5.4. Comportamiento asintótico de estimadores

Hasta ahora hemos tratado con una muestra aleatoria $(X_1, \ldots, X_n)$ de tamaño n genérico de una variable X con función de masa/densidad $f(x; \theta)$, con las que construíamos estimadores del parámetro θ. Ahora queremos analizar justamente el papel de n, el tamaño de la muestra, en la confianza que se puede tener en esas estimaciones. Porque esperamos que, cuanto mayor sea el tamaño de la muestra, dispondremos de más información sobre la fuente aleatoria X, y en particular, que si los estimadores están "bien diseñados", seremos capaces de ir afinando cada vez más la estimación del parámetro de interés.

Para ello, en lo que sigue precisaremos un poco tanto el modelo como la notación, y consideraremos una *sucesión* (T_n) *de estimadores*, uno para cada n:

$$T_n = h_n(X_1, \ldots, X_n),$$

actuando sobre muestras aleatorias $(X_1, \ldots, X_n)$ de tamaño n. Obsérvese que ahora hacemos explícita la dependencia en n del estimador.

Aunque no es imprescindible, sí resulta habitual que la sucesión de estimadores (T_n) se obtenga con funciones h_n que tienen la misma "forma" (promedio, máximo, etc.):

$$h_n(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{j=1}^{n} x_j, \quad h_n(x_1, \ldots, x_n) = \text{máx}(x_1, \ldots, x_n), \quad \text{etc.}$$

Interesa, pues, analizar el comportamiento de la sucesión de variables aleatorias (T_n) como estimadores (sucesivos) de θ, con el objetivo de establecer si efectivamente la estimación es cada vez mejor según aumenta el tamaño de la muestra.

Detallamos ahora el sentido en el que interpretaremos esa mejor aproximación.

Definición 5.13 Sea X una variable aleatoria con función de densidad/masa $f(x; \theta)$, donde $\theta \in \Theta$. Decimos que una sucesión (T_n) de estimadores del parámetro θ (donde cada T_n es una función de $(X_1, \ldots, X_n)$) es **consistente** cuando, para todo $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbf{P}_\theta(|T_n - \theta| \geq \varepsilon) = 0\,, \quad \textit{para todo } \theta \in \Theta\,.$$

Es decir, cuando para n grande, la probabilidad de que T_n yerre siquiera ε del verdadero valor θ (sea cual sea éste), sea casi nula. Para una tal sucesión de estimadores tendremos mucha confianza (para n grande) en que la estimación de θ con T_n será buena.

La condición anterior es natural, y casi exigible a una sucesión de "buenos" estimadores. Pero nos interesará, no sólo establecer si el límite anterior tiende a 0, sino también determinar *con qué velocidad* se va a 0 (en función de n): si como $1/n$, como $1/n^2$, o quizás a velocidad exponencial, con objeto de cuantificar lo probable que es que T_n se aparte de θ en una cierta cantidad.

Para establecer la consistencia de una sucesión de estimadores, y para determinar de paso esa velocidad de convergencia, caben varias alternativas. Por un lado, podemos apelar a desigualdades generales tipo Chebyshev si es que disponemos de buenas estimaciones para las sucesivas varianzas (o errores cuadráticos medios) de los T_n. Veremos algunos ejemplos en el apartado A que sigue. En otras (contadas) ocasiones, dispondremos de fórmulas explícitas para determinar la probabilidad $\mathbf{P}_\theta(|T_n - \theta| \geq \varepsilon)$, como veremos en el apartado B.

Pero sin duda el procedimiento más poderoso, y más habitual, consistirá en apoyarse en resultados asintóticos generales, como por ejemplo el teorema central del límite, para a partir de ellos deducir buenas estimaciones de la probabilidad anterior. Explicaremos brevemente este proceder en el apartado C, y luego dedicaremos las secciones 5.4.1 y 5.4.2 a ilustrar su uso en el análisis de estimadores obtenidos por momentos y por máxima verosimilitud.

☞ **Nota** 5.4.1. La convergencia de T_n a θ exhibida en la definición 5.13 se conoce como *convergencia en probabilidad*.

A. Consistencia y varianzas/errores cuadráticos medios

Digamos que los estimadores T_n son (todos ellos) estimadores insesgados del parámetro θ. Es decir, que

$$\mathbf{E}_\theta(T_n) = \theta$$

para todo $\theta \in \Theta$ y para todo n. Entonces, apelando a la desigualdad de Chebyshev (teorema 2.2) tenemos, para $\varepsilon > 0$ dado, que

$$\mathbf{P}_\theta(|T_n - \theta| \geq \varepsilon) \leq \frac{\mathbf{V}_\theta(T_n)}{\varepsilon^2}\,. \tag{5.27}$$

De manera que si la sucesión de varianzas de los T_n tiende a 0, la sucesión de los estimadores T_n será consistente.

En el caso general de estimadores no necesariamente insesgados, necesitamos que el error cuadrático medio tienda a cero (es decir, que tanto la varianza como el sesgo tiendan a 0), pues apelando a la desigualdad de Markov (teorema 2.2 de nuevo) tenemos que, para $\varepsilon > 0$,

$$\mathbf{P}_\theta(|T_n - \theta| \geq \varepsilon) \leq \frac{\mathbf{E}_\theta(|T_n - \theta|)}{\varepsilon} \leq \frac{\sqrt{\mathbf{E}_\theta((T_n - \theta)^2)}}{\varepsilon} = \frac{\sqrt{\mathrm{ECM}_\theta(T_n)}}{\varepsilon} \xrightarrow{n\to\infty} 0, \tag{5.28}$$

donde la segunda desigualdad es consecuencia de la de Cauchy–Schwarz (teorema 2.3).

Reunimos estas observaciones en el siguiente lema, que recoge criterios directos para comprobar la consistencia de una sucesión de estimadores.

Lema 5.14 a) *Sea $(T_n)_{n\geq 1}$ una sucesión de estimadores de θ tales que*

$$\lim_{n\to\infty} \mathrm{ECM}_\theta(T_n) = 0\,.$$

Entonces la sucesión $(T_n)_{n\geq 1}$ es consistente.

b) *En particular, si* $(T_n)_{n\geq 1}$ *es una sucesión de estimadores insesgados de* θ *tales que*

$$\lim_{n\to\infty} \mathbf{V}_\theta(T_n) = 0\,,$$

entonces la sucesión $(T_n)_{n\geq 1}$ *es consistente.*

Repare, lector, en que si además de saber que $\mathbf{V}_\theta(T_n) \to 0$ (o bien $\text{ECM}_\theta(T_n) \to 0$) cuando $n \to \infty$ (que nos daría la consistencia), supiéramos estimar cuán rápidamente tiende esa varianza/ECM a 0, entonces (5.27), o quizás (5.28), nos darían una estimación de la velocidad de convergencia a 0 de $\mathbf{P}_\theta(|T_n - \theta| \geq \varepsilon)$.

Ilustramos este enfoque en un par de ejemplos.

Ejemplo 5.4.1. *La media y la cuasivarianza muestrales, como estimadores consistentes de la media y la varianza de* X.

Llamemos $\mathbf{E}(X) = \mu$ a la media de X, y $\mathbf{V}(X) = \sigma^2$ a su varianza. Para indicar la dependencia en el tamaño de la muestra escribimos $\overline{X}_{(n)}$ y $S^2_{(n)}$ para la media y la cuasivarianza muestrales, respectivamente.

Sabemos que

$$\overline{X}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

es un estimador insesgado de μ, para cualquier n, y además que

$$\mathbf{V}(\overline{X}_{(n)}) = \sigma^2/n\,.$$

Así que, por el lema 5.14, $(\overline{X}_{(n)})$ es una sucesión consistente de estimadores de $\mathbf{E}(X)$. Esto requiere, por supuesto, que $\mathbf{E}(X^2) < +\infty$. La velocidad de convergencia sería del orden de $1/n$.

Como ilustración numérica, pongamos que X es una variable con media μ y varianza 1. La desigualdad de Chebyshev nos daría entonces

$$\mathbf{P}_\mu(|\overline{X}_{(n)} - \mu| \geq \varepsilon) \leq \frac{\mathbf{V}_\mu(\overline{X}_{(n)})}{\varepsilon^2} = \frac{1}{n\varepsilon^2}\,.$$

Tomando, por ejemplo, $\varepsilon = 1/10$ y $n = 900$, obtenemos que la probabilidad anterior es menor o igual que $1/9 = 11.11\,\%$.

