Probabilidad y Estadística (Borradores, Curso 23)
Estimadores puntuales

Sebastian Grynberg

23 de mayo de 2011

La libertad de los pueblos no consiste en palabras,
ni debe existir en los papeles solamente. (...)
Si deseamos que los pueblos sean libres,
observemos religiosamente el sagrado dogma de la igualdad.
(Mariano Moreno, 8/XII/1810)

Índice

1. Introducción

El objetivo principal de la teoría estadística es a partir de las observaciones de un fenómeno aleatorio hacer inferencias sobre la distribución de probabilidades subyacente. Esto es, la estadística provee una descripción del comportamiento de un fenómeno pasado, o alguna predicción de un fenómeno futuro de similar naturaleza.

Debido a que la inferencia estadística está basada en la construcción de modelos probabilísticos para los fenómenos observados, la estadística debería considerarse más como una *interpretación* de los fenómenos que como una *explicación* de los mismos.

1.1. Nociones y presupuestos básicos

Definición 1.1 (Muestra aleatoria)**.** Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y $X : \Omega \to \mathbb{R}$ una variable aleatoria. Una *muestra aleatoria de volumen* n de la variable aleatoria X es una sucesión $X_1, \ldots, X_n$ de variables aleatorias independientes cada una con la misma distribución de X.

Nota Bene. El espacio de probabilidades $(\Omega, \mathcal{A}, \mathbb{P})$ modela el fenómeno aleatorio que se desea estudiar y la muestra aleatoria de la variable X los resultados de las observaciones.

Modelos paramétricos. En todo lo que sigue vamos a suponer que

1. La función de distribución de la variable aleatoria X es *desconocida parcialmente*: se sabe que $F(x) = \mathbb{P}(X \leq x)$ pertenece a una familia, $\mathcal{F}$, de distribuciones conocidas que dependen de un parámetro θ desconocido: $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.

2. El conjunto de los posibles parámetros, Θ, es no vacío y está contenido en $\mathbb{R}^d$.

3. Las distribuciones de la familia $\mathcal{F}$ son *distinguibles*: $F_{\theta_1} \neq F_{\theta_2}$ cuando $\theta_1 \neq \theta_2$.

4. Las distribuciones de la familia $\mathcal{F}$ tienen "densidad". Si se trata de una familia de *distribuciones continuas* esto significa que para cada $\theta \in \Theta$, existe una función densidad de probabilidades (f.d.p.) $f(x|\theta)$ tal que $\frac{d}{dx}F_\theta(x) = f(x|\theta)$. Si se trata de una familia de *distribuciones discretas* esto significa que para cada $\theta \in \Theta$, existe una función de probabilidad (f.p.) $f(x|\theta)$ tal que $F_\theta(x) - F_\theta(x-) = f(x|\theta)$.

5. Es posible conseguir muestras aleatorias de la variable X del volumen que se desee.

Nota Bene. De los presupuestos básicos adoptados resulta que los modelos paramétricos adoptan la forma

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\},$$

donde θ es un parámetro desconocido que puede tomar valores en un espacio paramétrico $\Theta \subset \mathbb{R}^d$.

1.2. Algunas familias paramétricas

En esta sección repasaremos algunas de las familias de distribuciones que se utilizan comúnmente en el análisis de datos en problemas prácticos.

1. Familia Normal, $\mathcal{N}(\mu, \sigma^2)$. Decimos que X tiene distribución normal de parámetros $\mu \in \mathbb{R}$ y $\sigma^2 > 0$ cuando la f.d.p. de X está dada por

$$f(x|\mu,\, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

Vale que $\mathbb{E}[X] = \mu$ y $\mathbb{V}(X) = \sigma^2$.

2. Familia Gamma, $\Gamma(\nu, \lambda)$. Decimos que X tiene distribución gamma de parámetros $\nu > 0$ y $\lambda > 0$ cuando la f.d.p. de X está dada por

$$f(x|\nu, \lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x} \mathbf{1}\{x \geq 0\},$$

donde $\Gamma(\nu) := \int_0^\infty x^{\nu-1}e^{-x}dx$. Vale que $\mathbb{E}[X] = \nu/\lambda$ y $\mathbb{V}(X) = \nu/\lambda^2$.

Casos particulares de las familias Gamma son las familias exponenciales $Exp(\lambda) = \Gamma(1, \lambda)$ y las familias chi cuadrado $\chi^2_\nu = \Gamma(\nu/2, 1/2)$.

3. Familia Beta, $\beta(\nu_1, \nu_2)$. Decimos que X tiene distribución beta de parámetros $\nu_1 > 0$ y $\nu_2 > 0$ cuando la f.d.p. de X está dada por

$$f(x|\nu_1, \nu_2) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1 - 1}(1-x)^{\nu_2 - 1} \mathbf{1}\{0 < x < 1\}.$$

Vale que

$$\mathbb{E}[X] = \frac{\nu_1}{\nu_1 + \nu_2} \qquad \text{y} \qquad \mathbb{V}(X) = \frac{\nu_1 \nu_2}{(\nu_1 + \nu_2)^2(\nu_1 + \nu_2 + 1)}.$$

Notar que cuando los parámetros ν_1 y ν_2 son números naturales se tiene que

$$\frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} = \frac{(\nu_1 + \nu_2 - 1)!}{(\nu_1 - 1)!(\nu_2 - 1)!} = (\nu_1 + \nu_2 - 1)\binom{\nu_1 + \nu_2 - 2}{\nu_1 - 1}.$$

La distribución $\beta(\nu_1, \nu_2)$ se puede obtener como la distribución del cociente $X_1/(X_1 + X_2)$ donde $X_1 \sim \Gamma(\nu_1, 1)$ y $X_2 \sim \Gamma(\nu_2, 1)$.

Notar que $\beta(1, 1) = \mathcal{U}(0, 1)$.

4. Familia Binomial, Binomial(n, p). Decimos que X tiene distribución Binomial de parámetros $n \in \mathbb{N}$ y $0 < p < 1$ cuando su f.p. está dada por

$$f(x|n, p) = \binom{n}{x}(1-p)^{n-x}p^x, \qquad x = 0, 1, \ldots, n.$$

Vale que $\mathbb{E}[X] = np$ y $\mathbb{V}(X) = np(1-p)$.

Un caso particular de la familia Binomial es la familia Bernoulli de parámetro p, Bernoulli(p)=Binomial$(1, p)$.

5. Familia Pascal, Pascal(n, p). Decimos que X tiene distribución Pascal de parámetros $n \in \mathbb{N}$ y $0 < p < 1$ cuando su f.p. está dada por

$$f(x|n, p) = \binom{x-1}{n-1}p^n(1-p)^{x-n}, \qquad x = n, n+1, \ldots.$$

Vale que $\mathbb{E}[X] = n/p$ y $\mathbb{V}(X) = n(1-p)/p^2$.

6. Familia Poisson, Poisson(λ). Decimos que X tiene distribución Poisson de parámetro $\lambda > 0$ cuando su f.p. está dada por

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x = 0, 1, \ldots.$$

Vale que $E[X] = \lambda$ y $\mathbb{V}(X) = \lambda$.

2. Estimadores

El punto de partida de la investigación estadística está constituido por una muestra aleatoria, $\mathbf{X} = (X_1, \ldots, X_n)$, de la distribución desconocida F perteneciente a una familia paramétrica de distribuciones $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$[1]. Como las distribuciones de la familia $\mathcal{F}$ son *distinguibles* lo que se quiere saber es cuál es el parámetro $\theta \in \Theta$ que corresponde a la distribución F. En otras palabras, se quiere hallar $\theta \in \Theta$ tal que $F = F_\theta$.

Formalmente, "cualquier" función, $\hat{\theta} := \hat{\theta}(\mathbf{X})$, de la muestra aleatoria $\mathbf{X}$ que no depende de parámetros desconocidos se denomina una *estadística*.

