

Capítulo 5

Estimación (puntual) de parámetros

5.1. Generalidades sobre estimadores	3
5.1.1. Dependencia de la distribución respecto del parámetro . .	4
5.1.2. Sesgo de un estimador. Estimadores insesgados	9
5.1.3. Varianza de un estimador	14
5.1.4. Error cuadrático medio	16
5.2. Métodos de construcción de estimadores	21
5.2.1. Método de momentos	21
5.2.2. Método de máxima verosimilitud	26
5.3. Información y cota de Cramér–Rao	41
5.3.1. Información y cantidad de información	45
5.3.2. La cota de Cramér–Rao	52
5.3.3. Complementos sobre la cota de Cramér–Rao	56
5.3.4. Información y cota de Cramér–Rao, caso de varios parámetros	64
5.4. Comportamiento asintótico de estimadores	68
5.4.1. Comportamiento asintótico de los estimadores por momentos	74
5.4.2. Comportamiento asintótico de los estimadores de máxima verosimilitud	83

En este tema (y en los dos siguientes), X será una variable aleatoria cuya distribución de probabilidad depende de (y está determinada por) un cierto parámetro θ **que desconocemos** y que nos **interesa estimar**.

Así que sólo sabemos que la cantidad X se distribuye en una población siguiendo una distribución de probabilidad de una cierta familia (exponenciales, uniformes,

geométricas, etc.), pero no sabemos de qué distribución concreta se trata, es decir, desconocemos el parámetro.

El objetivo ahora es **estimar** el valor de θ a partir de una muestra (x_1, \dots, x_n) que se supone que se ha obtenido aleatoria e independientemente, es decir, de una realización de (X_1, \dots, X_n) , donde X_1, \dots, X_n son clones independientes de X . Veamos algunos ejemplos de estos objetivos de estimación.

- $X \sim \text{BER}(p)$. El parámetro es $\theta = p$.
Por ejemplo, p puede ser el porcentaje de gente que va a votar “sí” en un referendo. Queremos estimar p a partir de una muestra de tamaño n de respuestas.
- $X \sim \text{EXP}(\lambda)$. El parámetro podría ser $\theta = \lambda$. Pero también $\theta = 1/\lambda$, dado que $\mathbf{E}(X) = 1/\lambda$.
La variable X podría, por ejemplo, describir el tiempo de espera hasta el siguiente mensaje, o lo que tarda un cliente en ser atendido en la cola del Ikea.
- $X \sim \text{POISS}(\lambda)$. El parámetro podría ser $\theta = \lambda$ o también $\theta = e^{-\lambda}$.
Por ejemplo, analizamos el número de ocurrencias X de un fenómeno relativamente raro en un intervalo de tiempo relativamente corto.
- $X \sim \mathcal{N}(0, \sigma^2)$. El parámetro aquí sería $\theta = \sigma$ o $\theta = \sigma^2$.
Por ejemplo, X puede registrar errores en la medición con un cierto aparato de medida (que se supone que en media no comete errores, es decir, que está calibrado de manera que el error medio es nulo).
- $X \sim \text{UNIF}[0, a]$ y nos interesa estimar el parámetro $\theta = a$.

Como el lector puede apreciar en estos ejemplos, el parámetro θ que queremos estimar no tiene por qué ser el *parámetro oficial* del que depende la distribución de la variable aleatoria. En una $X \sim \text{EXP}(\lambda)$ nos puede interesar el parámetro oficial $\theta = \lambda$, o la esperanza $\theta = 1/\lambda$, o la probabilidad de superar el nivel 1, que sería $\theta = e^{-\lambda}$. Como toda la distribución de X depende de λ , cualquier otro parámetro viene dado por una función de λ , en general, invertible. Una estimación de $\theta = 1/\lambda$ nos da, claro, una estimación de λ . Pero, como veremos, *no son estimaciones equivalentes* en cuanto a sus propiedades (“sesgo”, “error cuadrático”, etc.).

Aunque disponemos de libertad para elegir el parámetro que se quiere estimar, lo más habitual es que sea el parámetro “oficial” de la distribución, o quizás la función de ese parámetro que da, por ejemplo, la media o la varianza de la distribución.

En ocasiones, la forma funcional de la distribución puede depender de dos parámetros (o incluso más, aunque esto es ya más inusual). Por ejemplo,

- $X \sim \mathcal{N}(\mu, \sigma^2)$, con $\mu \in \mathbb{R}$ y $\sigma > 0$. Aquí los parámetros oficiales de la distribución son su propia media $\mathbf{E}(X) = \mu$ y varianza $\mathbf{V}(X) = \sigma^2$.
- $X \sim \text{GAMMA}(\lambda, t)$, con $\lambda, t > 0$. Los parámetros oficiales son λ y t ; mientras que la media y la varianza vienen dadas por combinaciones de estos parámetros: $\mathbf{E}(X) = t/\lambda$ y $\mathbf{V}(X) = t/\lambda^2$.

5.1. Generalidades sobre estimadores

El protocolo general de estimación de parámetros va como sigue:

- Partiremos de una **muestra empírica** (u **observada**), es decir, de una lista (x_1, \dots, x_n) de valores concretos de X (obtenidos de forma aleatoria e independiente).
- Y, con ella, calcularemos mediante una cierta función $h: \mathbb{R}^n \rightarrow \mathbb{R}$ lo que se conoce (en este contexto) como una **estimación** de θ , y que nombraremos como $\hat{\theta}$:

$$\hat{\theta} = h(x_1, \dots, x_n) \quad \Leftarrow \quad \text{ESTIMACIÓN DE } \theta$$

Por ejemplo, la función h podría ser

- $h(x_1, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n x_j = \bar{x}$, la media muestral/empírica/observada,
- $h(x_1, \dots, x_n) = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2 = s^2$, la cuasivarianza muestral/empírica/observada,
- $h(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$, el máximo muestral/empírico/observado. Etc.

Éste es el procedimiento. El dato es la muestra observada (x_1, \dots, x_n) , a partir de la cual obtenemos un valor $\hat{\theta} = h(x_1, \dots, x_n)$ que, esperamos, sea una buena estimación del desconocido parámetro θ .

Pero, ¿qué función h debemos elegir para que la estimación de θ sea adecuada? Una vez elegida una función h , ¿cuán buenas serán las estimaciones de θ obtenidas? ¿Qué grado de confianza podemos tener en que el número $\hat{\theta} = h(x_1, \dots, x_n)$ esté realmente cercano al valor de θ ?

Digamos, por ejemplo, que tenemos la siguiente muestra de ceros y unos:

$$(0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1).$$

Esta muestra de tamaño 15 ha sido obtenida a partir de una variable $X \sim \text{BER}(p)$. Pero no conocemos p . Parece natural estimar p a través de la proporción de unos, o lo que es lo mismo, haciendo la media aritmética de los 15 datos. Obtendríamos así el valor $7/15 \approx 46.67\%$ como estimación \hat{p} del desconocido parámetro p . Obsérvese que esta estimación depende de la muestra. ¿Qué grado de confianza podemos tener en ella? El verdadero valor de p bien podría ser 20%, o quizás 50%, o quizás... Aunque desde luego es razonable pensar que esté cerca del 7/15. Además, intuimos/sabemos, ¿verdad, lector?, que, si la muestra hubiera sido de tamaño 1500, la estimación obtenida habría sido más representativa del valor de p . Pero, ¿cuánto más?

Como ejemplo alternativo, digamos que tenemos una lista de 26 alturas (en cm) de individuos varones de entre 40 y 50 años de una cierta región. Partiendo de la hipótesis de que la variable altura de ese tipo de individuos sea una $\mathcal{N}(\mu, \sigma^2)$, parece razonable suponer que la media muestral de la muestra nos dé información sobre μ , y que la ¿varianza muestral? lo haga sobre σ^2 . Pero ¿cuánto?, ¿y con qué fiabilidad?

Para dar respuesta a estas preguntas, necesitamos abstraer la cuestión y considerar la *contrapartida teórica* que supone el **modelo de muestreo aleatorio**:

- Tenemos una muestra aleatoria (X_1, \dots, X_n) de clones independientes de X .

- Dada una función h , la distribución de probabilidad del estadístico

$$T = h(X_1, \dots, X_n),$$

recoge todas las posibles estimaciones del parámetro θ y sus respectivas probabilidades de ocurrencia.

Del estadístico T decimos que es un **estadístico estimador** de θ , o simplemente un **estimador** de θ .

La muestra aleatoria (abstracta) (X_1, \dots, X_n) recoge todos los posibles “escenarios” de valores (x_1, \dots, x_n) con sus respectivas probabilidades. Entendemos que la muestra concreta (x_1, \dots, x_n) de la que obtendremos la estimación de θ es una *realización* o una *materialización* de (X_1, \dots, X_n) .

Con el estadístico $T = h(X_1, \dots, X_n)$ recogemos también todas las posibles estimaciones $\hat{\theta} = h(x_1, \dots, x_n)$ de θ con sus respectivas probabilidades. Y de nuevo entendemos que la estimación $h(x_1, \dots, x_n)$ es una realización del estimador T .



Nota 5.1.1. Hacemos notar que usamos aquí los mismos términos para la situación empírica y para el modelo teórico. Por ejemplo, con “muestra” designamos tanto al vector aleatorio (X_1, \dots, X_n) como a una cualquiera de sus realizaciones, la lista de números (x_1, \dots, x_n) ; con el término “media muestral” designamos tanto al *número* $\frac{1}{n} \sum_{i=1}^n x_i$ como a la *variable aleatoria* $\frac{1}{n} \sum_{i=1}^n X_i$, etc.

En desmesurado afán clarificador, podríamos ir añadiendo los adjetivos “empírico” u “observado” cada vez que estuviéramos tratando con realizaciones, pero en la práctica no será necesario, porque habitualmente el contexto deja claro si estamos con el modelo teórico o con sus realizaciones. No olvidemos tampoco la distinción tipográfica en el uso de mayúsculas (para variables aleatorias) y minúsculas (para números).

Parece natural exigir que un estadístico T tenga ciertas propiedades: por ejemplo, es deseable que “apunte” en la dirección correcta, es decir, que *en media* dé el valor correcto de θ . Si en el ejemplo de ceros y unos anterior tomáramos como estadístico el mínimo valor de la muestra, lo más seguro es que devolviera un 0, lo que con casi toda seguridad sería una mala estimación de p . Además nos gustaría que la dispersión del estadístico T respecto de la media (es decir, la varianza de T) fuera pequeña, porque si éste fuera el caso tendríamos una cierta (alta) confianza en que la estimación que se obtiene no se desvía mucho del verdadero y desconocido valor del parámetro.

En el resto de la sección vamos a estudiar estadísticos estimadores de parámetros como lo que son, objetos aleatorios, para poder comparar su valía y utilidad en función de las propiedades que tengan. Para ello, necesitaremos cierta notación.

5.1.1. Dependencia de la distribución respecto del parámetro

El objeto de estudio es una variable aleatoria X cuya distribución de probabilidad depende de un parámetro θ . Usaremos en lo que sigue de manera genérica y unificada la notación

$$f(x; \theta)$$

para representar la función de densidad o la función de masa (dependiendo de si la variable es continua o discreta) de X . La expresión $f(x; \theta)$ anterior tiene dos argumentos: x y θ . Por un lado, el argumento x recorre los valores de la variable X . El segundo argumento θ recorre los posibles valores del parámetro. Observe, lector, que, a valor de θ le corresponde una función de densidad o de masa distinta específica¹.

Llamaremos **espacio de parámetros** $\Theta \subset \mathbb{R}$ al conjunto de *posibles valores del parámetro θ de interés*. Habitualmente, Θ será un intervalo de la recta real, pero también podría ser un conjunto finito (o numerable). Si la distribución dependiera de, por ejemplo, dos parámetros, entonces Θ sería un cierto subconjunto (por ejemplo un semiplano, un rectángulo) de \mathbb{R}^2 .