Análogamente obtendríamos que la sucesión $(S^2_{(n)})$ de estimadores de σ^2, dada por

$$S^2_{(n)} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

es consistente, siempre que tuviéramos $\mathbf{E}(X^4) < +\infty$ y usando por ejemplo la cota de (4.6), que nos decía que

$$\mathbf{V}(S^2_{(n)}) \leq \frac{1}{n}\,\mathbf{E}\big((X - \mathbf{E}(X))^4\big)\,.$$

De nuevo, la velocidad de convergencia sería $1/n$. ♣

Ejemplo 5.4.2. *Estimadores del parámetro a para muestras de una* unif$[0,a]$.

Si tomamos $T_n = \frac{n+1}{n}\,\text{máx}(X_1,\dots,X_n)$, como ya sabemos del ejemplo 5.1.9, $\mathbf{E}_a(T_n) = a$ y $\mathbf{V}_a(T_n) = \frac{1}{n(n+2)}a^2$. Así que, por el lema 5.14, tendríamos que (T_n) es una sucesión consistente de estimadores de a; aunque ahora la velocidad de convergencia sería del orden de $1/n^2$.

Si hubiéramos tomado como estimador a $T_n = \text{máx}(X_1,\dots,X_n)$, entonces tendríamos que

$$\mathbf{E}_a(T_n) = \frac{n}{n+1}\,a = a - \overbrace{\frac{a}{n+1}}^{=\text{sesgo}} \qquad \text{y} \qquad \mathbf{V}_a(T_n) = \frac{n}{(n+1)^2(n+2)}\,a^2.$$

Ahora T_n no sería un estimador insesgado, pero tanto el sesgo como la varianza tienden a 0 cuando $n \to \infty$. Así que (T_n) es una sucesión consistente, con velocidad de convergencia del orden de $1/n^2$. ♣

B. Consistencia y distribución exacta del estimador

En ocasiones, podemos determinar la *distribución exacta* de T_n y obtener cotas mucho mejores de $\mathbf{P}_\theta(|T_n - \theta| \geq \varepsilon)$ que la dada por Chebyshev/Markov, precisando así mucho más la confianza en las estimaciones que ofrecen las sucesiones de estimadores consistentes.

Para los siguientes ejemplos usaremos resultados ya obtenidos anteriormente. De la discusión de las variables media y cuasivarianza muestral en el caso normal del capítulo 4, véase en particular el teorema 4.6 de Fisher–Cochran, obtenemos directamente las dos siguientes ilustraciones.

Ejemplo 5.4.3. *La media muestral como estimador de μ en una $X \sim \mathcal{N}(\mu,\sigma_0^2)$, con σ_0^2 conocida.*

Si $X \sim \mathcal{N}(\mu,\sigma_0^2)$ entonces $\overline{X}_{(n)} \sim \mathcal{N}(\mu,\sigma_0^2/n)$, así que

$$\mathbf{P}_\mu(|\overline{X}_{(n)} - \mu| \geq \varepsilon) = \mathbf{P}_\mu\Big(\frac{|\overline{X}_{(n)} - \mu|}{\sigma_0/\sqrt{n}} \geq \frac{\varepsilon}{\sigma_0/\sqrt{n}}\Big) = 2\Big(1 - \Phi\Big(\frac{\sqrt{n}\varepsilon}{\sigma_0}\Big)\Big)\,, \tag{5.29}$$

que es una fórmula explícita que únicamente requiere calcular (numéricamente) valores de la función de distribución Φ. Si, como en el ejemplo 5.4.1, tomamos $\sigma_0 = 1$, $n = 900$ y $\varepsilon = 1/10$, la probabilidad anterior es

$$2(1 - \Phi(3)) \approx 0.27\,\%.$$

Compárese con el 11.11 % que se obtenía con la estimación general vía Chebyshev, que no sacaba partido de que $\overline{X}_{(n)}$ es una variable normal. ♣

☞ **Nota** 5.4.2. Podemos reescribir la velocidad de convergencia que supone la expresión (5.29) en términos de funciones elementales usando la estimación $1 - \Phi(x) \leq \phi(x)/x$, para $x > 0$, de la cola de la normal estándar (nota 2.3.3), de la siguiente manera:

$$\mathbf{P}_\mu(|\overline{X}_{(n)} - \mu| \geq \varepsilon) = 2\Big(1 - \Phi\Big(\frac{\varepsilon\sqrt{n}}{\sigma_0}\Big)\Big) \leq \frac{2\sigma_0}{\varepsilon\sqrt{n}}\,\frac{1}{\sqrt{2\pi}}\,e^{-n\varepsilon^2/(2\sigma_0^2)}\,.$$

Recuérdese que la desigualdad de Chebyshev daba

$$\mathbf{P}_\mu(|\overline{X}_{(n)} - \mu| \geq \varepsilon) \leq \frac{\mathbf{V}_\mu(\overline{X}_{(n)})}{\varepsilon^2} = \frac{\sigma_0^2}{n\varepsilon^2}\,.$$

Como comparación, pongamos que $\sigma_0 = 1$ y tomemos $\varepsilon = n^{-1/4}$ (aunque, para esta comparación, en lugar de $1/4$ valdría cualquier exponente menor que $1/2$). Tendríamos entonces una estimación

$$\mathbf{P}_\mu(|\overline{X}_{(n)} - \mu| \geq n^{1/4}) \leq \sqrt{\frac{2}{\pi}}\,\frac{1}{n^{1/4}}\,e^{-\sqrt{n}/2}$$

vía el teorema central del límite, frente a la estimación

$$\mathbf{P}_\mu(|\overline{X}_{(n)} - \mu| \geq n^{1/4}) \leq \frac{1}{\sqrt{n}}$$

que daría Chebyshev. ¡No hay color!

EJEMPLO 5.4.4. *La cuasivarianza muestral como estimador de σ^2 para una variable $X \sim \mathcal{N}(\mu_0, \sigma^2)$, con μ_0 conocida.*

Como $(n-1)S^2_{(n)}/\sigma^2 \sim \chi^2_{n-1}$, tenemos que

$$\begin{aligned}
\mathbf{P}_{\sigma^2}(|S^2_{(n)} - \sigma^2| \geq \varepsilon) &= \mathbf{P}_{\sigma^2}\Big(\Big|\frac{(n-1)S^2_{(n)}}{\sigma^2} - (n-1)\Big| \geq \frac{\varepsilon(n-1)}{\sigma^2}\Big)\\
&= \mathbf{P}_{\sigma^2}\Big(\frac{(n-1)S^2_{(n)}}{\sigma^2} \geq (n-1)\Big(1 + \frac{\varepsilon}{\sigma^2}\Big)\Big) + \mathbf{P}_{\sigma^2}\Big(\frac{(n-1)S^2_{(n)}}{\sigma^2} \leq (n-1)\Big(1 - \frac{\varepsilon}{\sigma^2}\Big)\Big)\\
&= \big[1 - F_{\chi^2_{n-1}}\big((n-1)(1+\varepsilon/\sigma^2)\big] + F_{\chi^2_{n-1}}\big((n-1)(1-\varepsilon/\sigma^2)\big);
\end{aligned}$$

una expresión exacta, en términos de percentiles de la χ^2_{n-1}, válida *para todo n*.

Adelántandonos a los argumentos que detallaremos en el apartado C, observemos que una χ^2_{n-1} es una suma de $n-1$ variables independientes, todas con la misma distribución (de hecho, normales estándar al cuadrado), y que su esperanza es $n-1$ y su varianza $2(n-1)$. Por el teorema del límite central tenemos que, si n es grande,

$$\frac{\frac{(n-1)S^2_{(n)}}{\sigma^2} - (n-1)}{\sqrt{2(n-1)}} \approx \mathcal{N}(0,1)$$

(«≈» en el sentido de que las funciones de distribución son próximas), lo que nos da, tras unas ligeras manipulaciones, que

$$\mathbf{P}_{\sigma^2}(|S^2_{(n)} - \sigma^2| \geq \varepsilon) \approx 2\Big(1 - \Phi\Big(\frac{\varepsilon\sqrt{n-1}}{\sqrt{2}\,\sigma^2}\Big)\Big)$$

cuando n es grande, que es justamente el caso de interés. ♣

Ejemplo 5.4.5. *Estimador $M_n = \text{máx}(X_1, \dots, X_n)$ del parámetro a de* UNIF$[0, a]$.

El estimador M_n es sesgado, pero asintóticamente es insesgado, pues su media, como ya sabemos, es $\mathbf{E}_a(M_n) = \frac{n}{n+1}\,a$, que tiende a a cuando $n \to \infty$.

De $M_n = \text{máx}(X_1, \dots, X_n)$ conocemos la distribución exacta (ejemplo 5.1.7), que nos permite escribir

$$\mathbf{P}_a(|M_n - a| \geq \varepsilon) = \mathbf{P}_a(M_n \leq a - \varepsilon) = (1 - \varepsilon/a)^n \,.$$

Así que M_n no sólo es consistente, sino que además la probabilidad de error de estimación converge exponencialmente a 0.