Ejemplo 2.1. Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de la variable aleatoria X con función de distribución F_θ. Ejemplos de estadísticas son

(i) $X_{(1)} = \text{mín}(X_1, \ldots, X_n)$,

(ii) $X_{(n)} = \text{máx}(X_1, \ldots, X_n)$,

(iii) $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$,

(iv) $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

En (i) y (ii), mín$(\cdot)$ y máx$(\cdot)$ denotan, respectivamente, el mínimo y el máximo muestrales observados. Por otro lado, $\bar{X}$ y $\hat{\sigma}^2$ denotan, respectivamente, la media y la varianza muestrales. □

Cualquier estadística que asuma valores en el conjunto paramétrico Θ de la familia de distribuciones $\mathcal{F}$ se denomina un *estimador puntual* para θ. El adjetivo puntual está puesto para distinguirla de las *estimaciones por intervalo* que veremos más adelante.

En muchas situaciones lo que interesa es estimar una función $g(\theta)$. Por ejemplo, cuando se considera una muestra aleatoria $\mathbf{X}$ de una variable $X \sim \mathcal{N}(\mu, \sigma^2)$ donde μ y σ^2 son desconocidos entonces $\theta = (\mu, \sigma^2)$ y el conjunto de parámetros es $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R} \text{ y } \sigma^2 > 0\}$. Si el objetivo es estimar solamente μ, entonces $g(\theta) = \mu$.

Definición 2.2. Cualquier estadística que solamente asuma valores en el conjunto de los posibles valores de $g(\theta)$ es un *estimador para* $g(\theta)$.

Uno de los grandes problemas de la estadística es construir estimadores razonables para el parámetro desconocido θ o para una función $g(\theta)$. Existen diversos métodos para elegir entre todos los estimadores posibles de θ. Cada elección particular del estimador depende de ciertas propiedades que se consideran "deseables" para la estimación.

[1] **Notación.** Si $\mathcal{F}$ es una familia de distribuciones F_θ con "densidades" $f(x|\theta)$, $\theta \in \Theta$, escribimos

$$\mathbb{P}_\theta(X \in A) = \int_A f(x|\theta)dx \qquad \text{y} \qquad \mathbb{E}_\theta[r(X)] = \int r(x)f(x|\theta)dx$$

El subíndice θ indica que la probabilidad o la esperanza es con respecto a $f(x|\theta)$. Similarmente, escribimos $\mathbb{V}_\theta$ para la varianza.

2.1. Error cuadrático medio, sesgo y varianza

Uno de los procedimientos más usados para evaluar el desempeño de un estimador es considerar su error cuadrático medio. Esta noción permite precisar el sentido que se le otorga a los enunciados del tipo "*el estimador puntual* $\hat{\theta} = \hat{\theta}(\mathbf{X})$ *está próximo de* θ".

Definición 2.3 (Error cuadrático medio)**.** El *error cuadrático medio* (ECM) de un estimador $\hat{\theta}$ para el parámetro θ se define por

$$\mathrm{ECM}(\hat{\theta}) = \mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right]. \tag{1}$$

El ECM se puede descomponer de la siguiente manera[2]

$$\mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right] = \mathbb{V}_\theta(\hat{\theta}) + \mathbb{B}_\theta^2(\hat{\theta}), \tag{2}$$

donde $\mathbb{B}_\theta(\hat{\theta}) := \mathbb{E}_\theta[\hat{\theta}] - \theta$ es el llamado *sesgo* del estimador. El primer término de la descomposición (2) describe la "variabilidad" del estimador, y el segundo el "error sistemático": $\mathbb{E}_\theta[\hat{\theta}]$ describe alrededor de *qué* valor fluctúa $\hat{\theta}$ y $\mathbb{V}_\theta(\hat{\theta})$ mide *cuánto* fluctúa.

Definición 2.4 (Estimadores insesgados)**.** Diremos que un estimador $\hat{\theta}$ es *insesgado* para el parámetro θ si

$$\mathbb{E}_\theta[\hat{\theta}] = \theta.$$

para todo $\theta \in \Theta$, o sea $\mathbb{B}_\theta(\hat{\theta}) \equiv 0$. Si $\lim_{n\to\infty} \mathbb{B}_\theta[\hat{\theta}] = 0$ para todo $\theta \in \Theta$, diremos que el estimador $\hat{\theta}$ es *asintóticamente insesgado* para θ.

Nota Bene. En el caso en que $\hat{\theta}$ es un estimador insesgado para θ, tenemos que

$$\mathrm{ECM}(\hat{\theta}) = \mathbb{V}_\theta(\hat{\theta}),$$

o sea, el error cuadrático medio de $\hat{\theta}$ se reduce a su varianza.

---

[2]La descomposición (2) se obtiene escribiendo $\hat{\theta} - \theta$ en la forma $(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)$. Desarrollando cuadrados obtenemos $(\hat{\theta} - \theta)^2 = (\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta) + \left(\mathbb{E}_\theta[\hat{\theta}] - \theta\right)^2$. El resultado se obtiene observando que la esperanza $\mathbb{E}_\theta$ de los términos cruzados $(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta)$ es igual a 0:

$$\begin{aligned}
\mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right] &= \mathbb{E}_\theta\left[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta) + \left(\mathbb{E}_\theta[\hat{\theta}] - \theta\right)^2\right] \\
&= \mathbb{E}_\theta\left[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2\right] + 0 + \left(\mathbb{E}_\theta[\hat{\theta}] - \theta\right)^2 = \mathbb{V}_\theta(\hat{\theta}) + \mathbb{B}_\theta^2(\hat{\theta}).
\end{aligned}$$

Nota Bene. Una consecuencia destacable de la descomposición (2) para grandes muestras ($n >> 1$) es la siguiente: si a medida que se aumenta el volumen de la muestra, el sesgo y la varianza del estimador $\hat{\theta}$ tienden a cero, entonces, el estimador $\hat{\theta}$ converge en media cuadrática al verdadero valor del parámetro θ.

Ejemplo 2.5 (Estimación de media)**.** Sea $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ una familia de distribuciones. Para cada $\theta \in \Theta$ designemos mediante $\mu(\theta)$ y $\sigma^2(\theta)$ la media y la varianza correspondientes a la distribución F_θ, respectivamente. Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de alguna distribución perteneciente a $\mathcal{F}$. Denotemos mediante $\bar{X}$ el promedio de la muestra:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

En lo que sigue vamos a suponer que para cada $\theta \in \Theta$, $\mu(\theta) \in \mathbb{R}$ y $\sigma^2(\theta) < \infty$. Si la muestra aleatoria proviene de la distribución F_θ, tenemos que

$$\mathbb{E}_\theta\left[\bar{X}\right] = \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_i] = \mu(\theta).$$

Por lo tanto $\bar{X}$ es un estimador insesgado para $\mu(\theta)$ y su error cuadrático medio al estimar $\mu(\theta)$ es

$$\mathrm{ECM}(\bar{X}) = \mathbb{V}_\theta\left(\bar{X}\right) = \mathbb{V}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}_\theta[X_i] = \frac{1}{n}\sigma^2(\theta).$$

□

Ejemplo 2.6 (Estimación de varianza)**.** Sea $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ una familia de distribuciones. Para cada $\theta \in \Theta$ designemos mediante $\mu(\theta)$ y $\sigma^2(\theta)$ la media y la varianza correspondientes a la distribución F_θ, respectivamente, a las que supondremos finitas. Sea $X_1, \ldots, X_n$ una muestra aleatoria de alguna distribución perteneciente a $\mathcal{F}$. Sean $\bar{X}$ y $\hat{\sigma}^2$ la media y la varianza muestrales definidas en el Ejemplo 2.1:

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{y} \qquad \hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Para analizar el sesgo de la varianza muestral conviene descomponerla de la siguiente manera:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu(\theta))^2 - (\bar{X} - \mu(\theta))^2, \tag{3}$$

cualquiera sea $\theta \in \Theta$. [3] Si la muestra aleatoria, $X_1, \ldots, X_n$, proviene de la distribución F_θ, al tomar esperanzas en ambos lados de (3) se obtiene

$$\begin{aligned} \mathbb{E}_\theta[\hat{\sigma}^2] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta\left[(X_i - \mu(\theta))^2\right] - \mathbb{E}_\theta\left[(\bar{X} - \mu(\theta))^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}_\theta(X_i) - \mathbb{V}_\theta(\bar{X}). \end{aligned} \tag{4}$$