Estando, por otro lado, interesados sólo en aquellos valores x donde $f(x; \theta) > 0$. Para cada $\theta \in \Theta$, denotamos por **sop $_{\theta}$** al conjunto de valores de x donde $f(x; \theta) > 0$. Al conjunto **sop $_{\theta}$** se le llama **soporte de X si se da θ** . Incluimos un subíndice θ porque el soporte podría depender del parámetro, como veremos en algún ejemplo, aunque lo más habitual es que el soporte sea fijo y no dependa de θ . El **sop $_{\theta}$** será un intervalo o unión de intervalos (finitos o no) de \mathbb{R} para variables continuas, o un subconjunto numerable de \mathbb{R} en el caso de discretas (en los modelos más usuales, un subconjunto de $\{0, 1, 2, \dots\}$, y para variables finitas, un subconjunto finito de \mathbb{R}).

Listamos a continuación unos cuantos modelos que usaremos recurrentemente como ilustración en lo que sigue, detallando la expresión de la función de masa/densidad (indicando el soporte) y el espacio de parámetros. Aunque generalmente usaremos x para los valores de la variable, y θ para el parámetro, en ocasiones la tradición obliga a usar otros símbolos: por ejemplo, en la distribución geométrica el parámetro es p (y no θ), y los posibles valores son k (en lugar de x), por tratarse de enteros.

- **BER(p)**:

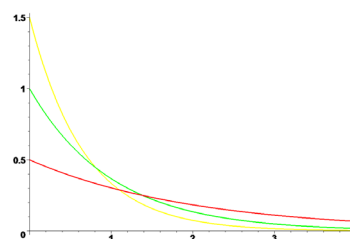
$$f(k; p) = \begin{cases} p, & \text{si } k = 1, \\ (1 - p), & \text{si } k = 0, \end{cases}$$

para $p \in \Theta = (0, 1)$. El soporte es **sop $_p$** = $\{0, 1\}$, para todo $p \in \Theta$.

- **EXP(λ)**:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{si } x \geq 0, \\ 0, & \text{si } x < 0, \end{cases}$$

para $\lambda \in \Theta = (0, +\infty)$. El soporte es **sop $_{\lambda}$** = $[0, +\infty)$, para todo $\lambda \in \Theta$. A la derecha de estas líneas dibujamos las funciones de densidad para los casos $\lambda = 1/2$, $\lambda = 1$ y $\lambda = 3/2$.



- **POISS(λ)**:

$$f(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \text{ entero}, k \geq 0,$$

para $\lambda \in \Theta = (0, +\infty)$. El soporte es **sop $_{\lambda}$** = $\{0, 1, 2, \dots\}$, para todo $\lambda \in \Theta$.

¹Hay quien escribe $f_{\theta}(x)$, o mejor, $f(x|\theta)$, la función de densidad *condicionada* al valor de θ .

- $\text{GEO}(p)$:

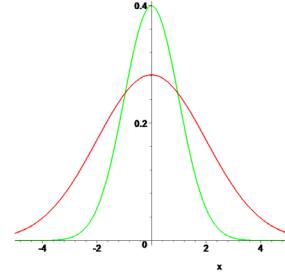
$$f(k; p) = p(1 - p)^{k-1}, \quad k \text{ entero}, k \geq 1,$$

para $p \in \Theta = (0, 1)$. El soporte es $\text{sop}_p = \{1, 2, \dots\}$, para todo $p \in \Theta$.

- $\mathcal{N}(0, \sigma^2)$.

$$f(x; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)}, \quad x \in \mathbb{R},$$

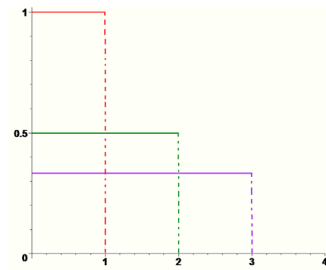
para $\sigma^2 \in \Theta = (0, +\infty)$. El soporte es $\text{sop}_{\sigma^2} = \mathbb{R}$. A la derecha, las funciones de densidad de los casos $\sigma^2 = 1$ y $\sigma^2 = 2$.



- $\text{UNIF}[0, a]$.

$$f(x; a) = \begin{cases} 1/a & \text{si } x \in (0, a), \\ 0 & \text{si } x \notin (0, a). \end{cases}$$

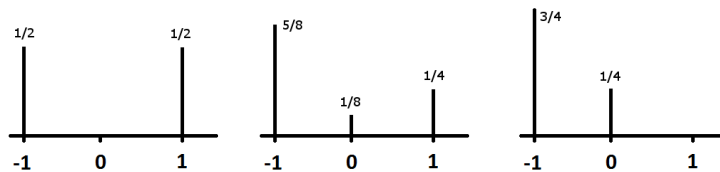
para $a \in \Theta = (0, +\infty)$. Aquí, $\text{sop}_a = [0, a]$ es un intervalo que depende del parámetro a . El dibujo recoge las funciones de densidad correspondientes a los casos $a = 1$, $a = 2$ y $a = 3$.



- Una variable discreta que toma tres valores:

$$f(k; \theta) = \begin{cases} (2 + \theta)/4, & \text{si } k = -1, \\ \theta/4, & \text{si } k = 0, \\ (2 - 2\theta)/4, & \text{si } k = 1, \end{cases}$$

para $\theta \in \Theta = [0, 1]$. El soporte es $\text{sop}_\theta = \{-1, 0, 1\}$, para todo $\theta \in \Theta$. Dibujamos a continuación las funciones de masa para los casos $\theta = 0$, $\theta = 1/2$ y $\theta = 1$:



- O quizás

$$f(k; \theta) = \begin{cases} (2 + \theta)/4, & \text{si } k = -1 - \theta, \\ \theta/4, & \text{si } k = 0, \\ (2 - 2\theta)/4, & \text{si } k = 1 + \theta, \end{cases}$$

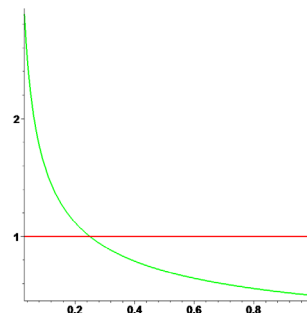
para $\theta \in \Theta = [0, 1]$. Ahora el soporte es $\text{sop}_\theta = \{-1 - \theta, 0, 1 + \theta\}$, y depende del parámetro $\theta \in \Theta$. Las funciones de masa para los casos $\theta = 0$, $\theta = 1/2$ y $\theta = 1$ tendrían un aspecto similar al del ejemplo anterior, aunque los soportes respectivos serían $\{-1, 0, 1\}$, $\{-3/2, 0, 3/2\}$ y $\{-2, 0, 2\}$.

• Podría darse también la situación en la que sólo hay unos cuantos valores alternativos del parámetro, es decir, que Θ es un conjunto finito y no un intervalo. Por ejemplo, cuando $\Theta = \{0, 1\}$ y, para $\theta = 0$, la función de densidad es

$$f(x; 0) = 1 \quad \text{si } x \in (0, 1) \quad (\text{y } f(x; 0) = 0 \text{ si } x \notin (0, 1)),$$

mientras que para $\theta = 1$ es

$$f(x; 1) = \frac{1}{2\sqrt{x}}, \quad \text{si } x \in (0, 1) \quad (\text{y } f(x; 1) = 0 \text{ si } x \notin (0, 1)).$$



El soporte es $\text{sop}_\theta = (0, 1)$ para ambos valores del parámetro.

• Si la distribución tiene dos parámetros, el espacio Θ será un subconjunto del plano. Por ejemplo, para la normal $\mathcal{N}(\mu, \sigma^2)$ se tiene:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R},$$

para $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, +\infty)$. El soporte es $\text{sop}_{\mu, \sigma^2} = \mathbb{R}$, para $(\mu, \sigma^2) \in \Theta$.

Recordemos que el parámetro de interés, el que queremos estimar, podría no ser el parámetro “oficial” de la distribución. Por ejemplo, en una $\text{EXP}(\lambda)$ pudiera interesarnos la esperanza $\theta = 1/\lambda$ en lugar de λ . En este caso, para los cálculos, podría ser más conveniente usar la representación

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}; & x \geq 0, \\ 0; & x < 0. \end{cases}$$

donde $\theta \in \Theta = (0, +\infty)$. O en una $\text{POISS}(\lambda)$, si en lugar de λ nos interesara estimar $\theta = e^{-\lambda} = \mathbf{P}(X = 0)$, usaríamos

$$f(k; \theta) = \theta \frac{(-1)^k \ln(\theta)^k}{k!}, \quad k \text{ entero, } k \geq 0,$$

donde $\theta \in \Theta = (0, 1)$.

A. Cálculo de medias de estadísticos

En lo que sigue calcularemos medias o esperanzas usando la función de densidad/masa $f(x; \theta)$. Para insistir en que para cada valor de θ tenemos una densidad/masa distinta (un modelo distinto), y por tanto el resultado dependerá del valor de θ , usaremos la notación \mathbf{E}_θ . Las medias a las que nos referimos son de dos tipos.

A1. Medias de (funciones de) la variable. Dada una variable X con función de densidad/masa $f(x; \theta)$, nos interesará calcular medias de variables $Y = g(X)$, donde g es cierta función.

Si X es una variable continua tendremos que

$$\mathbf{E}_\theta(g(X)) = \int_{\mathbb{R}} g(x) f(x; \theta) dx = \int_{\text{sop}_\theta} g(x) f(x; \theta) dx.$$

En el caso en que sea discreta, el cálculo será

$$\mathbf{E}_\theta(g(X)) = \sum_{x \in \text{sop}_\theta} g(x) f(x; \theta).$$

Usaremos, como corresponde, la notación \mathbf{V}_θ para la varianza de X :

$$\mathbf{V}_\theta(X) = \mathbf{E}_\theta(X^2) - \mathbf{E}_\theta(X)^2.$$

También utilizaremos la notación \mathbf{P}_θ para probabilidades calculadas con $f(x; \theta)$.

Insistimos, por si no ha quedado claro: en las notaciones anteriores, el subíndice θ significa que el cálculo se realiza *suponiendo* (o *condicionando a*) que el “verdadero” valor del parámetro fuera θ . Por ejemplo, la cantidad $\mathbf{E}_\theta(X^2)$ significa el valor de la media/esperanza de la variable X suponiendo que X tuviera la función de densidad/masa $f(x; \theta)$.

Por, ejemplo, si $X \sim \text{EXP}(\lambda)$ y $g(x) = x^2$, para todo $x \in \mathbb{R}$, entonces

$$\mathbf{E}_\lambda(g(X)) = \mathbf{E}_\lambda(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}.$$

A2. Medias de estadísticos. La función de densidad/masa de la muestra aleatoria $\mathbb{X} = (X_1, \dots, X_n)$ viene dada por

$$f(\mathbf{x}; \theta) = f(x_1, \dots, x_n; \theta) = \prod_{j=1}^n f(x_j; \theta)$$

para $\theta \in \Theta$ y $\mathbf{x} \in (\text{sop}_\theta)^n$. Fuera de $(\text{sop}_\theta)^n$, se tiene $f(\mathbf{x}; \theta) \equiv 0$.

En muchas ocasiones calcularemos medias de variables del tipo $h(\mathbb{X})$ (es decir, de estadísticos). Usaremos la notación

$$\mathbf{E}_\theta(h(\mathbb{X})) = \mathbf{E}_\theta(h(X_1, \dots, X_n))$$

para describir el valor de la media/esperanza de la variable aleatoria $h(X_1, \dots, X_n)$ *suponiendo* que se trata de muestras aleatorias (X_1, \dots, X_n) de la variable X que sigue una función de densidad/masa $f(x; \theta)$. Al variar θ obtenemos, claro, una función de θ . Perdón por la insistencia.