Pero, más interesante: cambiemos la escala del error y consideremos

$$\mathbf{P}_a(|M_n - a| \geq (a\varepsilon/n)) = \big(1 - \varepsilon/n\big)^n \underset{n\to\infty}{\longrightarrow} e^{-\varepsilon}$$

Esto nos dice que

$$\frac{n}{a}(a - M_n) \quad \text{converge en distribución a} \quad \textsc{exp}(1)\,.$$

Es decir, con esa normalización tenemos una distribución exacta en el límite, que precisa el error que se puede llegar a cometer cuando se estima a con M_n. ♣

Ejemplo 5.4.6. *La media muestral como estimador del parámetro $\theta = 1/\lambda$ para $X \sim$* EXP(λ).

Recuérdese (sección 2.3.4) que $X \sim$ GAMMA$(1/\theta, 1)$. Usando la proposición 2.15 y las propiedades de rescalado de las variables Gamma, obtenemos:

$$\sum_{i=1}^n X_i \sim \textsc{gamma}\Big(\frac{1}{\theta}, n\Big) \implies \frac{2}{\theta}\sum_{i=1}^n X_i \sim \textsc{gamma}\Big(\frac{1}{2}, n\Big) = \textsc{gamma}\Big(\frac{1}{2}, \frac{2n}{2}\Big) = \chi^2_{2n}\,.$$

De manera que podemos calcular explícitamente:

$$\begin{aligned}\mathbf{P}_\theta\Big(\Big|\,\overline{X}_{(n)} - \frac{1}{\theta}\Big| \geq \varepsilon\Big) &= \mathbf{P}_\theta\Big(\Big|\,\frac{2}{\theta}\sum_{i=1}^n X_i - 2n\Big| \geq 2n\,\frac{\varepsilon}{\theta}\Big) \\ &= \mathbf{P}\Big(\chi^2_{2n} > 2n\Big(1 + \frac{\varepsilon}{\theta}\Big)\Big) + \mathbf{P}\Big(\chi^2_{2n} < 2n\Big(1 - \frac{\varepsilon}{\theta}\Big)\Big)\end{aligned}$$

De nuevo una expresión exacta, en términos de percentiles de χ^2_{2n}. ♣

☞ **Nota** 5.4.3. En los ejemplos anteriores la velocidad de consistencia se ha expresado en términos de la $\mathcal{N}(0, 1)$, de la χ^2_n y de la EXP(1).

C. Consistencia y aproximaciones asintóticas generales

Supongamos que (T_n) es una sucesión consistente de estimadores de un parámetro θ. Esto nos dice que, para todo $a > 0$,

$$\mathbf{P}_\theta(|T_n - \theta| \leq a) \xrightarrow{n\to\infty} 1,$$

esto es, que para n grande, es muy probable T_n se "parezca" mucho al valor de θ. Vamos ahora a reescalar esas (tan pequeñas) diferencias entre T_n y θ, considerando una sucesión (k_n) y la probabilidad

$$\mathbf{P}_\theta(k_n\,|T_n - \theta| \leq a), \tag{5.30}$$

donde $a > 0$ está fijo. Si la sucesión (k_n) es constante, por muy grande que ésta sea, la probabilidad (5.30) seguirá tendiendo a 1 cuando $n \to \infty$. Por otro lado, si $k_n \to \infty$ muy rápidamente, esa probabilidad tenderá a 0 cuando $n \to \infty$. Cabe imaginar entonces que pudiera existir una sucesión (k_n) para la que la probabilidad anterior tienda a un número estrictamente entre 0 y 1, para ese $a > 0$ prefijado.

En la mayor parte de los estimadores que se manejan, ocurrirá que el argumento anterior se puede extender a todos los valores de a, y que existirá una sucesión (k_n), que tiende a $+\infty$ cuando $n \to \infty$, para la que

$$\mathbf{P}_\theta(k_n\,|T_n - \theta| \leq a) \to F(a) \qquad \text{cuando } n \to \infty, \tag{5.31}$$

donde $F(a)$ es una cierta *función de distribución continua*. Si éste es el caso, y somos capaces de calcular los valores de la función F, entonces podremos usar la convergencia dada en (5.31) para obtener estimaciones (asintóticas, cuando $n \to \infty$) de probabilidades del tipo $\mathbf{P}_\theta(|T_n - \theta| \leq a)$ sin más que deshacer los cambios de escala.

El ejemplo paradigmático de esta manera de proceder es, claro, el teorema central del límite (teorema 4.2), en el que para una variable X con media μ y varianza σ^2 tomamos $T_n = \overline{X}_{(n)}$, y calculamos probabilidades del tipo

$$\mathbf{P}_\mu(|\overline{X}_{(n)} - \mu| < \varepsilon)$$

aprovechando que, para todo $t \in \mathbb{R}$,

$$\mathbf{P}_\mu\Big(\sqrt{n}\,\Big(\frac{\overline{X}_{(n)} - \mu}{\sigma}\Big) < t\Big) \to \Phi(t) \qquad \text{cuando } n \to \infty,$$

donde Φ es la función de distribución de la normal estándar, como ya se ha ilustrado en varias ocasiones. Nótese que la sucesión (k_n) viene dada aquí por $\sqrt{n}/\sigma$.

En las dos siguientes secciones usaremos esta idea en el análisis asintótico de los dos tipos de estimadores más habituales, los que se obtienen por el método de los momentos, y los de máxima verosimilitud.

5.4.1. Comportamiento asintótico de los estimadores por momentos

Para una variable general con $\mathbf{E}(X^2) < +\infty$, el teorema central del límite nos permite precisar, con la normalización adecuada, cuán buena es la estimación de $\mathbf{E}(X)$ con la media muestral $\overline{X}_{(n)}$.

Si X tiene $\mathbf{E}(X) = \mu$ y $\mathbf{V}(X) = \sigma^2$, entonces

$$\Big(\frac{\overline{X}_{(n)} - \mu}{\sigma/\sqrt{n}}\Big) \quad \text{converge en distribución a una variable } \mathcal{N}(0,1)\,.$$

Resulta más conveniente escribir esta convergencia como sigue:

$$\sqrt{n}\,(\overline{X}_{(n)} - \mu) \quad \text{converge en distribución a una variable } \mathcal{N}(0,\sigma^2)\,.$$

Así que, para n grande,

$$\mathbf{P}(|\overline{X}_{(n)} - \mu| \geq \varepsilon) = \mathbf{P}(\sqrt{n}\,|\overline{X}_{(n)} - \mu|/\sigma \geq \sqrt{n}\varepsilon/\sigma) \approx 2\big(1 - \Phi(\sqrt{n}\varepsilon/\sigma)\big)\,.$$

En el caso en que X sea normal, el "$\approx$"es un "$=$".

Veamos un ejemplo.

Ejemplo 5.4.7. *La media muestral como estimador de p en una $X \sim \text{BER}(p)$.*

Recuérdese que $\mathbf{E}_p(X) = p$ y que $\mathbf{V}_p(X) = p(1-p)$.

La variable media muestral (para muestras de tamaño n)

$$\overline{X}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

tiene media $\mathbf{E}_p(\overline{X}_{(n)}) = p$ y varianza $\mathbf{V}_p(\overline{X}_{(n)}) = p(1-p)/n$.

De manera que, para n grande,

$$\mathbf{P}_p(|\overline{X}_{(n)} - p| \geq \varepsilon) = \mathbf{P}\Big(\sqrt{n}\,\frac{|\overline{X}_{(n)} - p|}{\sqrt{p(1-p)}} \geq \sqrt{n}\,\frac{\varepsilon}{\sqrt{p(1-p)}}\Big) \approx 2\,\Big(1 - \Phi\Big(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\Big)\Big)\,.$$

Por cierto, en este caso, como la función $x \mapsto x(1-x)$ en el intervalo $[0,1]$ alcanza su máximo en $x = 1/2$, resulta que $0 \leq p(1-p) \leq 1/4$ para $p \in [0,1]$, y por tanto podemos estimar la probabilidad anterior como sigue:

$$\mathbf{P}_p(|\overline{X}_{(n)} - p| \geq \varepsilon) \approx 2\,\Big(1 - \Phi\Big(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\Big)\Big) \leq 2\,\big(1 - \Phi(2\sqrt{n}\,\varepsilon)\big)\,,$$

que es una estimación *uniforme* en p. ■

Si el estimador en cuestión fuera, en lugar de la habitual media muestral, un promedio de, digamos, cuadrados, de la forma

$$\overline{X^2}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

aplicaríamos el mismo procedimiento salvo que ahora, claro, los parámetros de la tipificación son distintos. En concreto, como

$$\mathbf{E}\big(\overline{X^2}_{(n)}\big) = \mathbf{E}(X^2) \qquad \text{y} \qquad \mathbf{V}\big(\overline{X^2}_{(n)}\big) = \frac{\mathbf{V}(X^2)}{n},$$

tendríamos que

$$\sqrt{n}\,\big(\overline{X^2}_{(n)} - \mathbf{E}(X^2)\big) \quad \text{converge en distribución a una variable } \mathcal{N}(0, \mathbf{V}(X^2))\,.$$

Con frecuencia, el estadístico de interés para estimar un cierto parámetro es una *cierta función* g de la variable media muestral $\overline{X}_{(n)}$, o de la media muestral de los cuadrados $\overline{X^2}_{(n)}$, como por ejemplo

$$\frac{1}{\overline{X}_{(n)}} \qquad \text{o quizás} \qquad \sqrt{\overline{X^2}_{(n)}} = \Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big)^{1/2}.$$

Revise el lector los ejemplos vistos hasta ahora, y en particular los estimadores que se obtienen por el método de momentos del apartado 5.2.

