MAT-042: Estimación de momentos y máximo verosímil

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Definición 1 (Estadística):

Una estadística es una función de los datos $(X_1, \dots, X_n)$ que no depende de parámetros desconocidos.

Ejemplo:

Considere $\boldsymbol{X} = (X_1, \dots, X_n)$, de este modo

$$T_1(\boldsymbol{X}) = \overline{X}, \qquad T_2(\boldsymbol{X}) = S^2$$
$$T_3(\boldsymbol{X}) = X_{(n)}, \qquad T_4(\boldsymbol{X}) = \frac{X_i - \overline{X}}{S},$$

son estadísticas. Mientras que

$$T_5(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2,$$

para μ desconocido no es una estadística.

Idea:

El objetivo de la inferencia estadística es obtener información sobre la distribución de X a partir de los datos observados $\boldsymbol{x} = (x_1, \dots, x_n)$.

Supuesto:

Asumiremos que X es un miembro de la familia

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

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

Ejemplo (Modelo Poisson):

Considere $X_1, \dots, X_n$ variables aleatorias IID desde Poi(λ) con densidad conjunta

$$p_n(\boldsymbol{x}; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Es decir, tenemos el modelo:

$$\mathcal{P} = \{p_n(\boldsymbol{x}; \lambda) : \lambda \in (0, +\infty)\}.$$

Ejemplo (Modelo Normal):

Suponga una muestra aleatoria $X_1, \dots, X_n$ desde $\mathsf{N}(\mu, \sigma^2)$. Podemos escribir su densidad conjunta como:

$$\begin{aligned} f_n(\boldsymbol{x}; \mu, \sigma^2) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\Big\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \Big\} \\ &= (2\pi\sigma^2)^{n/2} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big\} \end{aligned}$$

Así, tenemos el modelo estadístico

$$\mathcal{P} = \{ f_n(\boldsymbol{x}; \mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0 \},$$

es decir, $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ y $\Theta = \mathbb{R} \times \mathbb{R}_+$.

Ejemplo (Modelo de regresión lineal simple):

Suponga el modelo,

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \dots, n,$$

con $\epsilon_i \overset{\text{IID}}{\sim} \mathsf{N}(0, \sigma^2)$, de ahí que

$$Y_i \sim \mathsf{N}(\alpha + \beta x_i, \sigma^2), \qquad i = 1, \dots, n.$$

De ahí que

$$f_n(\boldsymbol{y}; \alpha, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \Big\}$$

cuyo modelo asociado es dado por:

$$\mathcal{P} = \{f_n(\boldsymbol{x}; \alpha, \beta, \sigma^2) : \alpha, \beta \in \mathbb{R}, \sigma^2 > 0\}.$$

De este modo, $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)^\top$ con $\Theta = \mathbb{R}^2 \times \mathbb{R}_+$.

Suponga $X_1, \dots, X_n$ una muestra aleatoria desde el modelo estadístico

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

Deseamos entender el mecanismo o modelo verdadero que generó los datos.

Idea:

El objetivo es desarrollar procedimientos para 'seleccionar' $\boldsymbol{\theta}$ desde los datos observados[1]

Notación:

Denotaremos por $\mathcal{X}$ al espacio muestral asociado a la muestra aleatoria $\boldsymbol{X} = (X_1, \dots, X_n)$.

[1] Conocido como estimación.

Definición 2 (Estimador):

Una función $\boldsymbol{T} : \mathcal{X} \to \Theta$ es llamado un estimador (puntual).

Observación:

Un estimador es una regla o fórmula que permite usar los datos para construir un valor plausible de $\boldsymbol{\theta}$

El valor $\boldsymbol{T}(\boldsymbol{x})$ es llamado estimación de $\boldsymbol{\theta}$ y corresponde a una realización de la variable aleatoria $\boldsymbol{T}(\boldsymbol{X})$.

Observación:

Usualmente anotamos un estimador[2] como $\widehat{\boldsymbol{\theta}} = \boldsymbol{T}(X_1, \dots, X_n)$ y distinguimos el método usado por $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$, $\widehat{\boldsymbol{\theta}}_{\mathsf{MM}}$ o $\widehat{\boldsymbol{\theta}}_{\mathsf{LS}}$.

[2] Mientras que a una estimación, por $\widehat{\boldsymbol{\theta}} = \boldsymbol{T}(x_1, \dots, x_n)$.

Ejemplo:

Considere $X_1, \ldots, X_n$ muestra aleatoria desde $\text{Ber}(\theta)$. Suponga

$$\widehat{\theta}(X_1, \ldots, X_n) = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n} S_n,$$

es un estimador de θ, con $S_n = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta)$. Mientras que,

$$\widehat{\theta}(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n),$$

es la proporción (estimación) muestral de éxitos en la muestra.