El cálculo en sí, cuando X es continua, se plantea como sigue:

$$\begin{aligned}\mathbf{E}_\theta(h(\mathbb{X})) &= \int_{\mathbb{R}^n} h(\mathbf{x}) \cdot f(\mathbf{x}; \theta) \, d\mathbf{x} \\ &= \int_{\mathbb{R}^n} h(x_1, \dots, x_n) \cdot f(x_1, \dots, x_n; \theta) \, dx_1 dx_2 \cdots dx_n \\ &= \int_{(\mathbf{so}_\theta)^n} h(x_1, \dots, x_n) \cdot \left(\prod_{j=1}^n f(x_j; \theta) \right) \, dx_1 dx_2 \cdots dx_n.\end{aligned}$$

Mientras que si X es discreta, tendremos

$$\mathbf{E}_\theta(h(\mathbb{X})) = \sum_{\mathbf{x} \in \mathbf{so}_\theta^n} h(\mathbf{x}) f(\mathbf{x}; \theta) = \sum_{(x_1, \dots, x_n) \in \mathbf{so}_\theta^n} h(x_1, \dots, x_n) \prod_{j=1}^n f(x_j; \theta).$$

Sólo cuando la función h tenga una forma muy particular, será posible evaluar explícitamente las integrales/sumas múltiples anteriores. Y en estos casos, a veces no es preciso usar toda esa notación. Como ejemplos, si $h(x_1, \dots, x_n) = \sum_{j=1}^n \alpha_j x_j$, donde $\alpha_1, \dots, \alpha_n$ son números dados, usando la linealidad de la esperanza tenemos

$$\mathbf{E}_\theta(h(\mathbb{X})) = \mathbf{E}_\theta\left(\sum_{j=1}^n \alpha_j X_j\right) = \sum_{j=1}^n \alpha_j \mathbf{E}_\theta(X_j) = \mathbf{E}_\theta(X) \sum_{j=1}^n \alpha_j;$$

o si $h(x_1, \dots, x_n) = x_1 \cdots x_n$, entonces, usando la independencia,

$$\mathbf{E}_\theta(h(\mathbb{X})) = \mathbf{E}_\theta(X_1 \cdots X_n) = \mathbf{E}_\theta(X_1) \cdots \mathbf{E}_\theta(X_n) = \mathbf{E}_\theta(X)^n.$$

Si tomamos $h(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$, y por ejemplo X es una variable continua, entonces

$$\mathbf{E}_\theta(\max(X, \dots, X_n)) = \int_{(\mathbf{so}_\theta)^n} \max(x_1, \dots, x_n) \cdot \left(\prod_{j=1}^n f(x_j; \theta) \right) \, dx_1 dx_2 \cdots dx_n.$$

Pero en este caso suele ser mucho más directo calcular esta integral aprovechando que conocemos la distribución del máximo de una muestra (apartado 4.4).

5.1.2. Sesgo de un estimador. Estimadores insesgados

La primera propiedad deseable de un estimador T es que al menos en media no se equivoque. Decimos que $T = h(X_1, \dots, X_n)$ es un **estimador insesgado** del parámetro θ si

$$\boxed{\mathbf{E}_\theta(T) = \theta} \quad \text{para cualquier valor de } \theta \in \Theta.$$

En palabras, el estimador T es insesgado si la media de T es exactamente θ , cuando se supone que el parámetro con el que se generan las muestras es θ , y esto sea cual sea el valor de θ .

Recalcamos el “para todo $\theta \in \Theta$ ” en la definición anterior. Podría darse el caso de que la media de un cierto estimador coincidiera con θ para algún valor de θ , pero no para todos (véase el ejemplo 5.1.8). Esto no sería interesante, porque, recordemos, el parámetro θ es desconocido, y necesitamos que la ausencia de sesgo se verifique para cualquier valor del parámetro.

Si no se cumple que $\mathbf{E}_\theta(T) = \theta$ para todo $\theta \in \Theta$, decimos que T es un estimador sesgado de θ . La diferencia entre $\mathbf{E}_\theta(T)$ y θ es conocida como el **sesgo** del estimador:

$$(5.1) \quad \text{sesgo}_\theta(T) = \mathbf{E}_\theta(T) - \theta.$$

Observe, lector, que el sesgo es una función de $\theta \in \Theta$. Si $\text{sesgo}_\theta(T) = 0$ para todo $\theta \in \Theta$, T será insesgado; si $\text{sesgo}_\theta(T) > 0$ para todo $\theta \in \Theta$, entonces diremos que el estimador T está sesgado “al alza”, mientras que estará sesgado “a la baja” si $\text{sesgo}_\theta(T) < 0$ para todo $\theta \in \Theta$.

Veamos ahora algunos ejemplos. Los dos primeros son generales. Los parámetros que interesa estimar son, respectivamente, la esperanza $\mathbf{E}(X)$ y la varianza $\mathbf{V}(X)$ de una variable aleatoria X genérica. No usaremos aquí subíndices que hagan alusión a estos parámetros.

EJEMPLO 5.1.1. *Sea X una variable aleatoria cualquiera. Para estimar el parámetro $\mu = \mathbf{E}(X)$ usamos el estadístico media muestral \bar{X} .*

La media muestral es un estimador insesgado de μ , pues siempre se tiene

$$\mathbf{E}(\bar{X}) = \mu,$$

por la proposición 4.1. Así que si tenemos la muestra x_1, \dots, x_n , entendemos que \bar{x} es una estimación de $\mathbf{E}(X)$. ♣

EJEMPLO 5.1.2. *Sea X una variable aleatoria cualquiera. Para estimar el parámetro $\sigma^2 = \mathbf{V}(X)$ usamos la cuasivarianza muestral S^2 .*

La cuasivarianza muestral es un estimador insesgado de σ^2 , pues siempre tenemos (proposición 4.3) que

$$\mathbf{E}(S^2) = \sigma^2$$

(nótese que, por tanto, la varianza muestral es un estimador *sesgado* de σ^2).

Sin embargo, la cuasidesviación típica muestral S es un estimador sesgado de σ , pues $\mathbf{E}(S)^2 < \mathbf{E}(S^2)$ salvo si S es constante (que solo ocurre si X es constante).



Nota 5.1.2. La desigualdad de Cauchy–Schwarz (teorema 2.3) nos dice que $\mathbf{E}(S) = \mathbf{E}(S \cdot 1) < \mathbf{E}(S^2)^{1/2} \mathbf{E}(1^2)^{1/2} = \mathbf{E}(S^2)^{1/2}$, salvo si S es constante.

Si tenemos la muestra x_1, \dots, x_n , entendemos que $s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$ es una estimación de σ^2 . ♣

EJEMPLO 5.1.3. $X \sim \mathcal{N}(\mu, \sigma^2)$.

Apelando a los ejemplos generales 5.1.1 y 5.1.2, los estadísticos \bar{X} y S^2 son estimadores insesgados de μ y de σ^2 respectivamente, pues μ es la media de X , y σ^2 su varianza. ♣



Nota 5.1.3. Sesgo y no sesgo. Supongamos que se ha obtenido un número N grande de realizaciones independientes de (X_1, \dots, X_5) , pongamos (x_1^j, \dots, x_5^j) para $1 \leq j \leq N$. Nótese que $n = 5$. Para cada j obtenemos una realización de S^2 , pongamos s_j^2 , cada una de las cuales es una estimación de σ^2 .

Como estimación final de σ tenemos dos rutas:

- 1) Tomar un promedio de los s_j^2 y luego tomar la raíz cuadrada.
- 2) Tomar raíz cuadrada primero, y luego, tomar promedio de los s_j .

Con la primera obtenemos (por Grandes Números) una buena estimación de $\mathbf{E}(S^2)$, es decir, de σ^2 , y luego de σ . Con la segunda obtenemos (por Grandes Números) una buena estimación de $\mathbf{E}(S)$, que es inferior a σ .

Digamos, por ejemplo, que $X \sim \mathcal{N}(\mu, \sigma^2)$. Para muestras de tamaño n de X se tiene que

$$\mathbf{E}(S) = \sqrt{\frac{2}{n-1}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \sigma$$

(véase el lema 3.10). Para el caso $n = 5$, esto da $\mathbf{E}(S) = (3/8)\sqrt{2\pi}\sigma \approx 0.93\sigma$.

EJEMPLO 5.1.4. *Supongamos $X \sim \text{BER}(p)$. Desconocemos p y queremos estimarlo.*

Obsérvese que $\mathbf{E}_p(X) = p$ y que $\mathbf{V}_p(X) = p(1-p)$. Tomamos una muestra x_1, \dots, x_n .

Para estimar p insesgadamente podemos tomar $\frac{1}{n} \sum_{j=1}^n x_j$, la proporción de unos en la muestra x_1, \dots, x_n . Es decir, usar el estimador \bar{X} .

Como alternativa podríamos considerar S^2 , que es un estimador insesgado de $\mathbf{V}_p(X) = p(1-p)$, para a partir de ahí, “despejar” una estimación de p . Esta propuesta alternativa tiene dos problemas, 1) al despejar p de la ecuación $p - p^2 = \mathbf{V}_p(X)$ tendríamos dos raíces como estimación de p , 2) la estimación así obtenida sería sesgada. ¡Vaya! ♣

EJEMPLO 5.1.5. *Supongamos que $X \sim \text{EXP}(\lambda)$. Queremos estimar λ .*

Como $\mathbf{E}_\lambda(X) = 1/\lambda$, tendríamos que $T_1 = \bar{X}$ sería un estimador insesgado, no del parámetro oficial λ , sino de $1/\lambda$.



Nota 5.1.4. Si decidimos que el parámetro de la distribución es $\theta = 1/\lambda$, entonces tendríamos que $\theta \in (0, \infty)$, y la función de densidad se escribiría como $f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$ para $x > 0$. Con esta notación, tendríamos que $\mathbf{E}_\theta(X) = \theta$, y por tanto \bar{X} sería estimador insesgado de θ .

Pero si lo que pretendemos es estimar el parámetro oficial λ , entonces sería natural considerar el estimador $T_2 = 1/\bar{X}$ que, ¡atención!, sería un estimador *sesgado* de λ . De hecho, $\mathbf{E}_\lambda(1/\bar{X}) > \lambda$. Véase la nota 5.1.5 siguiente.



Nota 5.1.5. Usando de nuevo la desigualdad de Cauchy–Schwarz (teorema 2.3): como $1 = \mathbf{E}_\lambda(1) = \mathbf{E}_\lambda(\sqrt{\overline{X}} \cdot 1/\sqrt{\overline{X}}) < \mathbf{E}_\lambda(\overline{X}) \mathbf{E}_\lambda(1/\overline{X}) = \frac{1}{\lambda} \mathbf{E}_\lambda(1/\overline{X})$, tenemos que $\mathbf{E}_\lambda(1/\overline{X}) > \lambda$. Así que en media las estimaciones de λ con $T_2 = 1/\overline{X}$ serían sesgadas al alza.

Más detalle: como en este caso $n\overline{X}$ es una $\text{GAMMA}(\lambda, n)$ (recuérdese la relación entre exponenciales y variables Gamma de la sección 2.3.4), tenemos que

$$\mathbf{E}_\lambda(1/\overline{X}) = \frac{n}{n-1} \lambda,$$

por (2.25). Obsérvese cómo, efectivamente, $\mathbf{E}_\lambda(1/\overline{X}) > \lambda$, aunque cuando n es grande, el sesgo de T_2 es muy pequeño. El cálculo anterior nos dice que el estimador $\tilde{T}_2 = \frac{n-1}{n} \frac{1}{\overline{X}}$ sería un estimador insesgado de λ .

Consideremos, por otro lado, el estadístico

$$T_3 = \min(X_1, \dots, X_n).$$

Como $X \sim \text{EXP}(\lambda)$, se tiene que $\mathbf{P}_\lambda(X > t) = e^{-\lambda t}$, para todo $t > 0$. Así que, usando (4.11), tenemos que

$$\mathbf{P}_\lambda(T_3 > t) = \mathbf{P}_\lambda(X > t)^n = e^{-nt} \quad \text{para todo } t > 0.$$

Es decir, $T_3 \sim \text{EXP}(n\lambda)$: el mínimo de clones independientes de exponenciales es también una exponencial.

De manera que $\mathbf{E}_\lambda(T_3) = 1/(n\lambda)$ y $\mathbf{E}_\lambda(nT_3) = 1/\lambda$. Así que $\tilde{T}_3 = nT_3$ es un estimador insesgado de $1/\lambda$. ♣

EJEMPLO 5.1.6. *Tenemos $X \sim \text{POISS}(\lambda)$, con $\lambda > 0$. Planteamos dos estimaciones de parámetros:*

- a) *Estimar el parámetro oficial λ .*
- b) *Estimar el parámetro $\mu = \mathbf{P}(X = 0)$. Obsérvese que μ , que está entre 0 y 1, determina la distribución de probabilidad, pues $e^{-\lambda} = \mu$, es decir, $\lambda = \ln(1/\mu)$.*

Para el caso a), como $\lambda = \mathbf{E}_\lambda(X)$, tenemos que $\mathbf{E}_\lambda(\overline{X}) = \mathbf{E}_\lambda(X) = \lambda$. Así que $T_1 = \overline{X}$ es un estimador insesgado de λ .