En esos casos, el siguiente teorema resulta extremadamente útil.

Teorema 5.15 (Método delta) *Sea Z_n una sucesión de variables aleatorias en un espacio de probabilidad $(\Omega, \mathbf{P}, \mathcal{F})$ tal que, para ciertos $\alpha \in \mathbb{R}$ y $\beta > 0$,*

$$\sqrt{n}\,(Z_n - \alpha) \quad \text{converge en distribución a} \quad \mathcal{N}(0, \beta).$$

Supongamos que el soporte de todas las variables Z_n está contenido en un cierto intervalo $(a, b) \subset \mathbb{R}$, en el que se halla α.

Sea g una función que es C^2 en (a, b) y tal que $g'(\alpha) \neq 0$. Entonces

$$\sqrt{n}\,(g(Z_n) - g(\alpha)) \quad \text{converge en distribución a una variable } \mathcal{N}(0, |g'(\alpha)|^2\,\beta).$$

En el uso que haremos aquí de este resultado, como sucesión (Z_n) tomaremos habitualmente la sucesión de medias muestrales $(\overline{X}_{(n)})$ de una cierta variable X, cuya normalidad, tras la pertinente normalización (que usaría como α a la media de X, y como β, a su varianza), viene garantizada por el teorema central del límite. La conclusión será que funciones (razonables) de esa media muestral también exhiben normalidad asintótica.

Exhibimos primero la idea de la demostración, para luego, en la nota 5.4.4, dar una demostración completa con todos sus detalles para el lector interesado.

IDEA DE LA DEMOSTRACIÓN. Se tiene que

$$\sqrt{n}\,(Z_n - \alpha) \approx \sqrt{\beta}\,Y\,,$$

donde Y es una normal estándar. Obviemos $\approx$. Así que

$$Z_n = \alpha + \frac{\sqrt{\beta}}{\sqrt{n}}\,Y\,.$$

Obsérvese que Z_n será proximo a α para n grande.

Ahora

$$g(x) \approx g(\alpha) + g'(\alpha)(x-\alpha)\,,$$

si x próximo a α. Obviemos $\approx$.

Combinando, tenemos que

$$g(Z_n) = g(\alpha) + g'(\alpha)\,(Z_n-\alpha) = g(\alpha) + \frac{g'(\alpha)\sqrt{\beta}}{\sqrt{n}}\,Y\,.$$

Es decir, como la normal estándar es simétrica, $\sqrt{n}\,(g(Z_n)-g(\alpha))$ es una normal de media 0 y varianza $|g'(\alpha)|^2\,b$. ■

☞ **Nota** 5.4.4. Añadimos ahora los detalles que justifican estas (excesivamente ligeras) transiciones entre "$\approx$" e "$=$":

Demostración del teorema 5.15. Fijemos $t\in\mathbb{R}$. Queremos comprobar que

$$\lim_{n\to\infty} \mathbf{P}\big(\sqrt{n}\,(g(Z_n)-g(\alpha))\le t\big) = \Phi\Big(\frac{t}{\sqrt{\beta}\,g'(\alpha)}\Big)\,.$$

Llamemos I a un intervalo abierto que contenga a α, y donde $\frac{1}{2}|g''(x)|\le M$, para un cierto $M>0$. Por la simetría de la normal (centrada en 0) podemos suponer que $g'(\alpha)>0$, pues si no es el caso bastaría considerar $-g$.

Fijemos $x\in I$. Como

$$\int_\alpha^x \Big(\int_\alpha^s g''(u)\,du\Big)\,ds = \int_\alpha^x \big(g'(s)-g'(\alpha)\big)\,ds = g(x)-g(\alpha)-g'(\alpha)\,(x-\alpha),$$

se tiene que

$$g(x)-g(\alpha) = g'(\alpha)(x-\alpha) + \int_\alpha^x\int_\alpha^s g''(u)\,du\,ds$$

(de nuevo, el polinomio de Taylor de grado 1, con término de error), de donde

$$(\sharp)\qquad |g(x)-g(\alpha)-g'(\alpha)(x-\alpha)| \le 2M\int_\alpha^x\int_\alpha^s du\,ds = 2M\,\frac{(x-\alpha)^2}{2} = M(x-\alpha)^2\,.$$

Fijemos $\delta\in(0,1)$ y $\varepsilon>0$, ambos pequeños, que luego haremos tender a 0. Por hipótesis tenemos que, para todo $h>0$,

$$\lim_{n\to\infty}\mathbf{P}(\sqrt{n}\,|Z_n-\alpha|\le h) = \Phi(h/\sqrt{\beta})-\Phi(-h/\sqrt{\beta}) = 2\,\Phi(h/\sqrt{\beta})-1\,.$$

Recuérdese que $\Phi(z)$ es una función que crece hacia 1 cuando $z\to\infty$. Fijemos entonces h grande, tan grande como para que $2\Phi(h/\sqrt{\beta})-1\ge 1-\delta/2$. Y tomemos N suficientemente grande como para que, si $n\ge N$, se tenga que

$$\mathbf{P}(\sqrt{n}\,|Z_n-\alpha|\le h)\ge 1-\delta\,.$$

Consideremos, para $n\ge N$, el suceso

$$\Omega_n = \{\sqrt{n}\,|Z_n-\alpha|\le h\},$$

para el que $\mathbf{P}(\Omega_n)\ge 1-\delta$.

En Ω_n se tiene que $|Z_n-\alpha|\le h/\sqrt{n}$, así que aumentando N si fuera necesario, tenemos que para $n\ge N$ se tiene en Ω_n que $Z_n\in I$. Lo relevante de Ω_n es que $\mathbf{P}(\Omega_n)\ge 1-\delta$, y además que $Z_n\in I$ en Ω_n.

Por tanto, usando ($\sharp$) y la definición de Ω_n, para $n \geq N$ se cumple en Ω_n que

$$(\star) \qquad |g(Z_n) - g(\alpha) - g'(\alpha)(Z_n - \alpha)| \leq M|Z_n - \alpha|^2 \leq \frac{Mh^2}{n}\,,$$

Aumentemos N, si fuera necesario, para garantizar que

$$(\star\star) \qquad \frac{M\,h^2}{\sqrt{n}\,g'(\alpha)} \leq \varepsilon\,.$$

Entonces, para $n \geq N$, en el suceso

$$\widetilde{\Omega}_n := \Omega_n \cap \Big\{\sqrt{n}\,(Z_n - \alpha) \leq \frac{t}{g'(\alpha)} - \varepsilon\Big\}$$

se cumple además que

$$\sqrt{n}\,(g(Z_n) - g(\alpha)) \leq t\,,$$

pues

$$\sqrt{n}\,\big(g(Z_n) - g(\alpha)\big) \leq \sqrt{n}\,g'(\alpha)\,(Z_n - \alpha) + \frac{Mh^2}{\sqrt{n}} \leq \sqrt{n}\,g'(\alpha)\,(Z_n - \alpha) + \varepsilon g'(\alpha) \leq t\,,$$

usando ($\star$) en la primera desigualdad, ($\star\star$) en la segunda, y la definición de $\widetilde{\Omega}_n$ en la tercera.

