\top} \Big)^\top \{ \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2} \}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0},$$

con $\boldsymbol{A}_i = \operatorname{diag}(V(\mu_{i1}), \dots, V(\mu_{im_i}))$ y $\boldsymbol{R}_i(\boldsymbol{\alpha})$ es conocida como matriz de correlación de trabajo.

[3] Biometrika **73**, 13-22.

*Ejemplo (Quasi-verosimilitud):*

Suponga que $\text{E}(Y) = \mu$ y $\text{var}(Y) = \phi^{-1}V(\mu)$. Wedderburn (1974)[4] mediante una analogía con la función score para la familia exponencial definió

$$Q(\mu; Y) = \int_y^{\mu} \phi \, \frac{y - t}{V(\mu)} \, \text{d}t.$$

Para observaciones dependientes McCullagh y Nelder (1989)[5] consideran

$$Q_i(\boldsymbol{\mu}_i; \boldsymbol{Y}_i) = -\frac{1}{\phi}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^{\top} \Big( \int_0^1 s\{\boldsymbol{V}_i(\boldsymbol{t}(s))\}^{-1} \, \text{d}s \Big)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

donde la integral está definida a lo largo de la recta $\boldsymbol{t}(s) = \boldsymbol{Y}_i + (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)s$, para $0 \leq s \leq 1$. Una función de inferencia para $\boldsymbol{\beta}$ se obtiene como la primera derivada de

$$Q(\boldsymbol{\mu}; \boldsymbol{Y}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\mu}_i; \boldsymbol{Y}_i).$$

[4] Biometrika **61**, 439-447.

[5] Generalized Linear Models, Chapman & Hall, London.

# Funciones de inferencia

*Ejemplo (Función de inferencia cuadrática, QIF):*

Suponga que $\boldsymbol{R}^{-1} = \sum_{k=1}^{m} a_k \boldsymbol{M}_k$, con $\boldsymbol{M}_1, \dots, \boldsymbol{M}_m$, matrices conocidas, mientras que $a_1, \dots, a_m$ son constantes desconocidas. Haciendo $\boldsymbol{\Phi}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Phi}_i(\boldsymbol{\beta})$ con

$$\boldsymbol{\Phi}_i(\boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{D}_i^\top \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_1 \boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \boldsymbol{D}_i^\top \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_m \boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \end{pmatrix},$$

donde $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^\top$ para $i = 1, \dots, n$.

Podemos estimar $\boldsymbol{\beta}$, en el contexto de GEE, mediante minimizar la QIF:

$$Q_n(\boldsymbol{\beta}) = n\, \boldsymbol{\Phi}_n^\top(\boldsymbol{\beta}) \boldsymbol{C}_n^{-1}(\boldsymbol{\beta}) \boldsymbol{\Phi}_n(\boldsymbol{\beta}),$$

con $\boldsymbol{C}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Phi}_i(\boldsymbol{\beta}) \boldsymbol{\Phi}_i^\top(\boldsymbol{\beta})$.