

MAT-032: 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 $\mathbf{X} = (X_1, \dots, X_n)$, de este modo

$$\begin{aligned}T_1(\mathbf{X}) &= \overline{X}, & T_2(\mathbf{X}) &= S^2 \\T_3(\mathbf{X}) &= X_{(n)}, & T_4(\mathbf{X}) &= \frac{X_i - \overline{X}}{S},\end{aligned}$$

son estadísticas. Mientras que

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

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 $\mathbf{x} = (x_1, \dots, x_n)$.

Supuesto:

Asumiremos que X es un miembro de la familia

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

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

Ejemplo (Modelo Poisson):

Considere X_1, \dots, X_n variables aleatorias IID desde $\text{Poi}(\lambda)$ con densidad conjunta

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

Es decir, tenemos el modelo:

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

Ejemplo (Modelo Normal):

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

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

Así, tenemos el modelo estadístico

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

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

Ejemplo (Modelo de regresión lineal simple):

Suponga el modelo,

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

con $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, de ahí que

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

De ahí que

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

cuyo modelo asociado es dado por:

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

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

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$

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

Idea:

El objetivo es desarrollar procedimientos para '**seleccionar**' θ desde los datos observados¹

Notación:

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

¹Conocido como **estimación**.

Definición 2 (Estimador):

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

Observación:

Un estimador es una **regla** o **fórmula** que permite usar los datos para construir un valor plausible de θ

El valor $T(x)$ es llamado **estimación** de θ y corresponde a una realización de la variable aleatoria $T(X)$.

Observación:

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

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

Ejemplo:

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

$$\hat{\theta}(X_1, \dots, X_n) = \frac{1}{n}(X_1 + \dots + X_n) = \frac{1}{n}S_n,$$

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

$$\hat{\theta}(x_1, \dots, x_n) = \frac{1}{n}(x_1 + \dots + x_n),$$

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

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

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

Tenemos los estimadores LS:

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

Definición 3 (Error Cuadrático Medio):

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

$$\text{MSE}(T) = E_{\theta}\{(T - \theta)^2\}.$$

Observación:

Es fácil notar que

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

Definición 4 (Sesgo):

El sesgo de un estimador T es definido como:

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

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

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

Definición 5 (Insesgamiento):

Un estimador T para θ se dice insesgado, si

$$E(T) = \theta, \quad \forall \theta \in \Theta,$$

o equivalentemente,

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



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,

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

Observación:

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

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

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

Propiedades de estimadores puntuales

Ejemplo:

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

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

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

Usando la desigualdad de Chebyshev, sigue que

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

Es decir,

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



Método de momentos

Sea X_1, X_2, \dots, X_n una muestra aleatoria desde una distribución con función de densidad $f(x; \theta)$. 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,

$$\mu_1 = M_1 \qquad E(X) = \frac{1}{n} \sum_{i=1}^n X_i,$$

$$\mu_2 = M_2 \qquad \text{Es decir,} \qquad E(X^2) = \frac{1}{n} \sum_{i=1}^n X_i^2,$$

$$\vdots$$
$$\vdots$$

$$\mu_p = M_p \qquad E(X^p) = \frac{1}{n} \sum_{i=1}^n X_i^p,$$

donde $\theta = (\theta_1, \dots, \theta_p)^\top$.



Ejemplo:

Suponga X_1, X_2, \dots, X_n muestra aleatoria desde $\text{Gama}(a, b)$ con densidad

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

Sabemos que

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

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

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

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

Propiedades de estimadores puntuales

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

$$a = \mu b \quad (3)$$

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,

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

De este modo, por (3), tenemos

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

Finalmente,

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



Definición 7 (Estimador máximo verosímil):

Un estimador $\hat{\theta}_{\text{ML}}$ es llamado **estimador máximo verosímil (MLE)** de θ , si

$$L(\hat{\theta}_{\text{ML}}) \geq L(\theta), \quad \forall \theta \in \Theta.$$

Es decir, $\hat{\theta}_{\text{ML}}$ debe ser solución del siguiente problema de optimización

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

o equivalentemente,

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

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



Resultado 1 (Invarianza del MLE):

Si $\gamma = g(\theta)$ y g es biyectiva. Entonces $\hat{\theta}$ es el MLE para θ si y solo si $\hat{\gamma} = g(\hat{\theta})$ es el MLE para γ .

Observación:

Si $\ell(\theta)$ es continuamente diferenciable, el estimador máximo verosímil $\hat{\theta}_{\text{ML}}$ es dada como una solución de las [ecuaciones de verosimilitud](#):

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

donde $U(\theta) = \partial \ell(\theta) / \partial \theta$ corresponde a la [función score](#).



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,

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

Derivando con relación a θ , obtenemos

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

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

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

Es decir, $\hat{\theta} = \bar{x}$.



Ejemplo:

Considere X_1, \dots, X_n muestra aleatoria desde $N(\mu, \sigma^2)$. De este modo

$$L(\mu, \sigma^2; \mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}.$$

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

$$\hat{\mu}_{\text{ML}} = \bar{x}, \quad \hat{\sigma}_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Observación:

Por la propiedad de invarianza, tenemos:

$$\hat{\sigma} = \left\{ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right\}^{1/2}.$$

Ejemplo:

Suponga Y_1, \dots, Y_n muestra aleatoria desde $N(\theta, 1)$ y considere

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

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

$$\hat{\psi} = 1 - \Phi(-\hat{\theta}) = 1 - \Phi(-\bar{x})$$

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

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

donde

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

denota la **información de Fisher**.

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

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

con

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

⁴La aproximación mejora conforme $n \rightarrow \infty$.

Observaciones:

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

$$\hat{\theta} \approx \text{AN}_p(\theta, \mathcal{F}^{-1}(\theta)/n).$$

- ▶ 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**.

⁵Más formalmente, $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N_p(0, \mathcal{F}^{-1}(\theta))$.

⁶Corresponde a un caso particular de **funciones de inferencia**.