Principios de estimación

Ejemplo:

Considere los pares de observaciones $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ y suponga que siguen un modelo de regresión

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \dots, n.$$

Tenemos los estimadores LS:

$$\begin{aligned}
\widehat{\alpha}(\boldsymbol{Y}) &= \overline{Y} - \widehat{\beta}(\boldsymbol{Y})\,\overline{x}, \\
\widehat{\beta}(\boldsymbol{Y}) &= \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}.
\end{aligned}$$

Definición 3 (Error Cuadrático Medio):

Sea T un estimador del parámetro desconocido θ. Se define el error cuadrático médio como:

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

Observación:

Es fácil notar que

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

Definición 4 (Sesgo):

El sesgo de un estimador T es definido como:

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

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

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

Definición 5 (Insesgamiento):

Un estimador T para θ se dice insesgado, si

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

o equivalentemente,

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

Definición 6 (Consistencia):

Sea $T_1, T_2, \dots, T_n$ una secuencia de estimadores de θ.[3] Se dice que T es consistente si,

$$\lim_{n\to\infty} \mathsf{P}(|T_n - \theta| \leq \epsilon) = 1.$$

Observación:

También podemos definir un estimador consistente en forma tal, que:

$$\lim_{n\to\infty} \mathsf{MSE}(T_n) = \lim_{n\to\infty} \mathsf{E}\{(T_n - \theta)^2\} = 0.$$

[3] Basados en un estimador T de θ para muestras de tamaño n.

Ejemplo:

Sean $X_1, X_2, \ldots, X_n$ variables aleatorias IID, tal que $\mathsf{E}(X_i) = \mu$ y $\mathsf{var}(X_i) = \sigma^2 < +\infty$. Tenemos que

$$\mathsf{E}(\overline{X}_n) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(X_i) = \mu$$

$$\mathsf{var}(\overline{X}_n) = \mathsf{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{var}(X_i) = \frac{\sigma^2}{n}.$$

Usando la desigualdad de Chebyshev, sigue que

$$\mathsf{P}(|\overline{X}_n - \mu| < \epsilon) \geq 1 - \frac{\mathsf{var}(\overline{X}_n)}{\epsilon^2} = 1 - \frac{\sigma^2}{n\epsilon^2}.$$

Es decir,

$$\lim_{n\to\infty} \mathsf{P}(|\overline{X}_n - \mu| < \epsilon) \geq \lim_{n\to\infty}\left(1 - \frac{\sigma^2}{n\epsilon^2}\right) = 1.$$

Método de momentos

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria desde una distribución con función de densidad $f(x; \boldsymbol{\theta})$. El r-ésimo momento en torno de cero es dado por

$$M_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r.$$

El método de estimación de momentos se basa en construir el sistema de ecuaciones,

$$\begin{aligned} \mu_1 &= M_1 & & \qquad\qquad & \mathsf{E}(X) &= \frac{1}{n}\sum_{i=1}^{n} X_i, \\ \mu_2 &= M_2 & & \text{Es decir,} & \mathsf{E}(X^2) &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \\ &\vdots & & & &\vdots \\ \mu_p &= M_p & & & \mathsf{E}(X^p) &= \frac{1}{n}\sum_{i=1}^{n} X_i^p, \end{aligned}$$

donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$.

Ejemplo:

Suponga $X_1, X_2, \ldots, X_n$ muestra aleatoria desde Gama(a, b) con densidad

$$f(x; a, b) = \frac{b^a}{\Gamma(a)}\, x^{a-1} e^{-bx}, \qquad x > 0, a, b, > 0.$$

Sabemos que

$$\mathsf{E}(X) = \frac{a}{b}, \qquad \mathsf{E}(X^2) = \frac{a}{b^2} + \frac{a^2}{b^2}.$$

Es decir, usando el método de momentos obtenemos el sistema de ecuaciones:

$$\mu = \frac{a}{b} \tag{1}$$

$$\mu_2 = \frac{a}{b^2} + \frac{a^2}{b^2} \tag{2}$$

Resolviendo con relación a a desde **(1)**, sigue que

$$a = \mu b \tag{3}$$

Substituyendo en la Ecuación **(2)**, obtenemos

$$\mu_2 = \frac{\mu b}{b^2} + \frac{(\mu b)^2}{b^2} = \frac{\mu}{b} + \mu^2.$$

es decir,

$$\mu_2 - \mu = \frac{\mu}{b}, \qquad \Rightarrow \qquad b = \frac{\mu}{\mu_2 - \mu^2}.$$

De este modo, por **(3)**, tenemos

$$a = \mu b = \frac{\mu^2}{\mu_2 - \mu^2}.$$

Finalmente,

$$\widehat{a} = \frac{\overline{X}^2}{M_2 - \overline{X}^2}, \qquad \widehat{b} = \frac{\overline{X}}{M_2 - \overline{X}^2}.$$

Definición 7 (Estimador máximo verosímil):

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

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

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

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

o equivalentemente,

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

con $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ la función de log-verosimilitud.