Para la estimación b), planteamos dos alternativas.

b.1) Como λ se puede estimar con \overline{X} y $\mu = e^{-\lambda}$, podríamos considerar el estadístico $T_2 = e^{-\overline{X}}$ para estimar μ . Por comodidad, en lugar de realizar los cálculos con μ , los haremos con el parámetro original λ , para traducir al final.

Como $n\overline{X} \sim \text{POIS}(n\lambda)$ (ejemplo 4.2.2), tenemos que

$$\mathbf{P}_\lambda(\overline{X} = k/n) = e^{-n\lambda} \frac{(n\lambda)^k}{k!}, \quad \text{para cada entero } k \geq 0.$$

Así que

$$\mathbf{P}_\lambda(T_2 = e^{-k/n}) = e^{-n\lambda} \frac{(n\lambda)^k}{k!}, \quad \text{para cada entero } k \geq 0$$

y, por tanto.

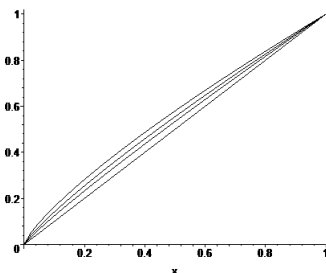
$$\begin{aligned}\mathbf{E}_\lambda(T_2) &= \sum_{k=0}^{\infty} e^{-k/n} e^{-n\lambda} \frac{(n\lambda)^k}{k!} = e^{-n\lambda} \sum_{k=0}^{\infty} \frac{(e^{-1/n} n\lambda)^k}{k!} \\ &= e^{-n\lambda} e^{n\lambda e^{-1/n}} = e^{-\lambda n(1-e^{-1/n})}.\end{aligned}$$

Deducimos finalmente que

$$\mathbf{E}_\mu(T_2) = \mu^{n(1-e^{-1/n})}.$$

- Como $\lim_{n \rightarrow \infty} n(1 - e^{-1/n}) = 1$, vemos que T_2 tiende a ser un estimador insesgado de μ cuando $n \rightarrow \infty$. Es decir, T_2 es “asintóticamente insesgado”.
- Para un n dado, T_2 es estimador sesgado (al alza) de μ , puesto que $\mu < 1$ y $n(1 - e^{-1/n}) < 1$ (que se sigue de que $1 + x \leq e^x$, para todo $x \in \mathbb{R}$).

En el dibujo de la derecha representamos las gráficas de $\mathbf{E}_\mu(T_2)$ para $n = 2$, $n = 3$ y $n = 10$, junto con la función $g(\mu) = \mu$, la bisectriz, que correspondería a sesgo 0. Obsérvese cuán rápidamente se va perdiendo el sesgo según crece n .



b.2) Como estimador alternativo, podemos considerar el estadístico T_3 que cuenta la proporción de ceros en (X_1, \dots, X_n) :

$$T_3 = \frac{1}{n} \# \{1 \leq i \leq n : X_i = 0\}.$$

La variable nT_3 es $\text{BIN}(n, \mu)$, de manera que T_3 es un estimador insesgado de μ . ♣

EJEMPLO 5.1.7. Supongamos que $X \sim \text{UNIF}[0, a]$. Desconocemos a , que es el parámetro que pretendemos estimar.

Consideremos el estimador

$$T_1 = \max(X_1, \dots, X_n).$$

Como por (4.8) para cada $x \in [0, a]$ se tiene que

$$\mathbf{P}_a(T_1 \leq x) = \mathbf{P}_a(X \leq x)^n = \left(\frac{x}{a}\right)^n,$$

la función de densidad de T_1 resulta ser

$$f_{T_1}(x; a) = \begin{cases} n \frac{x^{n-1}}{a^n}, & \text{si } 0 \leq x \leq a, \\ 0, & \text{en caso contrario,} \end{cases}$$

y, por tanto,

$$\mathbf{E}_a(T_1) = \int_0^a x n \frac{x^{n-1}}{a^n} dx = \frac{n}{n+1} a = a - \overbrace{a \frac{1}{n+1}}^{=\text{sesgo}_a}.$$

Así que T_1 es estimador sesgado (a la baja) de a , mientras que

$$\tilde{T}_1 = \frac{n+1}{n} T_1$$

es un estimador insesgado de a .

Podemos tomar como estimador alternativo del parámetro a al estadístico $T_2 = 2\bar{X}$, porque como $\mathbf{E}_a(X) = a/2$, resulta que $\mathbf{E}_a(T_2) = \mathbf{E}_a(2\bar{X}) = 2\mathbf{E}_a(\bar{X}) = 2\mathbf{E}_a(X) = a$, de manera que T_2 es un estimador insesgado del parámetro a . ♣

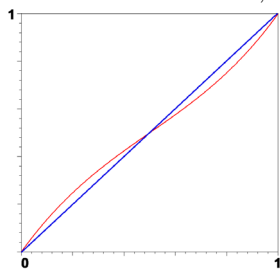
En este siguiente ejemplo se exhibe un estimador T de un parámetro θ para el que $\mathbf{E}_\theta(T)$ coincide con θ solo para algunos valores de $\theta \in \Theta$; de hecho, para un único² valor de θ .

EJEMPLO 5.1.8. Partimos de $X \sim \text{BER}(p)$. Desconocemos p y queremos estimarlo con una muestra de tamaño 3.

Proponemos el estimador

$$T(X_1, X_2, X_3) = \frac{\max(X_1, X_2, X_3) + \min(X_1, X_2, X_3)}{2}.$$

El estimador toma el valor 1 con probabilidad p^3 (solo cuando $X_1 = X_2 = X_3 = 1$), el valor 0 con probabilidad $(1-p)^3$ (cuando $X_1 = X_2 = X_3 = 0$), y el valor $1/2$ en el resto de los casos, es decir, con probabilidad $1 - p^3 - (1-p)^3$.



De manera que

$$\mathbf{E}_p(T) = p^3 + \frac{1}{2}(1-p^3 - (1-p)^3) = \frac{1}{2} + \frac{1}{2}(p^3 - (1-p)^3).$$

Obsérvese, en la figura de la izquierda, que $\mathbf{E}_p(T) = p$ para $p = 1/2$, pero $\mathbf{E}_p(T) \neq p$ para el resto de los valores de p . ♣

5.1.3. Varianza de un estimador

Para un estimador insesgado y, en general, para cualquier estimador, es deseable que su varianza sea pequeña. Si $\mathbf{V}_\theta(T)$ es pequeña, por la desigualdad de Chebyshev tendremos que es poco probable que sus realizaciones se diferencien mucho de $\mathbf{E}_\theta(T)$,

²Casi como un reloj parado que da la hora bien dos veces al día.

así que si además T es estimador insesgado, será muy probable que las estimaciones de θ estén próximas a θ :

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

De manera que, por ejemplo,

$$\mathbf{P}_\theta(|T - \theta| \leq 10\sqrt{\mathbf{V}_\theta(T)}) \geq 99\%.$$

Así que es muy probable (99 %) que la estimación dada por una realización de T no se aleje más de $10\sqrt{\mathbf{V}_\theta(T)}$ del valor del parámetro a estimar θ . Pero, claro, esto es relevante sólo si $\mathbf{V}_\theta(T)$ es pequeño.

Un estimador con mucha varianza es poco útil, porque tendremos poca confianza en sus estimaciones. Por eso se usan muestras de tamaño n , a ser posible, grande. Una muestra X_1 de tamaño $n = 1$ es un estimador insesgado de $\mathbf{E}_\theta(X)$, pero con una muestra de tamaño n el estimador \bar{X} es asimismo insesgado pero $\mathbf{V}_\theta(\bar{X}) = \mathbf{V}_\theta(X)/n$.

Si T_1 y T_2 son estimadores *insesgados* del mismo parámetro θ , decimos que T_1 es **más eficiente** que T_2 si, para todo $\theta \in \Theta$,

$$\mathbf{V}_\theta(T_1) < \mathbf{V}_\theta(T_2) \quad \text{para cualquier valor de } \theta \in \Theta.$$

Desde luego, si tanto T_1 como T_2 son estimadores insesgados de θ , preferimos aquél que tenga menos varianza. Para estimadores sesgados la preferencia ya no está tan clara. Véase el apartado 5.1.4.

EJEMPLO 5.1.9. UNIF[0, a]. Retomamos el ejemplo 5.1.7.

En el ejemplo 5.1.7 considerábamos dos estimadores del parámetro a de una variable $X \sim \text{UNIF}[0, a]$:

$$T_1 = \frac{n+1}{n} \max(X_1, \dots, X_n) \quad \text{y} \quad T_2 = 2\bar{X}.$$

Ambos son estimadores insesgados del parámetro a . Para T_2 tenemos que

$$\mathbf{V}_a(T_2) = 4 \mathbf{V}_a(\bar{X}) = \frac{4}{n} \mathbf{V}_a(X) = \frac{4}{n} \frac{a^2}{12} = \frac{a^2}{3n}.$$

Para T_1 tenemos

$$\mathbf{E}_a(T_1^2) = \left(\frac{n+1}{n}\right)^2 \int_0^a x^2 n \frac{x^{n-1}}{a^n} dx = \frac{(n+1)^2}{n(n+2)} a^2,$$

de manera que

$$\mathbf{V}_a(T_1) = \left[\frac{(n+1)^2}{n(n+2)} - 1 \right] a^2 = \frac{1}{n(n+2)} a^2.$$

Así que T_1 es más eficiente que T_2 , pues $\mathbf{V}_a(T_1) < \mathbf{V}_a(T_2)$ para todo $a > 0$ (y para cada $n \geq 1$). ♣

5.1.4. Error cuadrático medio

Si $T = h(X_1, \dots, X_n)$ es un estimador del parámetro θ , definimos su **error cuadrático medio** como

$$(5.2) \quad \text{ECM}_\theta(T) = \mathbf{E}_\theta((T - \theta)^2).$$

Si T es insesgado, entonces $\text{ECM}_\theta(T) = \mathbf{V}_\theta(T)$, pues en ese caso $\theta = \mathbf{E}_\theta(T)$ (véase también el lema 5.1).

Para estimadores generales, no necesariamente insesgados, ésta es la cantidad natural para medir cuán bueno es un estimador: es el error que, en media, se comete al estimar que θ es el valor de T . Si $\text{ECM}_\theta(T)$ es pequeño significa que es muy probable que las realizaciones de T estén próximas a θ .

Si T_1 y T_2 son estimadores del parámetro θ se dice que T_1 es **más eficiente** que T_2 si, para todo $\theta \in \Theta$,

$$\text{ECM}_\theta(T_1) < \text{ECM}_\theta(T_2).$$

En el caso de estimadores insesgados, esta definición de “más eficiente” coincide con la del apartado anterior.

Llamamos la atención, ¡de nuevo!, sobre la frase “para todo $\theta \in \Theta$ ”. Si para algunos valores de θ se tuviera que $\text{ECM}_\theta(T_1) < \text{ECM}_\theta(T_2)$, mientras que para otros ocurriera que $\text{ECM}_\theta(T_1) > \text{ECM}_\theta(T_2)$, entonces los estimadores no serían comparables desde este punto de vista de la eficiencia. Véase el ejemplo 5.1.12.