Por tanto,

$$\mathbf{P}\Big(\Omega_n \cap \Big\{\sqrt{n}\,(Z_n - \alpha) \leq \frac{t}{g'(\alpha)} - \varepsilon\Big\}\Big) \leq \mathbf{P}(\sqrt{n}\,(g(Z_n) - g(\alpha)) \leq t)\,.$$

Como $\mathbf{P}(\Omega \setminus \Omega_n) \leq \delta$, la parte del suceso $\{\sqrt{n}\,(Z_n - \alpha) \leq t/g'(\alpha) - \varepsilon\}$ que no está en Ω_n tiene probabilidad a lo sumo δ, de lo que deducimos que, para $n \geq N$, se tiene que

$$\mathbf{P}\Big(\sqrt{n}\,(Z_n - \alpha) \leq \frac{t}{g'(\alpha)} - \varepsilon\Big) \leq \mathbf{P}(\sqrt{n}\,(g(Z_n) - g(\alpha)) \leq t) + \delta\,.$$

Recordamos que esta desigualdad es válida para $n \geq N$, todo $t \in \mathbb{R}$, todo $\delta \in (0,1)$ y todo $\varepsilon > 0$.

Por consiguiente, haciendo n tender a ∞ y usando la convergencia en distribución de $\sqrt{n}\,(Z_n - \alpha)$ a $\mathcal{N}(0,\beta)$ se deduce que

$$\Phi\Big(\frac{t}{\sqrt{\beta}g'(\alpha)} - \frac{\varepsilon}{\sqrt{\beta}}\Big) \leq \liminf_{n\to\infty} \mathbf{P}\big(\sqrt{n}\,(g(Z_n) - g(\alpha)) \leq t\big) + \delta\,.$$

Haciendo ahora $\varepsilon \downarrow 0$ y usando que la función de distribución Φ es una función continua se deduce que

$$\Phi\Big(\frac{t}{\sqrt{\beta}g'(\alpha)}\Big) \leq \liminf_{n\to\infty} \mathbf{P}\big(\sqrt{n}\,(g(Z_n) - g(\alpha)) \leq t\big) + \delta\,.$$

Y haciendo ahora $\alpha \downarrow 0$ se concluye que

$$\Phi\Big(\frac{t}{\sqrt{\beta}g'(\alpha)}\Big) \leq \liminf_{n\to\infty} \mathbf{P}\big(\sqrt{n}\,(g(Z_n) - g(\alpha)) \leq t\big)\,.$$

Un argumento análogo nos da que

$$\Phi\Big(\frac{t}{\sqrt{\beta}g'(\alpha)}\Big) \geq \limsup_{n\to\infty} \mathbf{P}\big(\sqrt{n}\,(g(Z_n) - g(\alpha)) \leq t\big)\,.$$

y, por tanto, que

$$\Phi\Big(\frac{t}{\sqrt{\beta}g'(\alpha)}\Big) = \lim_{n\to\infty} \mathbf{P}\big(\sqrt{n}\,(g(Z_n) - g(\alpha)) \leq t\big)\,.$$

Finalmente como esto es válido para todo $t \in \mathbb{R}$, se concluye que $\sqrt{n}\,(g(Z_n) - g(\alpha))$ converge en distribución a $\mathcal{N}(0, |g'(\alpha)|^2\beta)$. ■

Veamos algunos ejemplos de aplicación de este método delta.

Ejemplo 5.4.8. *Estimador $1/\overline{X}_{(n)}$ del parámetro λ para $X \sim \text{EXP}(\lambda)$.*

Recordemos que $\mathbf{E}_\lambda(X) = 1/\lambda$ y que $\mathbf{V}_\lambda(X) = 1/\lambda^2$. Denotemos $\overline{X}_{(n)}$ a la media muestral para muestras de tamaño n. Por el teorema central del límite,

$$\sqrt{n}\left(\overline{X}_{(n)} - \frac{1}{\lambda}\right) \quad \text{converge en distribución a} \quad \mathcal{N}\Big(0, \frac{1}{\lambda^2}\Big).$$

Consideremos ahora $g(x) = 1/x$ para $x \in (0, +\infty)$. Obsérvese que

$$|g'(\mu)|^2 = |g'(1/\lambda)|^2 = \lambda^4.$$

Entonces, por el método delta, teorema 5.15,

$$\sqrt{n}\left(\frac{1}{\overline{X}_{(n)}} - \lambda\right) \quad \text{converge en distribución a} \quad \mathcal{N}\big(0, \lambda^2\big).$$

♣

Ejemplo 5.4.9. *Estimación de odds.*

Tenemos una $X \sim \text{BER}(p)$. Queremos estimar el parámetro $q = p/(1-p)$, que en la jerga (de apuestas) se conoce como "odds" a favor.

Como

$$\sqrt{n}\left(\overline{X}_{(n)} - p\right) \quad \text{converge en distribución a} \quad \mathcal{N}\big(0, p(1-p)\big),$$

usando $g(x) = x/(1-x)$ para $x \in (0,1)$ se tiene que

$$\sqrt{n}\left(\frac{\overline{X}_{(n)}}{1-\overline{X}_{(n)}} - \frac{p}{1-p}\right) \quad \text{converge en distribución a} \quad \mathcal{N}\Big(0, \frac{p}{(1-p)^3}\Big),$$

pues $g'(p) = 1/(1-p)^2$. ♣

Ejemplo 5.4.10. *Estimadores para el parámetro σ^2 de una $X \sim \text{RAY}(\sigma^2)$.*

El parámetro es $\theta = \sigma^2$. Tenemos, véase el ejemplo 5.2.4, dos estimadores por momentos del parámetro θ:

$$\widehat{M}_{(n)} = \frac{2}{\pi}\,\overline{X}^2 \qquad \text{y} \qquad M_{(n)} = \frac{1}{2}\,\overline{X^2}.$$

Vamos a usar las fórmulas explícitas de los siguientes momentos de la variable X:

$$\begin{aligned} \mathbf{E}_\theta(X) &= \sqrt{\frac{\pi\,\theta}{2}}, & \mathbf{E}_\theta(X^2) &= 2\theta, \\ \mathbf{V}_\theta(X) &= \Big(2 - \frac{\pi}{2}\Big)\theta, & \mathbf{V}_\theta(X^2) &= 4\theta^2. \end{aligned}$$

El análisis del segundo estimador $M_{(n)}$ es directo. Por el teorema central del límite, y usando los valores de $\mathbf{E}_\theta(X^2)$ y $\mathbf{V}_\theta(X^2)$ dados arriba,

$$\sqrt{n}\,\big(\overline{X^2}_{(n)} - 2\theta\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\big(0, 4\theta^2\big)\,,$$

y, por tanto,

$$\sqrt{n}\,\big(M_{(n)} - \theta\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\big(0, \theta^2\big)\,. \tag{5.32}$$

Para el primer estimador $\widehat{M}_{(n)}$, partimos del teorema central del límite, usando ahora los valores de $\mathbf{E}_\theta(X)$ y $\mathbf{V}_\theta(X)$,

$$\sqrt{n}\,\big(\overline{X}_{(n)} - \sqrt{\pi\theta/2}\,\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\big(0, (2-\pi/2)\theta\big)\,,$$

y aplicamos el método delta con la función $g(y) = 2y^2/\pi$, para $y \in (0,\infty)$. Como

$$|g'(\sqrt{\pi\theta/2})| = \frac{8\theta}{\pi},$$

se obtiene que

$$\sqrt{n}\,\big(\widehat{M}_{(n)} - \theta\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\big(0, \frac{4(4-\pi)}{\pi}\theta^2\big)\,. \tag{5.33}$$

Obsérvese, comparando (5.33) con (5.32), que el estimador $\frac{1}{2}\,\overline{X^2}$ es algo mejor asintóticamente, pues $4(4-\pi)/\pi \approx 1.092 > 1$. Aunque en realidad ya sabíamos, ejemplo 5.3.14, que este estimador es óptimo (eficiente) *para cada n*. Véase también el ejemplo 5.4.12. ♣

El método delta (teorema 5.15) exige, además de una cierta regularidad para la función $g(x)$, que en los ejemplos habituales se tiene siempre, que $g'(\alpha) \neq 0$. En el caso en el que $g'(\alpha) = 0$ tenemos un resultado alternativo, que vamos a enunciar a continuación. Siguiendo la idea de la demostración del método delta, observamos (sustituyendo '$\approx$' por '$=$') que

- $g(x) - g(\alpha) = \frac{1}{2}\,g''(\alpha)(x-\alpha)^2$,
- $Z_n - \alpha = \dfrac{\sqrt{\beta}}{\sqrt{n}}\,\mathcal{Z}$, con $\mathcal{Z}$ normal estándar;

de lo que "concluimos" que

$$g(Z_n) - g(\alpha) = \frac{1}{2}\,g''(\alpha)\,\frac{\beta}{n}\,\mathcal{Z}^2,$$

es decir,

$$n\,(g(Z_n) - g(\alpha)) = \frac{g''(\alpha)\,\beta}{2}\,\mathcal{Z}^2\,.$$

Teorema 5.16 (Método delta con $g'(\alpha) = 0$) *Sea Z_n una sucesión de variables aleatorias en un espacio $(\Omega, \mathbf{P})$ tal que*

$$\sqrt{n}\,(Z_n - \mu) \quad \text{converge en distribución a una variable } \mathcal{N}(0, \sigma^2).$$

Supongamos que el soporte de todas las variables Z_n está contenido en un cierto intervalo $(a, b) \subset \mathbb{R}$, en el que se halla α.