Resultado 1 (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 $\boldsymbol{\gamma}$.

Observación:

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

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

donde $\boldsymbol{U}(\boldsymbol{\theta}) = \partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ corresponde a la función score.

Método de máxima verosimilitud

Ejemplo:

Considere $X_1, \dots, X_n$ muestra aleatoria desde $\text{Ber}(\theta)$. En este caso,

$$L(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Entonces,

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} x_i \log\theta + \Big(n - \sum_{i=1}^{n} x_i\Big)\log(1-\theta) \\ & n\overline{x}\log\theta + (n - n\overline{x})\log(1-\theta). \end{aligned}$$

Derivando con relación a θ, obtenemos

$$\frac{\mathsf{d}\,\ell(\theta)}{\mathsf{d}\,\theta} = \frac{n\overline{x}}{\theta} - \frac{n - n\overline{x}}{1-\theta}.$$

Desde $\mathsf{d}\,\ell(\theta)/\,\mathsf{d}\,\theta = 0$, sigue que

$$n\overline{x}(1-\theta) - n(1-\overline{x})\theta = 0.$$

Es decir, $\widehat{\theta} = \overline{x}$.

Ejemplo:

Considere $X_1, \dots, X_n$ 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

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

resolviendo las ecuaciones anteriores para μ y σ^2, sigue que

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

Observación:

Por la propiedad de invarianza, tenemos:

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

Ejemplo:

Suponga $Y_1, \ldots, Y_n$ muestra aleatoria desde $\mathsf{N}(\theta, 1)$ y considere

$$\begin{aligned}\psi &= \mathsf{P}(Y_1 > 0) = 1 - \mathsf{P}(Y_1 \leq 0) = 1 - \mathsf{P}(Y_1 - \theta \leq 0 - \theta)\\ &= 1 - \mathsf{P}(Z \leq -\theta) = 1 - \Phi(-\theta).\end{aligned}$$

Sabemos que el MLE de θ es $\widehat{\theta} = \overline{x}$. De ahí que

$$\widehat{\psi} = 1 - \Phi(-\widehat{\theta}) = 1 - \Phi(-\overline{x})$$

Suponga θ unidimensional, la varianza de $\widehat{\theta}_{\text{ML}}$ puede ser 'aproximada'[4] por:

$$\text{var}(\widehat{\theta}_{\text{ML}}) \approx 1/\mathcal{F}(\widehat{\theta}_{\text{ML}}),$$

donde

$$\mathcal{F}(\theta) = \text{E}\left\{ -\frac{\text{d}^2\,\ell(\theta)}{\text{d}\,\theta^2} \right\} = \text{var}\left\{ \frac{\text{d}\,\ell(\theta)}{\text{d}\,\theta} \right\},$$

denota la información de Fisher.

En general, para $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$,

$$\text{Cov}(\widehat{\boldsymbol{\theta}}_{\text{ML}}) \approx \boldsymbol{\mathcal{F}}^{-1}(\widehat{\boldsymbol{\theta}}_{\text{ML}}),$$

con

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \text{E}\left\{ -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \right\} = \text{Cov}\{\boldsymbol{U}(\boldsymbol{\theta})\}.$$

[4] La aproximación mejora conforme $n \to \infty$.

Observaciones:

- El método de máxima verosimilitud permite obtener estimadores con propiedades optimales. Por ejemplo, conforme n 'crece'[5]

$$\widehat{\boldsymbol{\theta}} \approx \mathsf{AN}_p(\boldsymbol{\theta}, \boldsymbol{\mathcal{F}}^{-1}(\boldsymbol{\theta})/n).$$

- Aunque operativamente el método de momentos es muy simple. Puede no ser único, y sus propiedades no ser tan simples de estudiar.[6]

- Otros métodos de estimación:
  - Mínimos cuadrados (o más generalmente, Extremum estimators).
  - Procedimientos Bayesianos.

[5] Más formalmente, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \boldsymbol{\mathcal{F}}^{-1}(\boldsymbol{\theta}))$.
[6] Corresponde a un caso particular de funciones de inferencia.