Según el Ejemplo 2.5 $\bar{X}$ es un estimador insesgado para la media $\mu(\theta)$ y su varianza vale $\mathbb{V}_\theta(\bar{X}) = \frac{1}{n}\sigma^2(\theta)$, en consecuencia,

$$\mathbb{E}_\theta[\hat{\sigma}^2] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}_\theta(X_i) - \mathbb{V}_\theta(\bar{X}) = \sigma^2(\theta) - \frac{1}{n}\sigma^2(\theta) = \frac{n-1}{n}\sigma^2(\theta). \tag{5}$$

Esto demuestra que $\hat{\sigma}^2$ *no es un estimador insesgado* para la varianza $\sigma^2(\theta)$. La identidad $\mathbb{E}_\theta[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2(\theta)$ significa que si tomamos repetidas muestras de tamaño n y se promedian las varianzas muestrales resultantes, el promedio no se aproximará a la verdadera varianza, sino que de modo sistemático el valor será más pequeño debido al factor $(n-1)/n$. Este factor adquiere importancia en las muestras pequeñas. Si $n \to \infty$, el factor $(n-1)/n \to 1$ lo que demuestra que $\hat{\sigma}^2$ es un estimador *asintóticamente insesgado* para la varianza $\sigma^2(\theta)$.

Para eliminar el sesgo en $\hat{\sigma}^2$, basta multiplicar $\hat{\sigma}^2$ por $\frac{n}{n-1}$. De (5) sigue que

$$S^2 := \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{6}$$

es un *estimador insesgado* para la varianza. □

2.2. Comparación de estimadores

El error cuadrático medio puede usarse para comparar estimadores. Diremos que $\hat{\theta}_1$ es *mejor* que $\hat{\theta}_2$ si

$$\mathrm{ECM}(\hat{\theta}_1) \leq \mathrm{ECM}(\hat{\theta}_2), \tag{7}$$

para todo θ, con desigualdad estricta para al menos un valor de θ. En tal caso, el estimador $\hat{\theta}_2$ se dice *inadmisible*. Si existe un estimador $\hat{\theta}^*$ tal que para todo estimador $\hat{\theta}$ de θ con $\hat{\theta} \neq \hat{\theta}^*$

$$\mathrm{ECM}(\hat{\theta}^*) \leq \mathrm{ECM}(\hat{\theta}), \tag{8}$$

[3] La descomposición (3) se obtiene haciendo lo siguiente. Para cada i escribimos $(X_i - \bar{X})$ en la forma $(X_i - \mu(\theta)) - (\bar{X} - \mu(\theta))$. Desarrollando cuadrados obtenemos $(X_i - \bar{X})^2 = (X_i - \mu(\theta))^2 + (\bar{X} - \mu(\theta))^2 - 2(X_i - \mu(\theta))(\bar{X} - \mu(\theta))$. El resultado se obtiene observando que el promedio de los términos cruzados $(X_i - \mu(\theta))(\bar{X} - \mu(\theta))$ es igual a $(\bar{X} - \mu(\theta))^2$. (*Hacer la cuenta y verificarlo!*)

para todo θ, con desigualdad estricta para al menos un valor de θ, entonces $\hat{\theta}^*$ se dice *óptimo*.

Cuando la comparación se restringe a los estimadores son insesgados, el estimador óptimo, $\hat{\theta}^*$, se dice el estimador insesgado de varianza uniformemente mínima. Esta denominación resulta de observar que estimadores insesgados la relación (8) adopta la forma

$$\mathbb{V}_\theta(\hat{\theta}^*) \leq \mathbb{V}_\theta(\hat{\theta}),$$

para todo θ, con desigualdad estricta para al menos un valor de θ.

Ejemplo 2.7. Sean X_1, X_2, X_3 una muestra aleatoria de una variable aleatoria X tal que $\mathbb{E}_\theta[X] = \theta$ y $\mathbb{V}_\theta(X) = 1$. Consideremos los estimadores

$$\bar{X} = \frac{X_1 + X_2 + X_3}{3} \qquad \text{y} \qquad \hat{\theta} = \frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3.$$

Según el Ejemplo 2.5 $\mathbb{E}_\theta[\bar{X}] = \theta$ y $\mathbb{V}_\theta(\bar{X}) = \frac{1}{3}$. Tenemos también que

$$\mathbb{E}_\theta[\hat{\theta}] = \frac{1}{2}\mathbb{E}_\theta[X_1] + \frac{1}{4}\mathbb{E}_\theta[X_2] + \frac{1}{4}\mathbb{E}_\theta[X_3] = \frac{1}{2}\theta + \frac{1}{4}\theta + \frac{1}{4}\theta = \theta$$

y

$$\mathbb{V}_\theta(\hat{\theta}) = \frac{1}{4}\mathbb{V}_\theta(X_1) + \frac{1}{16}\mathbb{V}_\theta(X_2) + \frac{1}{16}\mathbb{V}_\theta(X_3) = \frac{1}{4} + \frac{1}{16} + \frac{1}{16} = \frac{6}{16}.$$

Como $\bar{X}$ y $\hat{\theta}$ son insesgados, resulta que $\bar{X}$ es mejor que $\hat{\theta}$, pues $\mathbb{V}_\theta(\bar{X}) < \mathbb{V}_\theta(\hat{\theta})$ para todo θ. □

Ejemplo 2.8. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable aleatoria $X \sim \mathcal{U}(0, \theta)$. Vamos a considerar $\hat{\theta}_1 = 2\bar{X}$ y $\hat{\theta}_2 = X_{(n)}$ como estimadores para θ y estudiaremos su comportamiento. Como $\mathbb{E}_\theta[X] = \theta/2$ y $\mathbb{V}_\theta(X) = \theta^2/12$, tenemos que

$$\mathbb{E}_\theta[\hat{\theta}_1] = \mathbb{E}_\theta[2\bar{X}] = \theta \qquad \text{y} \qquad \mathbb{V}_\theta(\hat{\theta}_1) = \frac{\theta^2}{3n}. \tag{9}$$

Por lo tanto, $\hat{\theta}_1$ es un estimador insesgado para θ. En consecuencia,

$$\text{ECM}(\hat{\theta}_1) = \mathbb{V}_\theta(\hat{\theta}_1) = \frac{\theta^2}{3n}. \tag{10}$$

Por otro lado, la función densidad de $X_{(n)}$ está dada por $f_\theta(x) = \frac{nx^{n-1}}{\theta^n}\mathbf{1}\{0 < x < \theta\}$, de donde se deduce que

$$\mathbb{E}_\theta[X_{(n)}] = \frac{n}{n+1}\,\theta \qquad \text{y} \qquad \mathbb{V}_\theta(X_{(n)}) = \frac{n\theta^2}{(n+1)^2(n+2)}. \tag{11}$$

Por lo tanto, $\hat{\theta}_2$ es un estimador asintóticamente insesgado para θ. Combinando las identidades (11) en (2), obtenemos

$$\begin{aligned} \mathrm{ECM}(\hat{\theta}_2) &= \mathbb{V}_\theta(\hat{\theta}_2) + \mathbb{B}^2_\theta(\hat{\theta}_2) = \frac{n\theta^2}{(n+1)^2(n+2)} + \left(\frac{n}{n+1}\,\theta - \theta\right)^2 \\ &= \frac{n\theta^2}{(n+1)^2(n+2)} + \frac{\theta^2}{(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)}. \end{aligned} \tag{12}$$

Es fácil, pero tedioso, ver que $\mathrm{ECM}(\hat{\theta}_2) < \mathrm{ECM}(\hat{\theta}_1)$ para todo θ y todo n. Por lo tanto, $X_{(n)}$ es mejor que $2\bar{X}$ para todo θ y todo n.

□

2.3. Consistencia

Lo mínimo que se le puede exigir a un estimador puntual, $\hat{\theta}(X_1, \ldots, X_n)$, es que, en algún sentido, se aproxime al verdadero valor del parámetro cuando el volumen de la muestra aumenta. En otras palabras, si $\theta \in \Theta$ es tal que $F = F_\theta$ y $X_1, X_2, \ldots$ es una sucesión de variables aleatorias independientes cada una con distribución F, en algún sentido, debe ocurrir que

$$\hat{\theta}(X_1, \ldots, X_n) \to \theta,$$

cuando $n \to \infty$.