El ECM de un estimador se puede escribir como la suma de la varianza y el sesgo (al cuadrado) del estimador:

Lema 5.1

$$\text{ECM}_\theta(T) = \mathbf{V}_\theta(T) + (\mathbf{E}_\theta(T) - \theta)^2.$$

DEMOSTRACIÓN. Llamemos, por simplificar la notación, $\mu = \mathbf{E}_\theta(T)$. Entonces,

$$\begin{aligned} \text{ECM}_\theta(T) &= \mathbf{E}_\theta((T - \theta)^2) = \mathbf{E}_\theta([(T - \mu) + (\mu - \theta)]^2) \\ &= \mathbf{E}_\theta((T - \mu)^2) + 2(\mu - \theta)\mathbf{E}_\theta(T - \mu) + (\mu - \theta)^2 = \mathbf{E}_\theta((T - \mu)^2) + (\mu - \theta)^2, \end{aligned}$$

donde hemos usado que $\mathbf{E}_\theta(T) = \mu$ para cancelar el segundo sumando de la penúltima expresión. ■

En general, son preferibles los estimadores insesgados. Sin embargo, el lema 5.1 nos dice, en particular, que en algún caso pudiera ser más adecuado un estimador sesgado con poco ECM que un estimador insesgado con mucha varianza, pues lo relevante en cuanto a error de estimación es la suma del sesgo (al cuadrado) más la varianza.

Obsérvese que si T es un estimador de un parámetro θ , entonces la desigualdad de Markov (teorema 2.2) da que

$$\mathbf{P}_\theta(|T - \theta| \geq \varepsilon) \leq \frac{\text{ECM}_\theta(T)}{\varepsilon^2}.$$

Es decir, que si $\text{ECM}_\theta(T)$ es pequeño entonces es (relativamente) poco probable que las estimaciones de θ obtenidas con T yerren demasiado.

EJEMPLO 5.1.10. *La cuasivarianza y la varianza muestrales como estimadores de la varianza de variables normales.*

Sea X una variable normal con media μ y varianza σ^2 . Como sabemos, S^2 es un estimador insesgado de σ^2 ; esto es un hecho general. La varianza de S^2 tiene, para una normal de parámetros μ y σ^2 , la siguiente expresión:

$$\mathbf{V}_{\mu, \sigma^2}(S^2) = \frac{2\sigma^4}{n-1}.$$

Para comprobarlo, podemos usar la expresión general de la varianza de S^2 del corolario 4.5, junto con el que $\mathbf{E}_{\mu, \sigma^2}((X - \mu)^4) = 3\sigma^4$ (nota 2.3.4). O más directamente, recordar, del teorema 4.6 de Fisher–Cochran, que $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$, y que la varianza de una χ_{n-1}^2 vale $2(n-1)$.

Así que el ECM de S^2 como estimador de σ^2 es también

$$(\star) \quad \text{ECM}_{\mu, \sigma^2}(S^2) = \frac{2\sigma^4}{n-1}.$$

Consideremos además el estadístico

$$D^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2.$$

Este estadístico D^2 (la varianza muestral) es estimador sesgado de σ^2 . En concreto, como $D^2 = \frac{n-1}{n} S^2$ tenemos que

$$\mathbf{E}_{\mu, \sigma^2}(D^2) = \frac{n-1}{n} \mathbf{E}_{\mu, \sigma^2}(S^2) = \frac{n-1}{n} \sigma^2 = \sigma^2 - \frac{1}{n} \sigma^2,$$

y por tanto que el sesgo de D^2 como estimador de σ^2 es $-\sigma^2/n$. Por otro lado,

$$\mathbf{V}_{\mu, \sigma^2}(D^2) = \frac{(n-1)^2}{n^2} \mathbf{V}_{\mu, \sigma^2}(S^2) = \frac{(n-1)^2}{n^2} \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2},$$

de manera que el ECM de D^2 como estimador de σ^2 es

$$(\star\star) \quad \text{ECM}_{\mu, \sigma^2}(D^2) = \frac{\sigma^4}{n^2} + \frac{2(n-1)\sigma^4}{n^2} = \left(\frac{2n-1}{n^2}\right) \sigma^4.$$

Comparando (\star) y $(\star\star)$, y como para todo $n > 1$ se tiene que

$$\left(\frac{2n-1}{n^2}\right) < \frac{2}{n-1},$$

resulta que D^2 , aunque sesgado, es un estimador de σ^2 más eficiente que S^2 . ♣

EJEMPLO 5.1.11. $\text{POISS}(\lambda)$. *Retomamos el ejemplo 5.1.6.*

Interesa estimar el parámetro $\mu = e^{-\lambda}$. En el citado ejemplo proponíamos los estimadores

$$T_2 = e^{-\bar{X}} \quad \text{y} \quad T_3 = \frac{1}{n} \# \{1 \leq i \leq n : X_i = 0\}.$$

Como $nT_3 \sim \text{BIN}(n, \mu)$, el estimador T_3 resulta ser insesgado, y

$$\text{ECM}_\mu(T_3) = \mathbf{V}_\mu(T_3) = \frac{1}{n^2} \mathbf{V}_\mu(\text{BIN}(n, \mu)) = \frac{1}{n^2} n \mu (1 - \mu) = \mu(1 - \mu) \frac{1}{n},$$

o en otros términos,

$$(5.3) \quad n \cdot \text{ECM}_\mu(T_3) = \mu(1 - \mu).$$

El estimador T_2 , por su parte, es sesgado al alza; aunque es asintóticamente insesgado. (En lo que sigue, como en el ejemplo 5.1.6, calcularemos con el parámetro λ , para traducir al final a μ). De hecho,

$$(5.4) \quad \text{sesgo}_\lambda(T_2) = \mathbf{E}_\lambda(T_2) - e^{-\lambda} = e^{-\lambda n(1-e^{-1/n})} - e^{-\lambda}.$$

Vamos con el cálculo de la varianza. Procediendo de manera análoga a la del ejemplo 5.1.6, se obtiene que

$$\mathbf{E}_\lambda(T_2^2) = e^{-\lambda n(1-e^{-2/n})},$$

y, por tanto,

$$(5.5) \quad \mathbf{V}_\lambda(T_2) = \mathbf{E}_\lambda(T_2^2) - \mathbf{E}_\lambda(T_2)^2 = e^{-\lambda n(1-e^{-2/n})} - e^{-2\lambda n(1-e^{-1/n})}.$$

Las expresiones (5.4) y (5.5) son un tanto complicadas, y conviene analizarlas asintóticamente, cuando $n \rightarrow \infty$. Usaremos que, para $u > 0$,

$$(5.6) \quad 1 - e^{-u} = u - \frac{u^2}{2} + O(u^3), \quad \text{cuando } u \rightarrow 0.$$

Llamemos $x_n = n(1 - e^{-1/n})$, que es una sucesión que tiende a 1 cuando $n \rightarrow \infty$ (recuérdese el ejemplo 5.1.6 o úsese (5.6)). La expresión (5.4) se puede reescribir como

$$\text{sesgo}_\lambda(T_2) = e^{-\lambda x_n} - e^{-\lambda},$$

lo que nos da, usando la definición de derivada, que

$$\frac{\text{sesgo}_\lambda(T_2)}{|x_n - 1|} \xrightarrow{n \rightarrow \infty} \lambda e^{-\lambda}.$$

Y como, por (5.6),

$$x_n = n(1 - e^{-1/n}) = 1 - \frac{1}{2n} + O\left(\frac{1}{n^2}\right) \quad \text{y por tanto} \quad 2n(1 - x_n) \xrightarrow{n \rightarrow \infty} 1,$$

deducimos que

$$2n |\text{sesgo}_\lambda(T_2)| \xrightarrow{n \rightarrow \infty} \lambda e^{-\lambda}, \quad \text{o bien que} \quad 4n^2 |\text{sesgo}_\lambda(T_2)|^2 \xrightarrow{n \rightarrow \infty} \lambda^2 e^{-2\lambda}.$$

Para la varianza, argumentamos como sigue. Introducimos la sucesión $y_n = \frac{n}{2}(1 - e^{-2/n})$, que tiende a 1 cuando $n \rightarrow \infty$, y para la que, de hecho, por (5.6),

$$y_n = 1 - \frac{1}{n} + O\left(\frac{1}{n^2}\right).$$

La expresión (5.5) se puede reescribir como

$$\mathbf{V}_\lambda(T_2) = e^{-2\lambda y_n} - e^{-2\lambda x_n},$$

donde x_n es la sucesión considerada anteriormente. Usando de nuevo la noción de derivada, tenemos que

$$\frac{\mathbf{V}_\lambda(T_2)}{|x_n - y_n|} \xrightarrow{n \rightarrow \infty} 2\lambda e^{-2\lambda},$$

y como $|x_n - y_n| = 1/(2n) + O(1/n^2)$, concluimos que

$$n \cdot \mathbf{V}_\lambda(T_2) \xrightarrow{n \rightarrow \infty} \lambda e^{-2\lambda}.$$

Finalmente,

$$n \cdot \text{ECM}_\lambda(T_2) = n \cdot \mathbf{V}_\lambda(T_2) + n \cdot \text{sesgo}_\lambda^2(T_2) \xrightarrow{n \rightarrow \infty} \lambda e^{-2\lambda}.$$

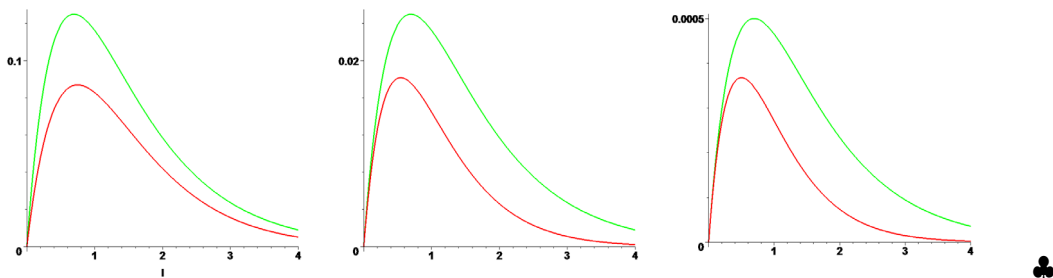
Listo. Ya podemos comparar. Tenemos (escribiendo en términos de λ) que

$$n \cdot \text{ECM}_\lambda(T_3) = e^{-\lambda}(1 - e^{-\lambda}) \quad \text{y} \quad n \cdot \text{ECM}_\lambda(T_2) \xrightarrow{n \rightarrow \infty} \lambda e^{-2\lambda}.$$

Como

$$\frac{\lambda e^{-2\lambda}}{e^{-\lambda}(1 - e^{-\lambda})} = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} = \frac{\lambda}{e^\lambda - 1} < 1,$$

pues $e^x > 1 + x$ para todo $x \in \mathbb{R}$, concluimos que, en cuanto n sea moderadamente grande, el estimador T_2 tiene menor ECM que T_3 , a pesar de que este último estimador sea insesgado. En realidad, la conclusión es cierta para todo $n \geq 2$, como se muestra en las siguientes gráficas, en las que se han dibujado los respectivos ECM (las expresiones exactas) para $n = 2$, $n = 10$ y $n = 500$:



Nota 5.1.6. Verificar la comparación que se destila de las gráficas anteriores, para todo $n \geq 2$, exige análisis más detallados que van más allá del uso de las cómodas estimaciones asintóticas como (5.6), y no lo haremos aquí.