Sea g una función, C^3 en (a, b), para la que $g'(\alpha) = 0$ pero $g''(\alpha) \neq 0$. Entonces

$$n\,(g(Z_n) - g(\alpha)) \quad \textit{converge en distribución a una variable} \quad \frac{g''(\alpha)}{2}\,\beta\,\chi_1^2\,.$$

Obsérvese que no hay valor absoluto en esta segunda derivada que aparece multiplicando a la chi cuadrado.

Demostración. La prueba es análoga a la del teorema 5.15. ■

Ejemplo 5.4.11. *Estimación de la varianza en una $X \sim \text{BER}(p)$.*

Queremos estimar el parámetro $p(1-p)$ (la varianza). Partimos, como antes, de que

$$\sqrt{n}\,\big(\overline{X}_{(n)} - p\big) \quad \text{converge en distribución a} \quad \mathcal{N}\big(0, p(1-p)\big)\,.$$

Ahora, el estimador natural es

$$\overline{X}_{(n)}(1 - \overline{X}_{(n)}).$$

Así que debemos considerar, en $x \in (0, 1)$, la función $g(x) = x(1-x)$, para la que $g'(x) = 1 - 2x$.

Esto nos da, si $p \neq 1/2$, usando el teorema 5.15,

$$\sqrt{n}\,\big(\overline{X}_{(n)}(1 - \overline{X}_{(n)}) - p(1-p)\big) \xrightarrow[n\to\infty]{\text{d}} \mathcal{N}\big(0, (1-2p)^2 p(1-p)\big)\,. \tag{5.34}$$

Sin embargo, para el caso $p = 1/2$, tenemos que $g'(1/2) = 0$. Obsérvese que $g''(x) = -2$. La convergencia de partida es ahora

$$\sqrt{n}\,\big(\overline{X}_{(n)} - 1/2\big) \quad \text{converge en distribución a} \quad \mathcal{N}\big(0, 1/4\big)\,.$$

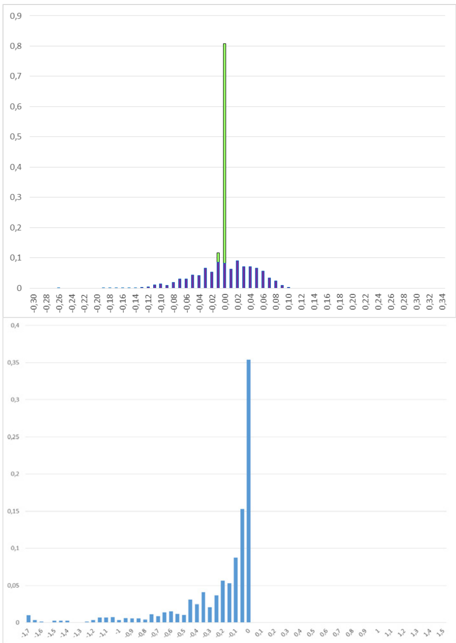
Y la conclusión, vía el teorema 5.16, es que

$$n\,\big(\overline{X}_{(n)}(1 - \overline{X}_{(n)}) - 1/4\big) \quad \text{converge en distribución a} \quad -\frac{1}{4}\,\chi_1^2\,. \tag{5.35}$$

♣

El resultado del método delta del teorema 5.15 será extremadamente útil en lo que sigue, por ejemplo en el diseño de intervalos (aproximados) de confianza; véase, por ejemplo, el apartado 6.3.4. Sin embargo, el uso en estadística del método delta en su versión del teorema 5.16, que hemos ilustrado en el ejemplo 5.4.11 anterior, es más inhabitual, pues requiere disponer de una información previa que, en muchas ocasiones, conforma un círculo vicioso. Recuerde, lector, que en el ejemplo 5.4.11 pretendemos estimar la varianza a partir de muestras; pero si *ya sabemos* que $p = 1/2$, que es el caso en el que aplicaríamos el teorema 5.16, entonces la varianza es $1/4$, y el experimento en sí pierde sentido.

Sin embargo, el teorema 5.16 desvela que, en ciertas situaciones, la variable transformada $g(Z_n)$ no es, asintóticamente, una normal, sino más bien una χ^2.



En las gráficas de la izquierda hemos representado los resultados recogen un análisis experimental de la variable

$$Z_n = \sqrt{n}\,\big(\overline{X}_{(n)}(1-\overline{X}_{(n)}) - p(1-p)\big)$$

que aparece en la ecuación (5.34), análisis con el que se intenta "comprobar" la convergencia en distribución que se enuncia en la citada ecuación. El primer histograma recoge muestras de Z_n para $n = 1500$, y para dos valores de p: $p = 50\,\%$ y $p = 60\,\%$. Para $p = 60\,\%$, el histograma (en morado) se asemeja a una normal; pero para la moneda equilibrada, las muestras tienden a concentrarse en el 0, como se aprecia en el histograma verde. Esto sucede porque el factor $\sqrt{n}$ no tiende suficientemente rápido a ∞ como para compensar la velocidad con la que tiende a 0 el otro factor. Sin embargo, al cambiar la escala de la normalización, poniendo un factor n, tal y como se sugiere en (5.35), obtenemos (en este caso de $p = 1/2$) el histograma que se representa en la segunda gráfica de la izquierda, que es (muy aproximadamente) el de una χ^2_1 (multiplicada por $-1/4$).

5.4.2. Comportamiento asintótico de los estimadores de máxima verosimilitud

Los estimadores de máxima verosimilitud tienen, en general, buenas propiedades asintóticas.

El primer resultado que vamos a ver, la convergencia recogida en (5.36), justifica en gran medida el concepto de máxima verosimilitud.

Veamos primero la idea. Estamos en el contexto habitual, de muestras de tamaño n de una variable X con función de densidad/masa $f(x;\theta)$, con $\theta \in \Theta$. Consideramos únicamente dos valores distintos, θ_0 y θ_1, que compiten entre sí.

Disponemos de una muestra $(x_1,\dots,x_n)$ que ha sido producida con la distribución correspondiente al parámetro θ_0. Calculamos las cantidades $\textsc{vero}(\theta_0;x_1,\dots,x_n)$ y $\textsc{vero}(\theta_1;x_1,\dots,x_n)$. Caben dos posibilidades:

a) $\textsc{vero}(\theta_0;x_1,\dots,x_n) > \textsc{vero}(\theta_1;x_1,\dots,x_n)$;

b) $\textsc{vero}(\theta_0;x_1,\dots,x_n) \le \textsc{vero}(\theta_1;x_1,\dots,x_n)$.

Si ocurriera a), entonces consideraríamos que θ_0 es más verosímil, y nos decantaríamos por él en esta comparación.

Si ocurriera a) la mayor parte de las veces (es decir, para casi todas las muestras generadas con el parámetro θ_0), entonces tendríamos alta confianza en que el procedimiento de medir verosimilitudes elige (casi siempre) bien el parámetro correcto, al menos en esta (limitada) comparación entre θ_0 y θ_1.

Para formalizar esta idea, consideramos muestras aleatorias $\mathbb{X} = (X_1,\dots,X_n)^\intercal$ de tamaño n de una variable X con función de densidad/masa $f(x;\theta)$, con $\theta \in \Theta$. A cada muestra aleatoria $\mathbb{X}$ le asociamos su verosimilitud

$$\textsc{vero}(\theta;X_1,\dots,X_n) = \prod_{i=1}^{n} f(X_i;\theta).$$

Sea $\theta_0 \in \Theta$. Entonces, para cualquier $\theta_1 \in \Theta$ fijo, $\theta_1 \neq \theta_0$, se tiene que

$$\mathbf{P}_{\theta_0}\big(\textsc{vero}(\theta_0;X_1,\dots,X_n) > \textsc{vero}(\theta_1;X_1,\dots,X_n)\big) \to 1 \qquad \text{cuando } n \to \infty. \tag{5.36}$$

En esta expresión importa, y mucho, el $>$ estricto. Es decir, si θ_0 es el *verdadero valor* del parámetro θ con el que se obtienen las muestras de la distribución, entonces, asintóticamente, es muy probable que concluyamos que el valor θ_0 es más verosímil que cualquier otro dado, θ_1, fijado previamente. Por ahora, obsérvese, sólo se compara θ_0 con un θ_1 prefijado, y no con todos los $\theta \in \Theta$.