Por ejemplo, es deseable que el estimador $\hat{\theta}$ tenga la siguiente propiedad, llamada *consistencia débil*: para cada $\epsilon > 0$ debe cumplir que

$$\lim_{n\to\infty} \mathbb{P}_\theta(|\hat{\theta}(X_1, \ldots, X_n) - \theta| > \epsilon) = 0. \tag{13}$$

Más exigente, es pedirle que tenga la siguiente propiedad, llamada *consistencia fuerte*:

$$\mathbb{P}_\theta\left(\lim_{n\to\infty} \hat{\theta}(X_1, \ldots, X_n) = \theta\right) = 1. \tag{14}$$

Normalidad asintótica. También se le puede pedir una propiedad similar a la del teorema central límite, llamada *normalidad asintótica*: existe $\sigma = \sigma(\theta) > 0$ tal que

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\frac{\sqrt{n}(\hat{\theta}(X_1, \ldots, X_n) - \theta)}{\sigma} \leq x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{15}$$

Nota Bene. Los problemas de consistencia y normalidad asintótica están relacionados con las leyes de los grandes números y el teorema central de límite. El siguiente ejemplo muestra dicha relación para el caso en que se quiere estimar la media de una distribución.

Ejemplo 2.9 (Estimación de media)**.** Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de una variable aleatoria cuya distribución pertenece a una familia $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$. Sean

$\mu(\theta)$ y $\sigma^2(\theta)$ la media y la varianza correspondientes a la distribución F_θ, respectivamente. Aplicando la desigualdad de Chebychev a $\bar{X}$ se obtiene que para cada $\epsilon > 0$

$$\mathbb{P}_\theta\left(\left|\bar{X} - \mu(\theta)\right| > \epsilon\right) \leq \frac{\mathbb{V}_\theta(\bar{X})}{\epsilon^2} = \frac{1}{n}\left(\frac{\sigma^2(\theta)}{\epsilon^2}\right) \to 0,$$

cuando $n \to \infty$.

Hasta aquí, lo único que hicimos es volver a demostrar la ley débil de los grandes números. Lo que queremos subrayar es que *en el contexto de la estimación de parámetros, la ley débil de los grandes números significa que el promedio de la muestra, $\bar{X}$, es un estimador débilmente consistente para la la media de la distribución, $\mu(\theta)$.*

La consistencia fuerte del promedio, como estimador para la media es equivalente a la *Ley fuerte de los grandes números* que afirma que: *Si $X_1, X_2, \ldots$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas y si existe $\mathbb{E}[X_i] = \mu$, entonces*

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X} = \mu\right) = 1.$$

La normalidad asintótica es equivalente al teorema central del límite. □

Nota Bene. De todas las propiedades de convergencia la consistencia débil es la mas simple, en el sentido de que puede establecerse con unas pocas herramientas técnicas. Para verificar la consistencia débil del promedio para estimar la media solamente usamos la desigualdad de Chebychev y las propiedades de la media y la varianza. El razonamiento utilizado en el Ejemplo 2.9 se puede extender un poco más allá.

Teorema 2.10. *Sea $\hat{\theta}$ un estimador de θ basado en una muestra aleatoria de volumen n. Si $\hat{\theta}$ es* asintóticamente insesgado *y su varianza tiende a cero, entonces $\hat{\theta}$ es débilmente consistente.*

Demostración. El resultado se obtiene usando la desigualdad de Chebychev y la identidad (2):

$$\mathbb{P}_\theta\left(\left|\hat{\theta} - \theta\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right] = \frac{1}{\epsilon^2}\left(\mathbb{V}_\theta(\hat{\theta}) + \mathbb{B}^2_\theta(\hat{\theta})\right) \to 0.$$

□

3. Método de máxima verosimilitud

El método de máxima verosimilitud es un "método universal" para construir estimadores puntuales. Su base intuitiva es la siguiente: *si al realizar un experimento aleatorio se observa un resultado, este debe tener alta probabilidad de ocurrir.*

Para hacer más precisa esa base intuitiva consideremos una muestra aleatoria, $\mathbf{X} = (X_1, \ldots, X_n)$, de una variable aleatoria discreta X con función de probabilidad $f(x|\theta)$,

$\theta \in \Theta$, donde Θ es el espacio paramétrico. La probabilidad de observar los resultados $X_1 = x_1, \ldots, X_n = x_n$ se calcula del siguiente modo:

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^{n} f(x_i|\theta). \tag{16}$$

Si los resultados observables deben tener una alta probabilidad de ocurrir y observamos que $X_1 = x_1, \ldots, X_n = x_n$, entonces lo razonable sería elegir entre todos los parámetros posibles, $\theta \in \Theta$, aquél (o aquellos) que maximicen (16). En consecuencia, se podría estimar θ como el valor (o los valores) de θ que hace máxima la probabilidad $\prod_{i=1}^{n} f(x_i|\theta)$.

3.1. Estimador de máxima verosimilitud (emv)

Definición 3.1 (EMV)**.** Sea X una variable aleatoria cuya distribución pertenece a la familia paramétrica $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$. Un *estimador de máxima verosimilitud* de θ, basado en los valores $\mathbf{x} = (x_1, \ldots, x_n)$ de una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$, es un valor $\hat{\theta}_{mv} \in \Theta$ que maximiza la función de verosimilitud

$$L(\theta|\mathbf{x}) := \prod_{i=1}^{n} f(x_i|\theta), \tag{17}$$

donde, dependiendo de la naturaleza de las distribuciones de la familia $\mathcal{F}$, $f(x|\theta)$ es la función de probabilidad o la función densidad de probabilidades de X.

Sobre la notación. Para destacar que el valor del estimador de máxima verosimilitud depende de los valores observados, $\mathbf{x} = (x_1, \ldots, x_n)$, en lugar de $\hat{\theta}_{mv}$ escribiremos $\hat{\theta}_{mv}(\mathbf{x})$:

$$\hat{\theta}_{mv} = \hat{\theta}_{mv}(\mathbf{x}) := \arg\max_{\theta \in \Theta} L(\theta|\mathbf{x}). \tag{18}$$

Ejemplo 3.2. Supongamos que tenemos una moneda que puede ser equilibrada o totalmente cargada para que salga cara. Lanzamos la moneda n veces y registramos la sucesión de caras y cecas. Con esa información queremos estimar qué clase de moneda tenemos.

Cada lanzamiento de la moneda se modela con una variable aleatoria X con distribución Bernoulli(θ), donde θ es la probabilidad de que la moneda salga cara. El espacio paramétrico es el conjunto $\Theta = \{1/2, 1\}$.

El estimador de máxima verosimilitud para θ, basado en los valores $\mathbf{x} = (x_1, \ldots, x_n)$ de una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ de la variable X, es el valor de $\hat{\theta}_{mv}(\mathbf{x}) \in \Theta = \{1/2, 1\}$ que maximiza la función de verosimilitud $L(\theta|\mathbf{x})$. Para encontrarlo comparamos los valores de la función de verosimilitud $L(1/2|\mathbf{x})$ y $L(1|\mathbf{x})$:

$$L(1/2|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|1/2) = (1/2)^n, \qquad L(1|\mathbf{x}) = \mathbf{1}\left\{\sum_{i=1}^{n} x_i = n\right\}.$$

En consecuencia, el estimador de máxima verosimilitud para θ, basado en los valores $\mathbf{x} = (x_1, \ldots, x_n)$ de una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ es

$$\hat{\theta}_{mv}(\mathbf{x}) = \frac{1}{2}\mathbf{1}\left\{\sum_{i=1}^{n} x_i < n\right\} + \mathbf{1}\left\{\sum_{i=1}^{n} x_i = n\right\}.$$

Por lo tanto, el estimador de máxima verosimilitud para θ basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ es

$$\hat{\theta}_{mv}(\mathbf{X}) = \frac{1}{2}\mathbf{1}\left\{\sum_{i=1}^{n} X_i < n\right\} + \mathbf{1}\left\{\sum_{i=1}^{n} X_i = n\right\}.$$