EJEMPLO 5.1.12. *Estimadores del parámetro p de $X \sim \text{BER}(p)$ con muestras de tamaño 2.*

Recuérdese que $\mathbf{E}_p(X) = p$ y $\mathbf{V}_p(X) = p(1 - p)$. Consideramos los siguientes estimadores del parámetro p :

$$T_1 = \frac{X_1 + X_2}{2} \quad \text{y} \quad T_2 = \min(X_1, X_2).$$

De T_1 ya sabemos que es insesgado y que su varianza es $\mathbf{V}_p(T_1) = p(1 - p)/2$. Por lo tanto,

$$\text{ECM}_p(T_1) = \frac{1}{2} p(1 - p).$$

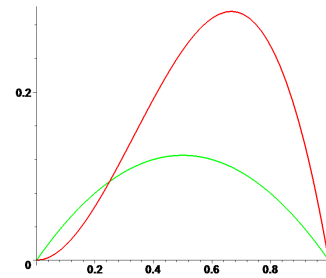
Por su parte, el estimador T_2 es también una variable Bernoulli, de parámetro p^2 (pues solo vale 1 cuando $X_1 = X_2 = 1$), así que

$$\mathbf{E}_p(T_2) = p^2 = p \underbrace{- p(1 - p)}_{=\text{sesgo}} \quad \text{y} \quad \mathbf{V}_p(T_2) = p^2(1 - p^2).$$

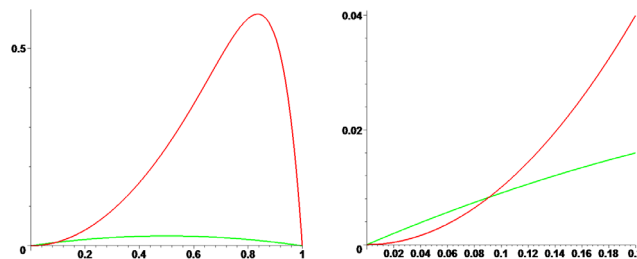
Esto nos da que

$$\text{ECM}_p(T_2) = p^2(1 - p^2) + p^2(1 - p)^2 = 2p^2(1 - p).$$

Las gráficas de las funciones $\text{ECM}_p(T_1)$ y $\text{ECM}_p(T_2)$ para $p \in (0, 1)$ se representan a la derecha. Obsérvese cómo se cruzan. Así que no podemos concluir nada sobre cuál de los dos estimadores es más eficiente, a menos que dividamos la discusión en dos regiones de Θ .



Para muestras de tamaño n genérico, tomando como estimador T_1 la media aritmética, y como T_2 el mínimo, los respectivos ECM vienen dados por $\text{ECM}(T_1) = p(1 - p)/n$ y $\text{ECM}(T_2) = p^n(1 - p^n) + p^2(1 - p^{n-1})^2$. Estas dos funciones se cruzan en el espacio de parámetros. Pero en cuanto n es moderadamente grande, el estimador T_1 , la media aritmética, resulta ser mucho mejor que T_2 para *casi todos* los valores de p , excepto los muy pequeños. Véanse en las figuras el caso $n = 10$:



La de la izquierda muestra cómo el ECM del mínimo es considerablemente más alto que el de la media aritmética en casi todos los valores de p , pero no en todos, como se ilustra con el zoom para valores pequeños de p de la derecha. ♣

5.2. Métodos de construcción de estimadores

Vamos ahora a presentar los dos métodos generales, de amplio espectro, de construcción de estimadores: método de momentos y método de máxima verosimilitud.

5.2.1. Método de momentos

La idea de este método es natural, y se basa en que, dada una muestra (x_1, \dots, x_n) de la variable X que sigue una distribución dada por $f(x; \theta)$,

- la media muestral \bar{x} “debe parecerse” a la media poblacional $\mathbf{E}_\theta(X)$ (la esperanza de X si se diera el parámetro θ);
- la media poblacional $\mathbf{E}_\theta(X)$ es de hecho una función de θ .

“Por consiguiente”, asimilando el “debe parecerse” con “coincide”,

- planteamos la ecuación (con incógnita θ)

$$(\dagger) \quad \mathbf{E}_\theta(X) = \bar{x},$$

- despejamos θ de la ecuación (\dagger) ,
- ésta es la **estimación** por momentos de θ , a la que nombramos como $\hat{\theta}$.

Veamos un ejemplo numérico sencillo. Digamos que tenemos la siguiente muestra de tamaño 8 de una $X \sim \text{BER}(p)$:

$$(0, 1, 1, 0, 1, 1, 1, 0).$$

La media muestral es $\bar{x} = 5/8$. Por otro lado, $\mathbf{E}_p(X) = p$. De manera que planteamos la ecuación

$$p = 5/8,$$

cuya solución es, obviamente, $\hat{p} = 5/8$. Ésta sería la estimación de p por momentos para la muestra dada.

La estimación así obtenida depende, claro, de la muestra, y en este caso particular del valor de \bar{x} . Distintas muestras, digamos de tamaño 8, producirían distintas estimaciones (en función de cuántos unos contengan), pero con una “fórmula” común para todas ellas (la media aritmética, en este caso).

En lo que sigue, y para análisis posteriores de cuán buenos resultan ser los estimadores obtenidos por este método, consideraremos la contrapartida teórica habitual, y elevaremos la notación, considerando variables aleatorias: tendremos muestras (X_1, \dots, X_n) de la variable X , y nos referiremos a la solución de la ecuación

$$(5.7) \quad \mathbf{E}_\theta(X) = \bar{X}$$

como el **estimador** por momentos de θ , en símbolo: $M_\theta(X_1, \dots, X_n)$.

Este método requiere, primero, disponer de una expresión explícita de $\mathbf{E}_\theta(X)$ en términos de θ , y luego resolver la ecuación (5.7) para obtener la estimación; esto último, por cierto, no es siempre inmediato (véase el ejemplo 5.2.5).

En algún caso, por ejemplo cuando resulta que $\mathbf{E}_\theta(X) = 0$ para todo θ , la ecuación (5.7) no tiene contenido, y apelaremos a otros momentos, como $\mathbf{E}_\theta(X^2)$ o a $\mathbf{V}_\theta(X)$ (o, incluso, si hiciera falta, a $\mathbf{E}_\theta(X^k)$ con $k \geq 3$). Ésta es la razón del nombre del método. De hecho, para cada momento de la variable se obtiene, en general, un estimador distinto. Si por ejemplo optáramos por utilizar el segundo momento, la ecuación correspondiente sería

$$\mathbf{E}_\theta(X^2) = \overline{x^2},$$

mientras que usando la varianza tendríamos que plantear

$$\mathbf{V}_\theta(X) = \mathbf{E}_\theta(X^2) - \mathbf{E}_\theta(X)^2 = \overline{x^2} - \overline{x}^2.$$

Cuando hay varios parámetros, como en las variables normales o en las variables gamma, se hace imprescindible recurrir a varios momentos.

Veamos ahora unos cuantos ejemplos, en los que ya empleamos la notación de variables aleatorias, que ilustran todas estas posibilidades. Empezamos con:

EJEMPLO 5.2.1. *Estimación por momentos en algunos de los modelos habituales.*

a) $X \sim \text{BER}(p)$. Dada una muestra (x_1, \dots, x_n) , como $\mathbf{E}_p(X) = p$, la ecuación es

$$p = \overline{x} \implies \hat{p} = \overline{x},$$

y el estimador es $M_p(X_1, \dots, X_n) = \overline{X}$. El *estimador* M_p es el estadístico que registra la proporción de unos en la muestra (X_1, \dots, X_n) ; la *estimación* \hat{p} , dada la lista (x_1, \dots, x_n) de ceros y unos, viene dada por la proporción de unos en esa lista.

b) $X \sim \text{EXP}(\lambda)$. Dada una muestra (x_1, \dots, x_n) , como $\mathbf{E}_\lambda(X) = 1/\lambda$, tenemos

$$\frac{1}{\lambda} = \overline{x} \implies \hat{\lambda} = \frac{1}{\overline{x}},$$

y por tanto el estimador es $M_\lambda(X_1, \dots, X_n) = 1/\overline{X}$.

c) $X \sim \text{POISS}(\lambda)$. Dada una muestra (x_1, \dots, x_n) , como $\mathbf{E}_\lambda(X) = \lambda$, tenemos

$$\lambda = \overline{x} \implies \hat{\lambda} = \overline{x},$$

y el estimador es $M_\lambda(X_1, \dots, X_n) = \overline{X}$.

d) $X \sim \text{UNIF}[0, a]$. Dada una muestra (x_1, \dots, x_n) , como $\mathbf{E}_a(X) = a/2$,

$$\frac{a}{2} = \overline{x} \implies \hat{a} = 2\overline{x};$$

el estimador por momentos de a es $M_a(X_1, \dots, X_n) = 2\overline{X}$.

e) $X \sim \mathcal{N}(\mu, \sigma^2)$. En este modelo tenemos un par de parámetros, μ y σ^2 . Tenemos que $\mathbf{E}_{\mu, \sigma^2}(X) = \mu$ y que $\mathbf{V}_{\mu, \sigma^2}(X) = \sigma^2$.


Dada una muestra (x_1, \dots, x_n) , planteamos el sistema de ecuaciones

$$\begin{cases} \mu = \bar{x}, \\ \sigma^2 = \overline{x^2} - \bar{x}^2, \end{cases}$$

del que se obtienen inmediatamente las estimaciones $\hat{\mu}$ y $\widehat{\sigma^2}$, y los correspondientes estimadores por momentos de los parámetros μ y σ^2 :

$$M_\mu(X_1, \dots, X_n) = \bar{X} \quad \text{y} \quad M_{\sigma^2}(X_1, \dots, X_n) = D^2.$$

(Recuérdese que $D^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 = \overline{X^2} - \bar{X}^2$). Obsérvese que M_{σ^2} no es S^2 .

 **Nota 5.2.1.** Si usamos los dos primeros momentos, y no la media y la varianza, se obtienen los mismos estimadores:

$$\begin{cases} \mathbf{E}_{\mu, \sigma^2}(X) = \mu = \bar{x} \\ \mathbf{E}_{\mu, \sigma^2}(X^2) = \sigma^2 + \mu^2 = \overline{x^2}, \end{cases} \quad \longrightarrow \quad M_\mu(X_1, \dots, X_n) = \bar{X} \text{ y } M_{\sigma^2}(X_1, \dots, X_n) = D^2.$$

f) $X \sim \text{GAMMA}(\lambda, t)$. El sistema de ecuaciones es, recordando los momentos de las variables Gamma, véase (2.23),

$$\begin{cases} \mathbf{E}_{\lambda, t}(X) = \frac{t}{\lambda} = \bar{x}, \\ \mathbf{V}_{\lambda, t}(X) = \frac{t}{\lambda^2} = \overline{x^2} - \bar{x}^2. \end{cases} \quad \Longrightarrow \quad \begin{cases} \hat{\lambda} = \frac{\bar{x}}{\overline{x^2} - \bar{x}^2}, \\ \hat{t} = \frac{\bar{x}^2}{\overline{x^2} - \bar{x}^2}. \end{cases}$$

Los correspondientes estimadores son

$$M_\lambda(X_1, \dots, X_n) = \frac{\bar{X}}{D^2} \quad \text{y} \quad M_t(X_1, \dots, X_n) = \frac{\bar{X}^2}{D^2}. \quad \clubsuit$$

Aunque sólo haya un parámetro, a veces el primer momento no da información.

EJEMPLO 5.2.2. Consideremos la función de densidad (de una variable aleatoria triangular, simétrica con moda y media en 0) dada por

$$f(x; \theta) = \begin{cases} \frac{1}{\theta}(1 - |x|/\theta), & \text{si } |x| < \theta, \\ 0, & \text{en otro caso,} \end{cases}$$

donde $\theta \in \Theta = (0, +\infty)$.

En este caso $\mathbf{E}_\theta(X) = 0$ para todo $\theta \in (0, +\infty)$. De hecho, por simetría, todos los momentos impares son cero. Los momentos pares son

$$\mathbf{E}_\theta(X^{2k}) = 2 \int_0^\theta \frac{1}{\theta} \left(1 - \frac{x}{\theta}\right) x^{2k} dx = \frac{1}{(2k+1)(k+1)} \theta^{2k},$$

y en particular, $\mathbf{E}_\theta(X^2) = \theta^2/6$. Por consiguiente, para una muestra (x_1, \dots, x_n) tendríamos, usando el segundo momento, la siguiente estimación:

$$\frac{\theta^2}{6} = \overline{x^2} \implies \hat{\theta} = \sqrt{6\overline{x^2}},$$

mientras que

$$M_\theta(X_1, \dots, X_n) = \sqrt{6\overline{X^2}}$$

sería el estimador usando el segundo momento. ♣

En principio se pueden usar cualesquiera momentos, y se obtendrán estimadores alternativos.

EJEMPLO 5.2.3. $X \sim \text{POISS}(\lambda)$.

Si se usa el primer momento, la media $\mathbf{E}_\lambda(X) = \lambda$, obtenemos el estimador $M_\lambda(X_1, \dots, X_n) = \overline{X}$. Usando que $\mathbf{V}_\lambda(X) = \lambda$, tendríamos el estimador alternativo $M_\lambda^*(X_1, \dots, X_n) = D^2$. ♣

EJEMPLO 5.2.4. $X \sim \text{RAY}(\sigma^2)$.