Damos ahora el argumento para probar (5.36), que requerirá ciertas hipótesis sobre $f(x;\theta)$, que iremos señalando oportunamente, y que recogeremos al final en un enunciado formal.

Supongamos ya de partida, *primera hipótesis*, que

$$\mathbf{sop}_{\theta_0} = \mathbf{sop}_{\theta_1}.$$

En este caso, la cantidad

$$\ln \frac{f(x;\theta_1)}{f(x;\theta_0)}$$

toma valores reales para todo x en ese soporte común (es decir, no se va a $\pm\infty$).

Tomando logaritmos, la convergencia (5.36) equivale a

$$\mathbf{P}_{\theta_0}\Big(\sum_{i=1}^{n} \ln \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)} < 0\Big) \to 1 \qquad \text{cuando } n \to \infty.$$

O mejor, a

$$\mathbf{P}_{\theta_0}\Big(\frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)} < 0\Big) \to 1 \qquad \text{cuando } n \to \infty.$$

La ley fuerte de los grandes números, aplicada a los clones independientes

$$Y_i = \ln \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}$$

nos dice que, para todo $\varepsilon > 0$,

$$\mathbf{P}_{\theta_0}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)} - \mathbf{E}_{\theta_0}\Big(\ln \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}\Big)\Big| > \varepsilon\Big) \to 0 \quad \text{cuando } n \to \infty. \tag{5.37}$$

Esta convergencia requiere que

$$\mathbf{E}_{\theta_0}\Big(\Big|\ln \frac{f(X;\theta_1)}{f(X;\theta_0)}\Big|\Big) < \infty,$$

que incluimos ya como *segunda hipótesis*.

La desigualdad de Jensen, aplicada a la función logaritmo, véase (2.4), nos dice que

$$(\star) \qquad \mathbf{E}(\ln Z) \leq \ln \mathbf{E}(Z) \qquad \text{para una variable } Z \text{ positiva,}$$

lo que nos lleva a

$$\begin{aligned} \mathbf{E}_{\theta_0}\Big(\ln \frac{f(X;\theta_1)}{f(X;\theta_0)}\Big) &\leq \ln\Big[\mathbf{E}_{\theta_0}\Big(\frac{f(X;\theta_1)}{f(X;\theta_0)}\Big)\Big] = \ln \int_{\mathbf{sop}_{\theta_0}} \frac{f(x;\theta_1)}{f(x;\theta_0)}\, f(x;\theta_0)\,dx \\ &= \ln \int_{\mathbf{sop}_{\theta_0}} f(x;\theta_1)\,dx = \ln \int_{\mathbf{sop}_{\theta_1}} f(x;\theta_1)\,dx = \ln 1 = 0. \end{aligned} \tag{5.38}$$

La igualdad en la desigualdad de Jensen $(\star)$ se produce sólo cuando la variable Z es constante. En nuestro caso, tendremos igualdad en (5.38) solo si, para un cierto valor c,

$$\ln \frac{f(x;\theta_1)}{f(x;\theta_0)} = c\,, \quad \text{para todo } x \in \mathbf{sop}_{\theta_0} = \mathbf{sop}_{\theta_1}\,.$$

Es decir, si para todo $x \in \mathbf{sop}_{\theta_0} = \mathbf{sop}_{\theta_1}$,

$$f(x;\theta_1) = e^c f(x;\theta_0) \quad \Longleftrightarrow f(x;\theta_1) = f(x;\theta_0),$$

pues ambas son funciones de densidad, de integral 1. Obsérvese que si éste fuera el caso, aunque $\theta_0 \neq \theta_1$, las respectivas funciones de densidad serían idénticas, y no tendría sentido intentar distinguir θ_0 de θ_1; desde luego, las muestras no lo conseguirán.

Tomamos pues, como *tercera hipótesis*, que $f(x;\theta_0)$ y $f(x;\theta_1)$ no sean idénticas (en la jerga, se dice que los parámetros son *distinguibles*), y proseguimos el argumento.

Tenemos entonces que

$$\mu = \mathbf{E}_{\theta_0}\Big(\ln \frac{f(X;\theta_1)}{f(X;\theta_0)}\Big) < 0,$$

y tomando $\varepsilon = -\mu/2$ en (5.37), deducimos que

$$\mathbf{P}_{\theta_0}\Big(\frac{3\mu}{2} < \frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)} < \frac{\mu}{2}\Big) \to 1 \quad \text{cuando } n \to \infty,$$

y en particular, que

$$\mathbf{P}_{\theta_0}\Big(\frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)} < 0\Big) \to 1 \quad \text{cuando } n \to \infty,$$

como queríamos. Hemos probado:

Teorema 5.17 *Sea X con función de densidad/masa $f(x;\theta)$, con $\theta \in \Theta$. Supongamos que para $\theta_0, \theta_1 \in \Theta$ se tiene que*

1) $\mathbf{sop}_{\theta_0} = \mathbf{sop}_{\theta_1}$,
2) $\mathbf{E}_{\theta_0}(|\ln(f(X;\theta_1)/f(X;\theta_0))|) < \infty$,
3) $f(x;\theta_0) \not\equiv f(x;\theta_1)$.

Entonces

$$\mathbf{P}_{\theta_0}\big(\textsc{vero}(\theta_0; X_1,\dots,X_n) > \textsc{vero}(\theta_1; X_1,\dots,X_n)\big) \to 1 \qquad \textit{cuando } n \to \infty.$$

Es decir, si θ_0 es el verdadero parámetro de la distribución, la probabilidad de que el procedimiento de máxima verosimilitud lo elija tiende a 1 cuando el tamaño de la muestra tiende a ∞.

De este resultado se obtiene el siguiente corolario, en el que se compara la verosimilitud para un conjunto finito de posibles valores del parámetro.

Corolario 5.18 *Sea X con función de densidad/masa $f(x;\theta)$, con $\theta \in \Theta$. Supongamos que para $\theta_0, \theta_1, \dots, \theta_N \in \Theta$ se tiene que*

1) $\mathbf{sop}_{\theta_0} = \mathbf{sop}_{\theta_j}$ *para cada* $j = 1, \ldots, N$*;*
2) $\mathbf{E}_{\theta_0}(|\ln(f(X;\theta_j)/f(X;\theta_0))|) < \infty$ *para cada* $j = 1, \ldots, N$*;*
3) $f(x;\theta_0) \not\equiv f(x;\theta_j)$ *para cada* $j = 1, \ldots, N$.

Entonces, cuando $n \to \infty$*,*

$$\mathbf{P}_{\theta_0}\big(\text{VERO}(\theta_0; X_1, \ldots, X_n) > \text{VERO}(\theta_j; X_1, \ldots, X_n),\ j = 1, \ldots, N\big) \to 1. \tag{5.39}$$

La comprobación es casi directa. Para un n fijo, el suceso recogido en (5.39) es una intersección de condiciones. Su complementario es la unión (finita) de los complementarios de cada una estas condiciones, y estos conjuntos complementarios tienen probabilidades que tienden, por el teorema 5.17, a 0 cuando $n \to \infty$.

Obsérvese que este corolario nos dice que si el espacio de parámetros Θ es finito, entonces para n grande, con probabilidad muy elevada, $\text{VERO}(\theta; X_, \ldots, X_n)$ tiene máximo estricto (único) en el verdadero valor θ_0 del parámetro.

Vamos ahora a utilizar el resultado 5.18 para establecer un resultado de *consistencia* para (la sucesión de) los estimadores por máxima verosimilitud del parámetro θ, que requerirá un par de hipótesis adicionales, que esencialmente se resumen en que la función de verosimilitud sea, para cada muestra, suficientemente regular y que tenga un único máximo, para que el propio concepto de estimador por máxima verosimilitud tenga sentido.

Teorema 5.19 (Consistencia de los estimadores de máxima verosimilitud) *Sea* X *una variable aleatoria con función de densidad/masa* $f(x;\theta)$*, con* $\theta \in \Theta$*, de manera que*

1) Θ *es un intervalo,* $\Theta = (a, b) \subset (-\infty, \infty)$.
2) $\mathbf{sop}_\theta = A$*, común para todo* $\theta \in \Theta$.
3) $\mathbf{E}_\theta(|\ln(f(X;\theta')/f(X;\theta))|) < \infty$ *para cada* $\theta \neq \theta' \in \Theta$.
4) $f(x;\theta) \not\equiv f(x;\theta')$ *para cada* $\theta \neq \theta' \in \Theta$.
5) *Para cada* $(x_1, \ldots, x_n) \in A^n$*, la función* $\text{VERO}(\theta; x_1, \ldots, x_n)$ *es* $C^1(a, b)$ *y tiene un único punto crítico* $\hat{\theta}$ *en* (a, b)*, que es un máximo.*

Sea $\mathbf{emv}_\theta^{(n)} := \mathbf{emv}_\theta(X_1, \ldots, X_n)$ *el estimador por máxima verosimilitud del parámetro* θ *de la variable* X *para una muestra aleatoria* $(X_1, \ldots, X_n)$ *de tamaño* n*. Entonces, para todo* $\varepsilon > 0$*,*

$$\mathbf{P}_\theta\big(|\mathbf{emv}_\theta(X_1, \ldots, X_n) - \theta| > \varepsilon\big) \to 0 \qquad \text{cuando } n \to \infty$$

para cualquier $\theta \in \Theta$.