Por ejemplo, si en 10 lanzamientos de la moneda se observaron 10 caras, el estimador de máxima verosimilitud para θ es $\hat{\theta}_{mv} = 1$; en cambio si se observaron 8 caras y 2 cecas, el estimador de máxima verosimilitud es $\hat{\theta}_{mv} = 1/2$. □

Ejemplo 3.3. Sea X una variable aleatoria con función densidad dada por

$$f(x|\theta) = \frac{1}{2}(1 + \theta x)\mathbf{1}\{x \in [-1, 1]\}, \qquad \theta \in [-1, 1].$$

Supongamos que queremos hallar el estimador de máxima verosimilitud para θ basado en la realización de una muestra aleatoria tamaño 1, X_1. Si se observa el valor x_1, la función de verosimilitud adopta la forma

$$L(\theta|x_1) = \frac{1}{2}(1 + \theta x_1)$$

El gráfico de $L(\theta|x_1)$ es un segmento de recta de pendiente x_1. Como se trata de una recta el máximo se alcanza en alguno de los extremos del intervalo $\Theta = [-1, 1]$:

1. si $x_1 < 0$, el máximo se alcanza en $\theta = -1$,
2. si $x_1 = 0$, el máximo se alcanza en cualquiera de los valores del intervalo Θ,
3. si $x_1 > 0$, el máximo se alcanza en $\theta = 1$.

Abusando de la notación tenemos que

$$\hat{\theta}_{mv}(x_1) = -\mathbf{1}\{x_1 < 0\} + \Theta\mathbf{1}\{x_1 = 0\} + \mathbf{1}\{x_1 > 0\}.$$

Por lo tanto,

$$\hat{\theta}_{mv}(X_1) = -\mathbf{1}\{X_1 < 0\} + \Theta\mathbf{1}\{X_1 = 0\} + \mathbf{1}\{X_1 > 0\}.$$

□

Ejemplo 3.4. Sea X una variable aleatoria con función densidad dada por

$$f(x|\theta) = \frac{1}{2}(1+\theta x)\mathbf{1}\{x \in [-1,1]\}, \qquad \theta \in [-1,1].$$

Supongamos que una muestra aleatoria de tamaño 2 arrojó los valores $1/2$ y $1/4$ y con esa información queremos hallar el estimador de máxima verosimilitud para θ. La función de verosimilitud adopta la forma

$$L(\theta|1/2,1/4) = \frac{1}{4}\left(1+\theta\frac{1}{2}\right)\left(1+\theta\frac{1}{4}\right),$$

y su gráfico es un segmento de parábola "cóncava" cuyas raíces son -4 y -2. Por lo tanto, $\hat{\theta}_{mv}(1/2,1/4) = 1$.

Supongamos ahora que una muestra aleatoria de tamaño 2 arrojó los valores $1/2$ y $-1/4$ y con esa información queremos hallar el estimador de máxima verosimilitud para θ. La función de verosimilitud adopta la forma

$$L(\theta|1/2,-1/3) = \frac{1}{4}\left(1+\theta\frac{1}{2}\right)\left(1-\theta\frac{1}{3}\right),$$

y su gráfico es un segmento de parábola "convexa" cuyas raíces son -2 y 3. Por lo tanto, $\hat{\theta}_{mv}(1/2,-1/3) = 0.5$. □

3.2. Cálculo del emv para familias regulares

Sea $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ una familia paramétrica de distribuciones y sea $\{f(x|\theta) : \theta \in \Theta\}$ la familia de funciones de densidad (o de probabilidad) asociada. Diremos que la familia $\mathcal{F}$ es *regular* si satisface las siguientes condiciones:

1. El conjunto paramétrico $\Theta \subset \mathbb{R}^d$ es abierto.

2. El soporte de las funciones $f(x|\theta)$ no depende del parámetro. Esto es, existe un conjunto $\mathbb{S}$ tal que $\text{sop}f(\cdot|\theta) := \{x \in \mathbb{R} : f(x|\theta) > 0\} = \mathbb{S}$ para todo $\theta \in \Theta$.

3. Para cada $x \in \mathbb{S}$, la función $f(x|\theta)$ tiene derivadas parciales respecto de todas las componentes θ_j, $j = 1,\ldots,d$.

Supongamos ahora que $\mathbf{X} = (X_1,\ldots,X_n)$ es una muestra aleatoria de tamaño n de una variable aleatoria X con función de densidad (o de probabilidad) $f(x|\theta)$, $\theta \in \Theta$, perteneciente a una familia regular de distribuciones. Debido a que la familia es regular cada uno de los valores observados pertenece al soporte común de las funciones $f(x|\theta)$: $\mathbf{x} = (x_1,\ldots,x_n) \in \mathbb{S}^n$. Por lo tanto, cualesquiera sean los valores observados, $\mathbf{x} = (x_1,\ldots,x_n)$, vale que

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) > 0.$$

Esto habilita a tomar logaritmos y utilizar la propiedad *"el logaritmo del producto es igual a la suma de los logaritmos"*. En consecuencia, para cada $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{S}^n$, la función $\log L(\theta|\mathbf{x})$ está bien definida y vale que

$$\log L(\theta|\mathbf{x}) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta). \tag{19}$$

Como el logaritmo natural $\log(\cdot)$ es una función monótona creciente, maximizar la función de verosimilitud $L(\theta|\mathbf{x})$ será equivalente a maximizar $\log L(\theta|\mathbf{x})$. La ventaja de maximizar el logaritmo de la función de verosimilitud es que, bajo las condiciones de regularidad enunciadas previamente, los productos se convierten en sumas, aligerando considerablemente el trabajo de cómputo del EMV ya que el EMV debe verificar el sistema de ecuaciones

$$\frac{\partial \log L(\theta|\mathbf{x})}{\partial \theta_j} = 0 \qquad j = 1, \ldots, d. \tag{20}$$

En vista de (19) el sistema de ecuaciones (20) se transforma en

$$\sum_{i=1}^{n} \frac{\partial \log f(x_i|\theta)}{\partial \theta_j} = 0, \qquad j = 1, \ldots, d. \tag{21}$$

Por este camino llegamos al siguiente resultado que provee la herramienta adecuada para el cálculo del EMV.

Lema 3.5. Sea X una variable aleatoria con función de densidad (o de probabilidad) $f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, perteneciente a una familia regular de distribuciones. El estimador de máxima verosimilitud de θ, basado en los valores $\mathbf{x} = (x_1, \ldots, x_n)$ de una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$, es solución del siguiente sistema de ecuaciones:

$$\sum_{i=1}^{n} \psi_j(\theta|x_i) = 0 \qquad j = 1, \ldots, d, \tag{22}$$

donde, para cada $x \in \mathbb{S}$, la funciones de θ, $\psi_j(\theta|x)$, $j = 1, \ldots, d$, se definen por

$$\psi_j(\theta|x) := \frac{\partial \log f(x|\theta)}{\partial \theta_j}. \tag{23}$$

Nota Bene. Por supuesto que las condiciones (22) son necesarias pero no suficientes para que θ sea un máximo. Para asegurarse que θ es un máximo deberán verificarse las condiciones de segundo orden. Además debe verificarse que no se trata de un máximo relativo sino absoluto.

Nota Bene. Si la función de densidad (o de probabilidad) $f(x|\theta)$ de la variable aleatoria X pertenece a una familia regular *uniparamétrica* de distribuciones, i.e., cuando el espacio paramétrico Θ es un subconjunto de la recta real $\mathbb{R}$, el sistema de ecuaciones (22) se reduce a una sola ecuación, denominada la *ecuación de verosimilitud*,

$$\sum_{i=1}^{n} \psi(\theta|x_i) = 0, \tag{24}$$

donde, para cada $x \in \mathbb{S}$, la función de θ, $\psi(\theta|x)$, se define por

$$\psi(\theta|x) := \frac{\partial \log f(x|\theta)}{\partial \theta}. \tag{25}$$

Ejemplo 3.6 (Distribuciones de Bernoulli)**.** Es fácil ver que la familia de distribuciones Bernoulli(θ), $\theta \in (0,1)$, es una familia uniparamétrica regular con funciones de probabilidad de la forma $f(x|\theta) = (1-\theta)^{1-x}\theta^x$, $x = 0, 1$. En consecuencia, para encontrar el estimador de máxima verosimilitud para θ basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ podemos usar el resultado del Lema 3.5.