Recuerde, lector, el apartado 3.3.4, y en particular la expresión (3.16). Los dos primeros momentos de una $\text{RAY}(\sigma^2)$ son

$$\mathbf{E}_{\sigma^2}(X) = \sqrt{\pi/2} \sigma, \quad \mathbf{E}_{\sigma^2}(X^2) = 2\sigma^2.$$

Esto nos da dos estimaciones del parámetro σ^2 para una muestra (x_1, \dots, x_n) dada:

$$\sqrt{\frac{\pi}{2}} \sigma = \overline{x} \implies \hat{\sigma}^2 = \frac{2}{\pi} \overline{x^2}, \quad \text{y} \quad 2\sigma^2 = \overline{x^2} \implies \hat{\sigma}^2 = \frac{\overline{x^2}}{2},$$

y los dos estimadores alternativos para el parámetro σ^2 siguientes:

$$\widehat{M}_{\sigma^2}(X_1, \dots, X_n) = \frac{2}{\pi} \overline{X^2} \quad \text{y} \quad M_{\sigma^2}(X_1, \dots, X_n) = \frac{1}{2} \overline{X^2}.$$

Como veremos más adelante (ejemplo 5.4.10), este segundo estimador M_{σ^2} (que proviene del segundo y no del primer momento) tiene mejores propiedades. ♣

Como se ilustra en el siguiente ejemplo, a veces no es fácil despejar θ en la ecuación $\mathbf{E}_\theta(X) = \overline{x}$.

EJEMPLO 5.2.5. Para $\lambda > 0$, consideremos la función de densidad

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{para } x > 1, \\ 1 - e^{-\lambda}, & \text{para } 0 \leq x \leq 1. \end{cases}$$

Se trata de una exponencial para $x > 1$, con el resto de la probabilidad repartida uniformemente en $[0, 1]$. Véase la figura de la derecha. Como

$$\mathbf{E}_\lambda(X) = \frac{1}{2} + e^{-\lambda} \left(\frac{1}{2} + \frac{1}{\lambda} \right),$$

la estimación $\hat{\lambda}$ viene dada implícitamente por

$$\bar{x} = \frac{1}{2} + e^{-\lambda} \left(\frac{1}{2} + \frac{1}{\lambda} \right).$$

No “sabemos” despejar λ de la ecuación anterior, y para cada muestra habrá que resolverla numéricamente.

Observe, lector, además, que la función $\lambda \mapsto \frac{1}{2} + e^{-\lambda} \left(\frac{1}{2} + \frac{1}{\lambda} \right)$ decrece desde $+\infty$ hasta $1/2$ cuando λ varía de 0 a ∞ . Así que, ¡atención!, si la media \bar{x} de una muestra de X fuera $< 1/2$ no se tendría estimación por momentos. ♣

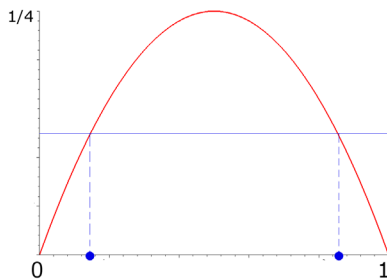
EJEMPLO 5.2.6. La variable X se escribe como la suma de dos variables X_1 y X_2 independientes, donde $X_1 \sim \text{BER}(p)$ y $X_2 \sim \text{BER}(1-p)$. El parámetro p es un número entre 0 y 1 .

Compruebe, lector, que la variable X toma valores $0, 1$ y 2 , con probabilidades $p(1-p)$, $p^2 + (1-p)^2$ y $p(1-p)$, respectivamente. Para empezar, $\mathbf{E}_p(X) = 1$, así que no podemos utilizar el primer momento para obtener estimaciones de p . Usando que X^2 toma los valores $0, 1$ y 4 con las mismas probabilidades de antes, obtenemos que

$$\mathbf{V}_p(X) = \mathbf{E}_p(X^2) - 1 = p^2 + (1-p)^2 + 4p(1-p) - 1 = 2p(1-p).$$

Dada una muestra (x_1, \dots, x_n) , la estimación por momentos (usando la varianza) se obtendría resolviendo la ecuación (de segundo grado en p)

$$p(1-p) = \frac{1}{2} (\overline{x^2} - \bar{x}^2).$$



Ahora, la función $p \mapsto p(1-p)$ en $p \in (0, 1)$ no es monótona: es simétrica en torno a $p = 1/2$, vale 0 en los extremos, y alcanza su máximo (el valor $1/4$) en $p = 1/2$. Así que, primero, si por ventura la varianza muestral fuera un número mayor que $1/2$, la ecuación anterior no tendría solución; y segundo, en el caso en el que la varianza muestral fuera menor que $1/2$, habría dos estimaciones para p por momentos (simétricas con respecto a $1/2$). En puridad, sólo tendríamos estimación única por momentos en el (muy improbable) caso de que la varianza muestral valiera justamente $1/2$. Véase la figura. ♣

5.2.2. Método de máxima verosimilitud

El de máxima verosimilitud es el método general, y más importante, de construir buenos estimadores de parámetros.

Para describirlo, empecemos con una variable X *discreta* y con función de masa $f(x; \theta)$, donde $x \in A = \text{sop}(X)$ y $\theta \in \Theta$. Tenemos la muestra aleatoria (X_1, \dots, X_n) formada por clones de X independientes entre sí.

A priori, la probabilidad, conocido θ , de obtener una realización específica (potencial) (x_1, \dots, x_n) de (X_1, \dots, X_n) viene dada por

$$(\star) \quad \prod_{j=1}^n f(x_j; \theta).$$

En la práctica (estadística) la situación es justo la contraria, disponemos de una muestra observada (x_1, \dots, x_n) , pero desconocemos θ .

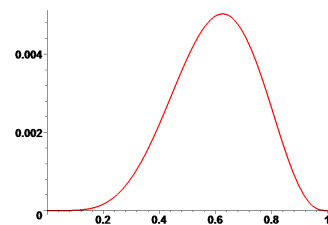
El método de máxima verosimilitud consiste, *para* (x_1, \dots, x_n) *dado*, en tomar como estimación de θ al valor de θ , digamos $\hat{\theta}$, que hace máxima la expresión (\star) . Entendemos que $\hat{\theta}$ es el valor de θ *más verosímil* dada la muestra observada (x_1, \dots, x_n) , es decir, el que asocia más probabilidad a la ocurrencia (x_1, \dots, x_n) .

*A priori: probabilidad de realizaciones;
a posteriori, verosimilitud de parámetros.*

Veamos un ejemplo numérico. Digamos que tenemos la siguiente muestra de tamaño 8 de una $X \sim \text{BER}(p)$:

$(0, 1, 1, 0, 1, 1, 1, 0).$

La expresión (\star) , en este caso, es $p^5(1-p)^3$, pues hay 5 unos y 3 ceros. Véase la gráfica de la derecha. El máximo se alcanza justamente en $p = 5/8 = 62.5\%$. Este máximo se puede calcular numéricamente o, como veremos en un momento, y como ya estará sospechando el lector (a la vista de que $5/8$ es la proporción de unos en la muestra), analíticamente.



A. Función de verosimilitud y estimación por máxima verosimilitud

Planteamos el método en general. La variable aleatoria X tiene función de masa/densidad $f(x; \theta)$, con $\theta \in \Theta$. Disponemos de una muestra aleatoria (x_1, \dots, x_n) de X . Definimos la **función de verosimilitud** (de la muestra) como

$$(5.8) \quad \text{VERO}(\theta; x_1, \dots, x_n) = \prod_{j=1}^n f(x_j; \theta).$$

Aquí, la variable de la función es θ ; lo que va detrás del punto y coma, la muestra $(x_1, \dots, x_n) \in \mathbf{sof}_\theta$, se supone fija, y desempeña el papel de parámetro. La definición (5.8) es válida tanto para variables discretas como para variables continuas³.

En el caso en el que la función $\theta \mapsto \text{VERO}(\theta; x_1, \dots, x_n)$ tenga un *máximo global (único)* en $\hat{\theta} \in \Theta$, diremos que $\hat{\theta}$ es la **estimación por máxima verosimilitud**.

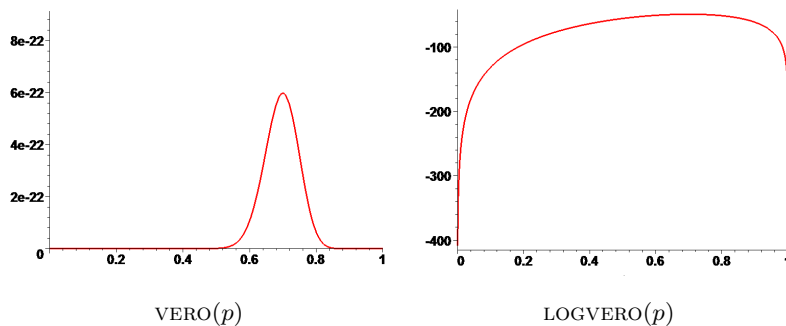
La función de verosimilitud es un producto de unos cuantos (quizás muchos) factores, muchas veces números menores o iguales que 1. Si se pretende encontrar el máximo numéricamente, tendremos dificultades (computacionales) porque los valores de la función suelen ser extraordinariamente pequeños; y si se intenta un enfoque analítico, por ejemplo derivando, será también aparatoso, pues hay que derivar productos. Por eso es habitual⁴ considerar la función de **log-verosimilitud**, que no es más que el logaritmo de la anterior:

$$(5.9) \quad \text{LOGVERO}(\theta; x_1, \dots, x_n) = \ln(\text{VERO}(\theta; x_1, \dots, x_n)) = \sum_{j=1}^n \ln f(x_j; \theta).$$

Ahora, obsérvese, la expresión involucra sumas, y no productos.

Si la función VERO alcanza un máximo global en $\hat{\theta} \in \Theta$, entonces la función LOGVERO tiene también un máximo global en ese punto.

Como ilustración de las complicaciones computacionales a las que antes aludíamos, vea el lector las dos siguientes figuras, y atienda en especial a la escala vertical.



En ellas hemos representado la función de verosimilitud y su logaritmo para una muestra de tamaño $n = 80$ procedente de una variable de Bernoulli $\text{BER}(p)$. La muestra en sí (una lista de ceros y unos) contiene 56 unos. En ambos casos, el máximo se alcanza en el mismo valor, que resulta ser $56/80$ (véase el detalle en el ejemplo 5.2.7). Pero mientras que los valores de la función de verosimilitud son del orden de 10^{-22} , los de su logaritmo son del orden de cientos (aunque negativos).

En lo que sigue, escribiremos muchas veces simplemente $\text{VERO}(\theta)$ ó $\text{LOGVERO}(\theta)$ para las funciones de verosimilitud o de log-verosimilitud, sin referencia a la muestra (x_1, \dots, x_n) , que se supone fijada.

³Salvo que, en el caso continuo, los valores de la función de densidad no se pueden interpretar como probabilidades, como sí ocurre en el caso discreto.

⁴Como LOGVERO suele tomar valores negativos, hay quien usa $-\text{LOGVERO}$ y calcula el mínimo.

B. Estimador por máxima verosimilitud

Como en el caso del método de momentos, la estimación por máxima verosimilitud depende de la muestra (x_1, \dots, x_n) . Para análisis posteriores, relacionados con la bondad de estas estimaciones, conviene elevar el enfoque y la notación considerando el modelo de muestreo aleatorio (X_1, \dots, X_n) , y denotando por $\mathbf{emv}_\theta(X_1, \dots, X_n)$ al **estimador por máxima verosimilitud** de θ .

Insistimos en la notación: en cálculos con muestras (x_1, \dots, x_n) , usaremos $\hat{\theta}$ para la estimación de θ por máxima verosimilitud dada la muestra; y designaremos como $\mathbf{emv}_\theta(X_1, \dots, X_n)$ al estadístico estimador de θ por máxima verosimilitud.