Es decir, la sucesión de estimadores $(\mathbf{emv}_\theta^{(n)})$ *es consistente.*

Ya hemos visto algunos ejemplos en los que la función de verosimilitud puede tener más de un máximo (lo que deja en situación ambigua la propia definición de estimación por máxima verosimilitud), no ser derivable, etc. Estas situaciones quedan descartadas con las condiciones 1), 2) y 5), que por otra parte se cumplen en los ejemplos habituales.

Demostración del teorema 5.19. Consideramos un valor genérico del parámetro, que nombramos como $\theta_0 \in (a,b)$. Podemos encontrar un entorno de θ_0 completamente incluido en (a,b), es decir, existe un entero K tal que, para todo $k \geq K$,

$$\Big(\theta_0 - \frac{1}{k}, \theta_0 + \frac{1}{k}\Big) \subset (a,b).$$

Consideramos ahora, para cada $k \geq K$ y cada $n \geq 1$, el conjunto

$$(\star) \quad S_{n,k} = \big\{(x_1,\dots,x_n) \in A^n : \text{VERO}(\theta_0; x_1,\dots,x_n) > \text{VERO}(\theta_0 - \tfrac{1}{k}; x_1,\dots,x_n) \\ \text{y } \text{VERO}(\theta_0; x_1,\dots,x_n) > \text{VERO}(\theta_0 + \tfrac{1}{k}; x_1,\dots,x_n)\big\}.$$

Para cada muestra $(x_1,\dots,x_n) \in A^n$, su función de verosimilitud tiene un único máximo, por la condición 5). Supongamos que $(x_1,\dots,x_n) \in S_{n,k}$. Esto supone que la función de verosimilitud de *esa* muestra es mayor en θ_0 que en los otros dos puntos, $\theta_0 - 1/k$ y $\theta_0 + 1/k$.Y por tanto, el punto donde se alcanza ese (único) máximo, es decir, la estimación $\hat\theta(x_1,\dots,x_n)$ de θ por máxima verosimilitud, ha de estar en el intervalo $(\theta_0 - 1/k, \theta_0 - 1/k)$.

De manera que, para todo $(x_1,\dots,x_n) \in S_{n,k}$ se tiene que

$$|\hat\theta(x_1,\dots,x_n) - \theta_0| < \frac{1}{k},$$

y por tanto

$$(\star\star) \qquad S_{n,k} \subseteq \big\{(x_1,\dots,x_n) \in A^n : |\hat\theta(x_1,\dots,x_n) - \theta_0| < 1/k\big\}.$$

Consideramos ahora el suceso, que también nombramos como $S_{n,k}$, conformado por las muestras $(X_1,\dots,X_n)$ que cumplen las condiciones descritas en $(\star)$. Por el corolario 5.18, aplicado a θ_0, $\theta_0 + 1/k$ y $\theta_0 - 1/k$, se tiene que, para cualquier $k \geq K$,

$$\mathbf{P}_{\theta_0}(S_{n,k}) \to 1 \qquad \text{cuando } n \to \infty.$$

Y, por tanto, para cualquier $k \geq K$,

$$\mathbf{P}_{\theta_0}\big(|\mathbf{emv}_\theta(X_1,\dots,X_n) - \theta_0| \geq 1/k\big) = 1 - \mathbf{P}_{\theta_0}\big(|\mathbf{emv}_\theta(X_1,\dots,X_n) - \theta_0| < 1/k\big) \\ \leq 1 - \mathbf{P}_{\theta_0}(S_{n,k}) \to 0 \qquad \text{cuando } n \to \infty,$$

donde en la última desigualdad hemos usado la inclusión de $(\star\star)$.

Esto es cierto sea cual sea θ_0, lo que concluye la demostración. ■

En realidad, los estimadores por máxima verosimilitud tienen, asintóticamente, otras buenas propiedades, más allá de la consistencia recogida en el teorema 5.19. Denotamos de nuevo por $\mathbf{emv}_\theta^{(n)} \equiv \mathbf{emv}_\theta(X_1, \dots, X_n)$ al estadístico que nos da el estimador máxima verosimilitud del parámetro θ de la variable X para una muestra aleatoria $(X_1, \dots, X_n)$ de tamaño n.

En condiciones generales, que cubren muchos de los casos de interés, pero no con total generalidad, se tiene que este estimador es:

1) Asintóticamente insesgado. Es decir,

$$\lim_{n\to\infty} \mathbf{E}_\theta(\mathbf{emv}_\theta^{(n)}) = \theta\,.$$

2) Asintóticamente eficiente. Es decir,

$$\lim_{n\to\infty} n\mathbf{V}_\theta(\mathbf{emv}_\theta^{(n)}) = \frac{1}{I_X(\theta)}\,.$$

así que

$$\mathbf{V}_\theta(\mathbf{emv}_\theta^{(n)}) \approx \frac{1}{n\,I_X(\theta)}\,, \quad \text{para } n \text{ grande}\,.$$

3) Normalidad asintótica. Esto es,

$$\sqrt{n}\left(\mathbf{emv}_\theta^{(n)} - \theta\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\Big(0, \frac{1}{I_X(\theta)}\Big) \tag{5.40}$$

Es decir, para todo $t \in \mathbb{R}$ se tiene que

$$\lim_{n\to\infty} \mathbf{P}_\theta\Big(\mathbf{emv}_\theta^{(n)} \leq \theta + \frac{t}{\sqrt{n}\sqrt{I_X(\theta)}}\Big) = \Phi(t)\,.$$

De esta última propiedad se puede deducir la consistencia de la sucesión de estimadores del teorema 5.19.

☞ **Nota** 5.4.5. El detalle de cómo (5.40) da la consistencia de la sucesión de estimadores va como sigue. Fijemos $M > 0$. Sea N tal que para $n \geq N$ se tenga que $\sqrt{n}\,\varepsilon > M$. Entonces,

$$\mathbf{P}_\theta(|\mathbf{emv}_\theta^{(n)} - \theta| < \epsilon) = \mathbf{P}_\theta(\sqrt{n}\,|\mathbf{emv}_\theta^{(n)} - \theta| < \sqrt{n}\,\varepsilon) \geq \mathbf{P}_\theta(\sqrt{n}\,|\mathbf{emv}_\theta^{(n)} - \theta| < M),$$

así que

$$\liminf_{n\to\infty} \mathbf{P}_\theta(|\mathbf{emv}_\theta^{(n)} - \theta| < \epsilon) \geq \lim_{n\to\infty} (\sqrt{n}\,|\mathbf{emv}_\theta^{(n)} - \theta| < M) = 2\Phi(M/I_X(\theta)) - 1\,,$$

y, como esto es cierto para todo $M > 0$, se deduce, haciendo $M \uparrow \infty$, que

$$\liminf_{n\to\infty} \mathbf{P}_\theta(|\mathbf{emv}_\theta^{(n)} - \theta| < \epsilon) = 1\,,$$

y, por tanto, que

$$\lim_{n\to\infty} \mathbf{P}_\theta(|\mathbf{emv}_\theta^{(n)} - \theta| < \epsilon) = 1\,.$$

EJEMPLO 5.4.12. *Comprobación de normalidad asintótica para el estimador máximo verosímil del parámetro θ de* RAY(θ).

El **emv** de θ es (ejemplo 5.2.10)

$$\mathbf{emv}_n(X_1,\dots,X_n)=\frac{1}{2}\,\overline{X^2}_{(n)}\,.$$

Como ya vimos en el ejemplo 5.4.10, usando el método delta,

$$\sqrt{n}\,\Big(\frac{1}{2}\,\overline{X^2}_{(n)}-\theta\Big)\quad\text{converge en distribución a}\quad\mathcal{N}(0,\theta^2)\,.$$

Obsérvese cuán expeditivamente se obtiene el mismo resultado sin más que apelar a la normalidad asintótica general para estimadores de máxima verosimilitud de (5.40), y recordar (ejemplo 5.3.6) que $I_X(\theta)=1/\theta^2$, ♣