En primer lugar hallamos la expresión de la función $\psi(\theta|x) = \frac{\partial \log f(x|\theta)}{\partial \theta}$. Observando que

$$\log f(x|\theta) = \log(1-\theta)^{1-x}\theta^x = (1-x)\log(1-\theta) + x\log\theta,$$

y derivando respecto de θ obtenemos

$$\psi(\theta|x) = \frac{1}{1-\theta}(x-1) + \frac{1}{\theta}x$$

Por lo tanto, la ecuación de verosimilitud (24) adopta la forma

$$\frac{1}{1-\theta}\sum_{i=1}^{n}(x_i - 1) + \frac{1}{\theta}\sum_{i=1}^{n} x_i = 0. \tag{26}$$

Un poco de álgebra muestra que para cada pareja $a \neq b$ vale que:

$$\frac{1}{1-\theta}a + \frac{1}{\theta}b = 0 \Leftrightarrow \theta = \frac{b}{b-a}. \tag{27}$$

Sigue de (27), poniendo $a = \sum_{i=1}^{n}(x_i - 1) = \sum_{i=1}^{n} x_i - n$ y $b = \sum_{i=1}^{n} x_i$, que la solución de la ecuación (26) es

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Con un poco más de trabajo, se puede verificar que dicha solución maximiza el logaritmo de la verosimilitud.

En resumen, si $\mathbf{x} = (x_1, \ldots, x_n)$ son los valores observados de una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$, el estimador de máxima verosimilitud para θ es el promedio (o media) muestral

$$\hat{\theta}_{mv} = \hat{\theta}_{mv}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Por lo tanto, *el estimador de máxima verosimilitud para θ, basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ de una variable con distribución Bernoulli(θ), es el promedio muestral*

$$\hat{\theta}_{mv}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{28}$$

□

Nota Bene. *El estimador de máxima verosimilitud* para θ, basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$, de una variable aleatoria con distribución Bernoulli(θ),

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

es una variable aleatoria. Subrayamos este hecho para que no se pierda de vista que los estimadores puntuales son funciones de la muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ y por lo tanto son variables aleatorias. En el Ejemplo 3.6, el parámetro θ es la media de la distribución que produce la muestra y el estimador de máxima verosimilitud para θ es el promedio muestral. Por lo tanto, $\hat{\theta}_{mv}$ es un estimador *insesgado, consistente y asintóticamente normal*.

Nota Bene. Si la muestra aleatoria arrojó los valores $1, 1, \ldots, 1$, es fácil ver que $\hat{\theta}_{mv} = 1$, en cambio si arrojó $0, 0, \ldots, 0$ resulta que $\hat{\theta}_{mv} = 0$. Estos resultados también coinciden con el promedio de los valores observados. Por lo tanto, el resultado obtenido en (28) se puede extender al caso en que $\Theta = [0, 1]$.

Ejemplo 3.7 (Distribuciones de Bernoulli)**.** Bajo el supuesto de que los valores de la secuencia

$$0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0. \tag{29}$$

fueron arrojados por una muestra aleatoria de tamaño 20 de una variable aleatoria $X \sim$ Bernoulli(θ), el estimador de máxima verosimilitud arrojará como resultado la siguiente estimación para el parámetro θ:

$$\hat{\theta}_{mv}(0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0) = \frac{11}{20} = 0.55$$

Con esta estimación podríamos decir que la ley que produce esos valores es la distribución de Bernoulli (0.55). Por lo tanto, si queremos "reproducir" el generador de números aleatorios que produjo esos resultados, debemos simular números aleatorios con distribución de Bernoulli de parámetro 0.55. □

Ejemplo 3.8 (Distribuciones normales con varianza conocida)**.** Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de una variable aleatoria $X \sim \mathcal{N}(\theta, \sigma^2)$, con varianza $\sigma^2 > 0$ conocida y media $\theta \in \mathbb{R}$. La familia de distribuciones normales $\mathcal{N}(\theta, \sigma^2)$, $\theta \in \mathbb{R}$, es una familia regular uniparamétrica con densidades de la forma

$$f(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}.$$

Usando el resultado del Lema 3.5 se puede ver que *el estimador de máxima verosimilitud para θ es*

$$\hat{\theta}_{mv}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

En efecto, como

$$\psi(\theta|x) = \frac{\partial \log f(x|\theta)}{\partial\theta} = \frac{x-\theta}{\sigma^2}$$

la ecuación de verosimilitud (24) equivale a

$$\sum_{i=1}^{n}(x_i - \theta) = 0.$$

El resultado se obtiene despejando θ. □

Ejemplo 3.9 (Distribuciones normales)**.** La familia de distribuciones normales

$$\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R},\, \sigma^2 > 0\}$$

es una familia regular con parámetro bidimensional $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Para encontrar el estimador de máxima verosimilitud del parámetro (μ, σ^2) basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ usaremos los resultados del Lema 3.5. La densidad de cada variable X es

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

con lo cual

$$\log f(x|\mu, \sigma^2) = \log(2\pi)^{-\frac{1}{2}} - \frac{1}{2}\log\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}.$$

En consecuencia,

$$\frac{\partial \log f(x|\mu, \sigma^2)}{\partial\mu} = \frac{x-\mu}{\sigma^2}$$

y

$$\frac{\partial \log f(x|\mu, \sigma^2)}{\partial\sigma^2} = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2(\sigma^2)^2}.$$

Luego el sistema de ecuaciones (22) se transforma en el sistema

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

que tiene como solución

$$\begin{aligned} \mu &= \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2. \end{aligned}$$

Se puede comprobar que en ese punto de coordenadas (μ, σ^2) se alcanza el máximo absoluto de la función $\log L(\mu, \sigma^2|\mathbf{x})$.

Resumiendo, cuando la muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ arroja los valores $\mathbf{x} = (x_1, \ldots, x_n)$, el estimador de máxima verosimilitud para (μ, σ^2) es el punto del conjunto paramétrico $\mathbb{R} \times (0, \infty)$ cuyas coordenadas son el promedio y la varianza muestrales: $\hat{\mu}_{mv}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ y $\widehat{\sigma^2}_{mv}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

Por lo tanto, el *estimador de máxima verosimilitud* para (μ, σ^2), basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ de variables normales, $\mathcal{N}(\mu, \sigma^2)$, es el punto en $\mathbb{R} \times (0, \infty)$ de coordenadas aleatorias

$$\hat{\mu}_{mv}(\mathbf{X}) = \bar{X}, \qquad \widehat{\sigma^2}_{mv}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2. \tag{30}$$

□

3.2.1. Familias exponenciales

Muchos modelos estadísticos pueden considerarse como casos particulares de una familia más general de distribuciones.

Definición 3.10 (Familias exponenciales)**.** Decimos que la distribución de una variable aleatoria X pertenece a una *familia exponencial unidimensional* de distribuciones, si podemos escribir su función de probabilidad o su función densidad como

$$f(x|\theta) = e^{a(\theta)T(x)+b(\theta)+S(x)}, \qquad x \in \mathbb{S}, \tag{31}$$

donde, a y b son funciones de θ; T y S son funciones de x y $\mathbb{S}$ no depende de θ.

Nota Bene. Si las funciones a y b son derivables y el espacio paramétrico Θ es abierto, las densidades (31) constituyen una familia regular uniparamétrica y en consecuencia, para encontrar el estimador de máxima verosimilitud de θ, basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$, se puede usar el resultado del Lema 3.5.

