

MAT-206: Sesión 5, Suficiencia y función de verosimilitud

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM



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

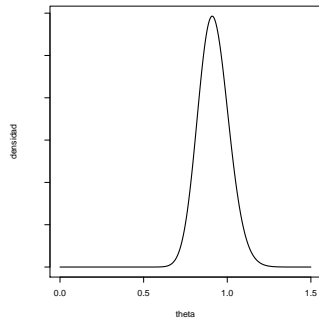
$$\begin{aligned} f(\mathbf{x}; \theta) &= \prod_{i=1}^n \theta \exp(-\theta x_i) = \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right) \\ &= \theta^n \exp(-\theta n\bar{x}). \end{aligned}$$

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

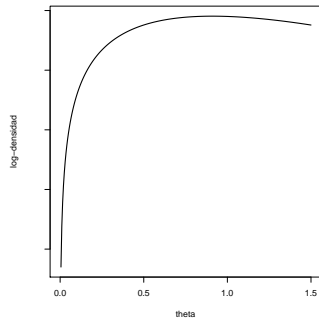
IDEA:

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





(a)



(b)

$T : \mathcal{X}^n \rightarrow \mathbb{R}$ reduce una colección de n observaciones a un único número y por tanto no puede ser inyectiva. Es decir, en general $T(X_1, \dots, X_n)$ provee **menos** información sobre θ que (X_1, \dots, X_n) .

Para algunos modelos una estadística T será igualmente informativa sobre θ que la muestra (X_1, \dots, X_n) . Tales estadísticas son llamadas **estadísticas suficientes**¹

¹Es suficiente usar T en lugar de (X_1, \dots, X_n) .



Definición 1:

Sea X_1, \dots, X_n variables aleatorias IID desde el modelo $\{P_\theta : \theta \in \Theta\}$. Una estadística $T : \mathcal{X}^n \rightarrow \mathbb{R}$ se dice **suficiente** para θ , si

$$P(X_1 \leq x_1, \dots, X_n \leq x_n | T = t),$$

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



Ejemplo:

Suponga X_1, \dots, X_n variables aleatorias IID desde $\text{Ber}(\theta)$, 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 $t \in \mathcal{T} = \{0, 1, \dots, n\}$. Ahora, note que la distribución conjunta de X_1, \dots, X_n es dada por

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

Por otro lado, sabemos que

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

con probabilidad

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



De este modo,

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

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



Resultado 1 (Factorización de Fisher-Neyman):

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

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

Para una demostración simple revisar Casella y Berger (2002, p. 276)²

²*Statistical Inference (2nd Edition)*. Duxbury, Pacific Grove.



Ejemplo:

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

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

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

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

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



Ejemplo:

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

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

Ahora,

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

Es decir, podemos escribir la densidad conjunta como

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

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



Ejemplo:

Suponga $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. La densidad conjunta puede ser escrita como:

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

como $\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} = \text{tr}(\mathbf{x} \mathbf{x}^\top \boldsymbol{\Sigma}^{-1})$, tenemos

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\left\{\mathbf{T}_1^\top(\mathbf{x}) \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \frac{1}{2} \text{tr}(\mathbf{T}_2(\mathbf{x}) \boldsymbol{\Sigma}) - \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \frac{1}{2} \log |2\pi \boldsymbol{\Sigma}|\right\},$$

con $\mathbf{T}(\mathbf{X}) = (\mathbf{T}_1(\mathbf{X}), \mathbf{T}_2(\mathbf{X}))$, es decir $\mathbf{T}_1(\mathbf{X}) = \mathbf{X}$ y $\mathbf{T}_2(\mathbf{X}) = \mathbf{X} \mathbf{X}^\top$, son estadísticas suficientes.



Ejemplo:

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

$$f(\mathbf{x}; \theta) = \exp \left[\sum_{i=1}^n T(X_i) \eta(\theta) - nb(\theta) \right] \tilde{h}(\mathbf{x}).$$

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



La **información de Kullback-Leibler (KL)** entre las funciones de densidad $g(x)$ y $f(x)$ es dada por:³

$$I(g : f) = \int \log \left(\frac{g(x)}{f(x)} \right) g(x) \, dx = \mathbb{E}_G \left[\log \left(\frac{g(x)}{f(x)} \right) \right].$$

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).

³En ocasiones anotamos $I(G : F) = \int \log(g/f) \, dG$.



Ejemplo:

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

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

Ahora, para

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

sigue que

$$\begin{aligned} E_G(\log f(x)) &= E_G \left[-\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} (X - \mu)^2 \right] \\ &= -\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} [\phi^2 + (\theta - \mu)^2]. \end{aligned}$$



Por otro lado,

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

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

$$\begin{aligned} I(g : f) &= \mathbb{E}_G(\log g(x)) - \mathbb{E}_G(\log f(x)) \\ &= \frac{1}{2} \left\{ \log \frac{\sigma^2}{\phi^2} + \frac{\phi^2 + (\theta - \mu)^2}{\sigma^2} - 1 \right\} \end{aligned}$$



Suponga X_1, \dots, X_n variables aleatorias IID desde una CDF desconocida $G(x)$. Asumiremos que $G(x)$ corresponde al **modelo estadístico verdadero** y sea $F(x)$ el **modelo asumido**

Supondremos también que asociadas a G y F 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.



Tenemos

$$I(g : f) = E_G \left[\log \frac{g(x)}{f(x)} \right] = E_G[\log g(x)] - E_G[\log f(x)],$$

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

Observación:

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



Además,

$$\mathbb{E}_G[\log f(x)] = \int \log f(x) dG(x),$$

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

$$\begin{aligned}\mathbb{E}_{\hat{G}_n}[\log f(x)] &= \int \log f(x) d\hat{G}_n(x) = \sum_{i=1}^n \hat{g}_n(x_i) \log f(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \log f(x_i).\end{aligned}$$

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

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



Definición 2 (Función de verosimilitud):

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

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

es definida como

$$L(\theta; x) = f(x; \theta), \quad \theta \in \Theta.$$

Observación:

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

