MAT-206: Suficiencia y función de verosimilitud

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Suficiencia

Considere $X_1, \dots, X_n$ variables aleatorias IID desde Exp(θ), de este modo

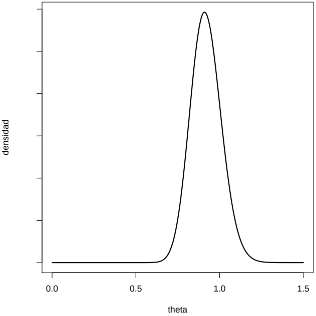
$$\begin{aligned} f(\boldsymbol{x};\theta) &= \prod_{i=1}^{n} \theta \exp(-\theta x_i) = \theta^n \exp\Big(-\theta \sum_{i=1}^{n} x_i\Big) \\ &= \theta^n \exp(-\theta n \overline{x}). \end{aligned}$$

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

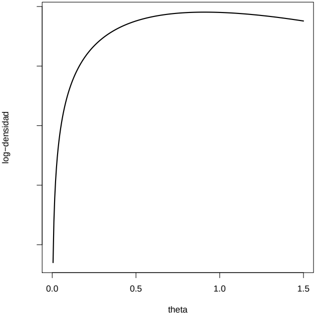
Idea:

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

Suficiencia



(a)



(b)

Suficiencia

$T : \mathcal{X}^n \to \mathbb{R}$ 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[1]

[1] Es suficiente usar T en lugar de $(X_1, \ldots, X_n)$.

Definición 1:

Sea $X_1, \dots, X_n$ 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

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

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 Ber(θ), 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

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

Por otro lado, sabemos 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.

Resultado 1 (Factorización de Fisher-Neyman):

Suponga que $X_1, \ldots, X_n$ 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

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

Demostración:

En Casella y Berger (2002, p. 276)[2], se presenta una demostración para el caso discreto. En el caso continuo una prueba usando Teoría de la Medida es dada en Lehmann (1986, p. 54)[3]

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

[3] Testing Statistical Hypotheses. Wiley, New York.

Suficiencia

Ejemplo:

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:

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

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

Ahora,

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

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).

Ejemplo:

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

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

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

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\big\{\boldsymbol{T}_1^\top(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\operatorname{tr}(\boldsymbol{T}_2(\boldsymbol{x})\boldsymbol{\Sigma}) - \tfrac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\log|2\pi\boldsymbol{\Sigma}|\big\},$$

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

Ejemplo:

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.

Información de Kullback-Leibler

La información de Kullback-Leibler (KL) entre las funciones de densidad $g(x)$ y $f(x)$ es dada por:[4]

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

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).

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

Ejemplo:

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

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

Por otro lado,

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

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

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

Función de verosimilitud

Suponga $X_1, \ldots, X_n$ 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

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

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.

Función de verosimilitud

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 estimador usando en la CDF empírica $\widehat{G}_n$ basada en los datos observados $X_1, \dots, X_n$. Es decir,

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

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

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

Definición 2 (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.