Debido a que el logaritmo de la densidad (31) es

$$\log f(x|\theta) = a(\theta)T(x) + b(\theta) + S(x)$$

tenemos que

$$\psi(\theta|x) = \frac{\partial \log f(x|\theta)}{\partial \theta} = a'(\theta)T(x) + b'(\theta)$$

y en consecuencia, la ecuación de verosimilitud (24) adopta la forma

$$a'(\theta)\sum_{i=1}^{n} T(x_i) + nb'(\theta) = 0.$$

Por lo tanto, el estimador de máxima verosimilitud para θ satisface la ecuación

$$\frac{-b'(\theta)}{a'(\theta)} = \frac{1}{n}\sum_{i=1}^{n} T(x_i). \tag{32}$$

Ejemplo 3.11 (Distribuciones exponenciales)**.** Sea X una variable aleatoria con distribución Exponencial(λ), $\lambda > 0$. Podemos escribir

$$f(x|\lambda) \;\; = \;\; \lambda e^{-\lambda x} = e^{-\lambda x + \log \lambda}$$

Por lo tanto, la distribución de X pertenece a una familia exponencial unidimensional con $a(\lambda) = -\lambda$, $b(\lambda) = \log \lambda$, $T(x) = x$, $S(x) = 0$ y $\mathbb{S} = (0, \infty)$. La ecuación de verosimilitud (32) adopta la forma

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{33}$$

cuya solución es $\lambda = 1/\bar{x}$. Se puede verificar que el valor de λ así obtenido maximiza el logaritmo de la verosimilitud.

Si la muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ arrojó los valores $\mathbf{x} = (x_1, \ldots, x_n)$, el estimador de máxima verosimilitud para λ es

$$\hat{\lambda}_{mv}(\mathbf{x}) = (\bar{x})^{-1}.$$

Por lo tanto, el *estimador de máxima verosimilitud* para λ, basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ de variables con distribución Exponencial(λ), es

$$\hat{\lambda}_{mv}(\mathbf{X}) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1}.$$

□

Ejemplo 3.12 (Distribuciones normales con media conocida)**.** Sea X una variable aleatoria con distribución normal $\mathcal{N}(\mu, \sigma^2)$, donde la media μ es conocida y la varianza $\sigma^2 > 0$. Podemos escribir

$$f(x|\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = e^{-\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2}\log\sigma^2 - \log\sqrt{2\pi}}$$

Por lo tanto, la distribución de X pertenece a una familia exponencial unidimensional con $a(\sigma^2) = -\frac{1}{2\sigma^2}$, $b(\sigma^2) = -\frac{1}{2}\log\sigma^2$, $T(x) = (x-\mu)^2$, $S(x) = -\log\sqrt{2\pi}$ y $\mathbb{S} = \mathbb{R}$. La ecuación de verosimilitud (32) adopta la forma

$$\frac{1/2\sigma^2}{1/2(\sigma^2)^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{34}$$

cuya solución es $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$. Se puede verificar que el valor de σ^2 así obtenido maximiza el logaritmo de la verosimilitud.

Si la muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ arrojó los valores $\mathbf{x} = (x_1, \ldots, x_n)$, el estimador de máxima verosimilitud para σ^2 es

$$\widehat{\sigma^2}_{mv}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Por lo tanto, el *estimador de máxima verosimilitud* para σ^2, basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ de variables con distribución $\mathcal{N}(\mu, \sigma^2)$, es

$$\widehat{\sigma^2}_{mv}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

□

3.2.2. Malas noticias!

"Esta calle es más angosta de lo que pensás".
(Proverbio Zen)

Ejemplo 3.13 (Fiabilidad)**.** Sea $T_1, \ldots, T_n$ una muestra aleatoria del tiempo de duración sin fallas de una máquina cuya función intensidad de fallas es $\lambda(t) = \beta t^{\beta-1}\mathbf{1}\{t > 0\}$, donde el parámetro de "desgaste" $\beta > 0$ es desconocido. La densidad de cada tiempo T es

$$f(t|\beta) = \beta t^{\beta-1} e^{-t^\beta}\mathbf{1}\{t > 0\} \tag{35}$$

Observando que

$$\log f(t|\beta) = \log\beta + (\beta - 1)\log t - t^\beta$$

y derivando respecto de β se obtiene

$$\frac{\partial \log f(x|\beta)}{\partial \beta} = \frac{1}{\beta} + \log t - t^{\beta} \log t.$$

Por lo tanto, la ecuación de verosimilitud (24) adopta la forma

$$\frac{n}{\beta} + \sum_{i=1}^{n} \log t_i - \sum_{i=1}^{n} t_i^{\beta} \log t_i = 0 \tag{36}$$

La mala noticia es que la ecuación (36) no tiene una solución analítica explícita. □

El ejemplo anterior muestra que en algunos casos la ecuación de verosimilitud no presenta solución analítica explícita. En tales casos, los estimadores de máxima verosimilitud pueden obtenerse mediante métodos numéricos.

Método de Newton-Raphson. El método de Newton-Raphson es un procedimiento iterativo para obtener una raíz de una ecuación

$$g(\theta) = 0, \tag{37}$$

donde $g(\cdot)$ es una función suave. La idea es la siguiente: supongamos que θ es una raíz de la ecuación (37). Desarrollando $g(\cdot)$ en serie de Taylor en torno de un punto θ_0, obtenemos que

$$g(\theta) \approx g(\theta_0) + (\theta - \theta_0) g'(\theta_0).$$

En consecuencia, si θ_0 está cerca de una raíz θ de la ecuación (37), debería ocurrir lo siguiente

$$\theta \approx \theta_0 - \frac{g(\theta_0)}{g'(\theta_0)}. \tag{38}$$

De la ecuación (38) obtenemos el procedimiento iterativo

$$\theta_{j+1} = \theta_j - \frac{g(\theta_j)}{g'(\theta_j)} \tag{39}$$

que se inicia con un valor θ_0 y produce un nuevo valor θ_1 a partir de (39) y así siguiendo, hasta que el proceso se estabilice, o sea, hasta que $|\theta_{j+1} - \theta_j| < \epsilon$ para un $\epsilon > 0$ "pequeño" y prefijado.

Ejemplo 3.14 (Continuación del Ejemplo 3.13)**.** Para resolver la ecuación (36) usaremos el procedimiento de Newton-Raphson aplicado a la función

$$g(\beta) = \frac{n}{\beta} + \sum_{i=1}^{n} \log t_i - \sum_{i=1}^{n} t_i^{\beta} \log t_i.$$

Como

$$g'(\beta) \;=\; -\frac{n}{\beta^2} - \sum_{i=1}^{n} t_i^{\beta} (\log t_i)^2,$$

el procedimiento iterativo (39) adopta la forma

$$\beta_{j+1} = \beta_j + \frac{\frac{n}{\beta} + \sum_{i=1}^{n} \log t_i - \sum_{i=1}^{n} t_i^{\beta} \log t_i}{\frac{n}{\beta^2} + \sum_{i=1}^{n} t_i^{\beta} (\log t_i)^2}. \tag{40}$$

Generando una muestra aleatoria de tamaño $n = 20$ de una variable aleatoria T con densidad dada por (35) con $\beta = 2$ e inicializando el procedimiento iterativo (40) con $\beta_1 = \bar{T}$ obtuvimos que $\hat{\beta}_{mv} = 2.3674$.

Generando una muestra aleatoria de tamaño $n = 10000$ de una variable aleatoria T con densidad dada por (35) con $\beta = 2$ e inicializando el procedimiento iterativo (40) con $\beta_1 = \bar{T}$ obtuvimos que $\hat{\beta}_{mv} = 1.9969$. □

3.3. Cálculo del emv para familias no regulares

Venía rápido, muy rápido y se le soltó un patín ...

Ahora mostraremos algunos ejemplos correspondientes a familias no regulares. En estos casos hay que analizar dónde se realiza el máximo "a mano".

Ejemplo 3.15 (Distribuciones de Bernoulli con parámetros discretos)**.** Supongamos que los valores observados en la secuencia (29) que aparece en el Ejemplo 3.7 fueron arrojados por una muestra aleatoria de tamaño $n = 20$ de una variable aleatoria X con distribución Bernoulli(p), donde $p = 0.45$ o $p = 0.65$. La familia de distribuciones no es regular debido a que el espacio paramétrico $\{0.45, 0.65\}$ no es abierto. En esta situación no puede utilizarse la metodología del Lema 3.5 pues conduce a resultados totalmente disparatados. Lo único que se puede hacer es comparar los valores $L(0.45|\mathbf{x})$, $L(0.65|\mathbf{x})$ y quedarse con el valor de $p \in \{0.45, 0.65\}$ que haga máxima la probabilidad de observar el resultado $\mathbf{x}$:

$$\begin{aligned} L(0.45|\mathbf{x}) &= (0.45)^{11}(0.55)^9 = (7.0567...)10^{-7} \\ L(0.65|\mathbf{x}) &= (0.65)^{11}(0.35)^9 = (6.8969...)10^{-7}. \end{aligned}$$

Por lo tanto, el estimador de máxima verosimilitud, basado en las observaciones (29), será

$$\hat{p}_{mv}(0,1,1,1,0,1,0,1,1,0,1,1,0,0,1,0,1,0,1,0) = 0.45.$$

□

Ejemplo 3.16 (Distribución uniforme)**.** La familia $\{\mathcal{U}(0,\theta) : \theta > 0\}$ de distribuciones uniformes no es una familia regular debido a que el soporte de la densidad de la distribución $\mathcal{U}(0,\theta)$ es $[0,\theta]$ (y depende claramente del valor del parámetro θ). En esta situación tampoco

puede utilizarse la metodología del Lema 3.5. En este caso $\Theta = (0, \infty)$ y las funciones de densidad son de la forma

$$f(x|\theta) = \frac{1}{\theta}\mathbf{1}\{0 \leq x \leq \theta\}.$$

La función de verosimilitud es

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\theta}\mathbf{1}\{0 \leq x_i \leq \theta\} = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}\{0 \leq x_i \leq \theta\} \\ &= \frac{1}{\theta^n}\mathbf{1}\left\{\max_{i=1,\ldots,n} x_i \leq \theta\right\}. \end{aligned}$$

Si $\theta < \max_i x_i$, entonces $L(\theta|\mathbf{x}) = 0$. Si $\theta \geq \max_i x_i$, entonces $L(\theta|\mathbf{x}) = \theta^{-n}$, una función decreciente en θ. En consecuencia, su máximo se alcanza en

$$\theta = \max_{i=1,\ldots,n} x_i.$$

Por lo tanto, *el estimador de máxima verosimilitud para θ, basado en una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$ de una variable aleatoria $X \sim \mathcal{U}(0, \theta)$, es el máximo de la muestra*

$$\hat{\theta}_{mv}(\mathbf{X}) = X_{(n)} := \max_{i=1,\ldots,n} X_i.$$

□

Ejemplo 3.17 (Distribución uniforme)**.** La familia $\{\mathcal{U}(\theta - 1/2, \theta + 1/2) : \theta \in \mathbb{R}\}$ de distribuciones uniformes no es una familia regular debido a que el soporte de la densidad de la distribución $\mathcal{U}(\theta - 1/2, \theta + 1/2)$ es $[\theta - 1/2, \theta + 1/2]$ (y depende claramente del valor del parámetro θ). En este caso $\Theta = \mathbb{R}$ y las funciones de densidad son de la forma

$$f(x|\theta) = \mathbf{1}\{\theta - 1/2 \leq x \leq \theta + 1/2\}.$$

La función de verosimilitud es

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \mathbf{1}\{\theta - 1/2 \leq x_i \leq \theta + 1/2\} \\ &= \mathbf{1}\left\{\max_{i=1,\ldots,n} x_i - 1/2 \leq \theta \leq \min_{i=1,\ldots,n} x_i + 1/2\right\} \\ &= \mathbf{1}\left\{x_{(n)} - 1/2 \leq \theta \leq x_{(1)} + 1/2\right\}, \end{aligned}$$

pues

$$\theta - 1/2 \leq x_i \leq \theta + 1/2, \qquad i = 1, \ldots, n,$$

si y solamente si

$$\theta \leq x_i + 1/2 \qquad \text{y} \qquad x_i - 1/2 \leq \theta, \qquad i = 1, \ldots, n,$$

Como $L(\theta|\mathbf{x})$ se anula para $\theta < x_{(n)}$ y para $\theta > x_{(1)} + 1/2$ y es constantemente 1 en el intervalo $[x_{(n)} - 1/2, x_{(1)} + 1/2]$, tenemos que cualquier punto de ese intervalo es un estimador de máxima verosimilitud para θ. En particular,

$$\hat{\theta}(\mathbf{x}) = \frac{x_{(1)} + x_{(n)}}{2}$$

es un estimador de máxima verosimilitud para θ. Etc... □

3.4. Principio de invariancia

En lo que sigue presentamos una propiedad bastante importante del método de máxima verosimilitud.

Teorema 3.18 (Principio de invariancia)**.** *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable aleatoria X cuya distribución pertenece a la familia paramétrica $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$. Sea $g : \Theta \to \Lambda$ una función biunívoca de Θ sobre Λ. Si $\hat{\theta}$ es un estimador de máxima verosimilitud para θ, entonces $g(\hat{\theta})$ es un estimador de máxima verosimilitud para $\lambda = g(\theta)$.*

Demostración. Como $\lambda = g(\theta)$ es una función biunívoca de Θ sobre Λ, la función de verosimilitud $L(\theta|\mathbf{x})$ se puede expresar en función de λ ya que $\theta = g^{-1}(\lambda)$. Denominemos a la función de verosimilitud, como función de λ, por $L^*(\lambda|\mathbf{x})$. Es claro que

$$L^*(\lambda|\mathbf{x}) = L(g^{-1}(\lambda)|\mathbf{x}).$$

Sea $\hat{\theta}_{mv} \in \Theta$ un estimador de máxima verosimilitud para θ y sea $\hat{\lambda} := g(\hat{\theta}_{mv}) \in \Lambda$ su imagen por g. Hay que mostrar que vale lo siguiente:

$$L^*(\hat{\lambda}|\mathbf{x}) = \max_{\lambda \in \Lambda} L^*(\lambda|\mathbf{x})$$

Pero esto es inmediato, debido a que

$$\begin{aligned} L^*(\hat{\lambda}|\mathbf{x}) &= L(g^{-1}(\hat{\lambda})|\mathbf{x}) = L(\hat{\theta}_{mv}|\mathbf{x}) = \max_{\theta \in \Theta} L(\theta|\mathbf{x}) = \max_{\lambda \in \Lambda} L(g^{-1}(\lambda)|\mathbf{x}) \\ &= \max_{\lambda \in \Lambda} L^*(\lambda|\mathbf{x}). \end{aligned}$$

Por lo tanto,

$$\widehat{g(\theta)}_{mv} = g(\hat{\theta}_{mv}).$$

□

Ejemplo 3.19. Sea $X_1, \ldots, X_n$ una muestra aleatoria de la variable aleatoria $X \sim \mathcal{N}(\mu, 1)$. En el Ejemplo 3.8 vimos que $\hat{\mu}_{mv} = \bar{X}$ es el estimador de máxima verosimilitud para μ. Queremos estimar

$$g(\mu) = \mathbb{P}_\mu(X \leq 0) = \Phi(-\mu).$$

Por el principio de invariancia, tenemos que

$$g(\hat{\mu}_{mv}) = \Phi(-\bar{X})$$

es el estimador de máxima verosimilitud para $\mathbb{P}_\mu(X \leq 0)$. □

Nota Bene En general, si $\lambda = g(\theta)$, aunque g no sea biunívoca, se define el estimador de máxima verosimilitud de λ por

$$\hat{\lambda} = g(\hat{\theta}_{mv}).$$

4. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Bolfarine, H., Sandoval, M. C.: Introdução à Inferência Estatística. SBM, Rio de Janeiro. (2001).

2. Borovkov, A. A.: Estadística matemática. Mir, Moscú. (1984).

3. Cramer, H.: Métodos matemáticos de estadística. Aguilar, Madrid. (1970).

4. Hoel P. G.: Introducción a la estadística matemática. Ariel, Barcelona. (1980).

5. Maronna R.: Probabilidad y Estadística Elementales para Estudiantes de Ciencias. Editorial Exacta, La Plata. (1995).