C. Cálculo de estimadores por máxima verosimilitud

La definición de la estimación del parámetro θ por máxima verosimilitud exige que la función de verosimilitud (5.8) de la muestra tenga un *único máximo global*. De manera que el primer paso de nuestro análisis, como ilustraremos en los ejemplos que siguen, ha de ser la comprobación a priori de que efectivamente hay máximo global en el espacio de parámetros Θ .

Una vez hecho esto, el cálculo en sí de la estimación se puede obtener, en muchas ocasiones, analíticamente.

Una situación habitual (como comprobará el lector en los ejemplos 5.2.7–5.2.10 que siguen) es aquella en la que el espacio de parámetros Θ es un intervalo $\Theta = (a, b)$, con $-\infty \leq a < b \leq +\infty$, y en la que comprobamos que la función $\text{VERO}(\theta)$

- es derivable en (a, b) ,
- se anula en $\theta = a$ y $\theta = b$, en el sentido de que $\lim_{\theta \downarrow a} \text{VERO}(\theta) = 0$ y $\lim_{\theta \uparrow b} \text{VERO}(\theta) = 0$,
- y tiene un único punto crítico en (a, b) .

En esta situación, que como hemos dicho es frecuente, la función continua y positiva $\text{VERO}(\theta)$ ha de tener un máximo global en (a, b) , pues se anula en sus extremos. Como $\text{VERO}(\theta)$ es derivable, ese máximo global ha de alcanzarse en un punto crítico, que hemos supuesto es único. Así que este único punto crítico es el único máximo global de $\text{VERO}(\theta)$ y, por tanto, es la estimación de máxima verosimilitud.

De manera que, en la situación descrita, el cálculo de la estimación de θ por máxima verosimilitud se realiza con la siguiente:

Receta genérica para el cálculo de \mathbf{emv}_θ

1. Se forma $\text{VERO}(\theta; x_1, \dots, x_n)$.
2. Se toma su logaritmo: $\text{LOGVERO}(\theta; x_1, \dots, x_n) = \ln(\text{VERO}(\theta; x_1, \dots, x_n))$.
3. Se deriva respecto de θ y se iguala a 0:

$$(\dagger) \quad \partial_\theta(\text{LOGVERO}(\theta; x_1, \dots, x_n)) = 0.$$

4. Se resuelve la ecuación (\dagger) , despejando $\hat{\theta}$ en términos de (x_1, \dots, x_n) .

Usaremos, en lo que sigue de capítulo, el símbolo ∂_θ , en lugar de $\partial/\partial\theta$, para denotar derivadas respecto de θ . Aunque sólo haya un parámetro, no usaremos ni d_θ ni $d/d\theta$.

Para los casos de distribuciones que dependan de dos (o más) parámetros, la receta correspondiente (para hallar puntos críticos) requeriría calcular las derivadas parciales de la función de verosimilitud (o de su logaritmo), igualarlas a cero, y resolver el sistema de ecuaciones. Véase el ejemplo 5.2.11.

Pero, ¡atención!, lector, no siempre nos encontramos en la (exigente) situación anterior. En ocasiones, porque la función de verosimilitud no es derivable en todo punto (véase el ejemplo 5.2.13); en otros casos (ejemplo 5.2.16), porque la función de verosimilitud tiene varios máximos locales. O incluso podría darse el caso de que no tuviera sentido derivar la función de verosimilitud, porque, por ejemplo, el espacio de parámetros fuera finito (ejemplo 5.2.15).

D. Ejemplos

Tratamos primero ejemplos donde la función de densidad/masa depende de un parámetro, luego ejemplos con varios parámetros, y cerramos analizando ejemplos donde el estimador de máxima verosimilitud no se obtiene a través de puntos críticos, o donde no hay estimador de máxima verosimilitud pues hay varios máximos de la función de verosimilitud.

D1. Funciones de densidad/masa con un parámetro.

EJEMPLO 5.2.7. *Máxima verosimilitud para $X \sim \text{BER}(p)$.*

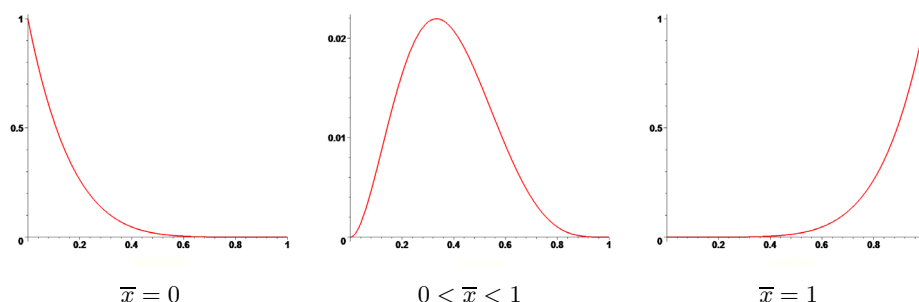
La variable $X \sim \text{BER}(p)$ toma sólo los valores 0 y 1. El parámetro $p \in [0, 1]$.

Dada la muestra observada (x_1, \dots, x_n) , la función de verosimilitud es, para cada $p \in [0, 1]$,

$$\text{VERO}(p; x_1, \dots, x_n) = p^{\#\{x_i=1\}} (1-p)^{\#\{x_i=0\}} = p^{n\bar{x}} (1-p)^{n-n\bar{x}}.$$

A la derecha hemos escrito la función de verosimilitud en términos de la media muestral, que en este caso es sólo la proporción de unos en la muestra.

La función $\text{VERO}(p)$ es derivable y no negativa. Si la muestra fuera tal que $\bar{x} = 0$ (es decir, si la muestra sólo contuviera ceros), entonces $\text{VERO}(p) = (1-p)^n$, que es decreciente. En el caso en el que $\bar{x} = 1$ (muestra con sólo unos), entonces $\text{VERO}(p) = p^n$, que es creciente. En el resto de los casos, cuando $0 < \bar{x} < 1$, la función $\text{VERO}(p)$ se anula en $p = 0$ y en $p = 1$, por lo que tiene un máximo en $(0, 1)$. Véanse las tres posibles situaciones en las siguiente figuras:



Claramente, el primer caso, $\bar{x} = 0$, da como estimación $\hat{p} = 0$, mientras que $\bar{x} = 1$ nos llevaría a $\hat{p} = 1$; lo que por otra parte es natural: si la muestra solo contiene ceros, lo más verosímil es suponer que la variable X *nunca* toma el valor uno.

En el resto de los casos, localizamos el máximo buscando puntos críticos de la función de log-verosimilitud:

$$\text{LOGVERO}(p) = n\bar{x} \ln(p) + (n - n\bar{x}) \ln(1 - p),$$

y por tanto

$$\partial_p \text{LOGVERO}(p) = n\bar{x} \frac{1}{p} - (n - n\bar{x}) \frac{1}{1 - p} = 0 \implies \frac{\bar{x}}{p} = \frac{1 - \bar{x}}{1 - p},$$

de donde se obtiene la estimación

$$\hat{p} = \bar{x}.$$

Recuérdese que, en este caso, la media muestral es la proporción de unos en la muestra. Por cierto, en los casos extremos vistos antes, la estimación para p (0 y 1, en cada caso) también se obtiene a través de la media muestral. Es decir, que si tenemos una muestra con, digamos, un 60 % de unos, entonces la estimación por máxima verosimilitud de p es, justamente, ese 60 %.

En este caso de la Bernoulli de parámetro p , el (estadístico) estimador de p resulta ser $\text{emv}_p(X_1, \dots, X_n) = \bar{X}$. ♣

EJEMPLO 5.2.8. Máxima verosimilitud para $X \sim \text{GEO}(p)$.

En este caso, la realización (x_1, \dots, x_n) consiste de enteros positivos $x_j \geq 1$, y la función de verosimilitud es

$$\text{VERO}(p) = p^n (1 - p)^{\sum_{j=1}^n x_j - n} = p^n (1 - p)^{n(\bar{x} - 1)}.$$

Si $\bar{x} = 1$, es decir, si $x_j = 1$, para $1 \leq j \leq n$, entonces $\text{VERO}(p) = p^n$ y el máximo se alcanza en $p = 1$, lo que nos daría estimación $\hat{p} = 1$, como es natural.

Si $\bar{x} > 1$, entonces la función $\text{VERO}(p)$ se anula en $p = 0$ y en $p = 1$, y tiene un máximo en $(0, 1)$. El único punto crítico, que es la estimación de máxima verosimilitud, \hat{p} , se obtiene derivando con respecto a p (e igualando a 0) la función

$$\text{LOGVERO}(p) = n \ln(p) + n(\bar{x} - 1) \ln(1 - p).$$

El resultado es

$$0 = \partial_p \text{LOGVERO}(p) = \frac{n}{p} - \frac{n(\bar{x} - 1)}{1 - p} \implies \hat{p} = \frac{1}{\bar{x}}.$$

De manera que si la muestra (de la geométrica) tiene media muestral, por ejemplo, 3, entonces la estimación por máxima verosimilitud de p es $1/3$.

Elevando notación, tenemos que $\mathbf{emv}_p(X_1, \dots, X_n) \equiv 1/\bar{X}$. ♣

EJEMPLO 5.2.9. *Máxima verosimilitud para $X \sim \text{EXP}(\lambda)$.*

La función de verosimilitud para una muestra observada (x_1, \dots, x_n) (que consiste necesariamente de números positivos) es

$$\text{VERO}(\lambda; x_1, \dots, x_n) = \lambda^n e^{-\lambda \sum_{j=1}^n x_j} = \lambda^n e^{-\lambda n \bar{x}}.$$

Como $\bar{x} > 0$, se tiene que $\text{VERO}(\lambda) \rightarrow 0$ cuando $\lambda \downarrow 0$ y cuando $\lambda \uparrow \infty$, así que tiene máximo global, que ha de ser un punto crítico.

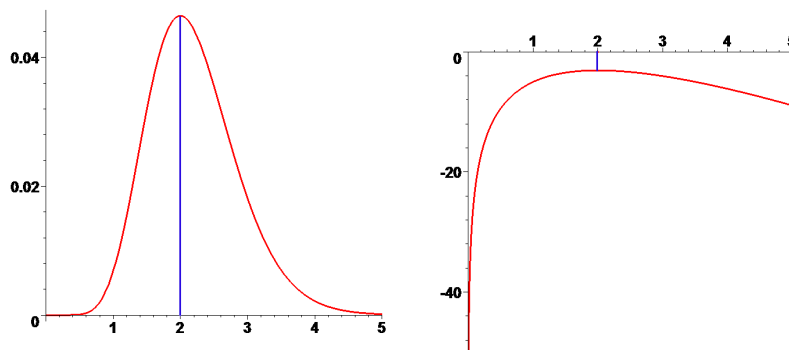
Tomando logaritmos,

$$\text{LOGVERO}(\lambda) = n \ln(\lambda) - \lambda n \bar{x}.$$

Derivando e igualando a 0, deducimos que

$$\partial_\lambda \text{LOGVERO}(\lambda) = \frac{n}{\lambda} - n \bar{x} \implies \hat{\lambda} = \frac{1}{\bar{x}}$$

es el único punto crítico, que ha de ser el máximo global de VERO y de LOGVERO , y, por tanto, la estimación de máxima verosimilitud. Dibujamos a continuación el aspecto de la función de verosimilitud y de logverosimilitud para $n = 10$ y $\bar{x} = 1/2$.



Por tanto, $\mathbf{emv}_\lambda(X_1, \dots, X_n) = 1/\bar{X}$. ♣

EJEMPLO 5.2.10. *Máxima verosimilitud para $X \sim \text{RAY}(\sigma^2)$.*

La función de densidad de $X \sim \text{RAY}(\sigma^2)$, viene dada por

$$f(x; \sigma^2) = \begin{cases} \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, & \text{si } x > 0, \\ 0, & \text{si } x \leq 0. \end{cases}$$

El parámetro de la distribución a estimar es σ^2 , así que pongamos $\theta = \sigma^2$ y escribamos la función de verosimilitud en términos de θ .

Fijada una muestra observada (x_1, \dots, x_n) (necesariamente números positivos)

$$\text{VERO}(\theta; x_1, \dots, x_n) = \frac{1}{\theta^n} \left(\prod_{j=1}^n x_j \right) e^{-\sum_{j=1}^n x_j^2 / (2\theta)} = \frac{1}{\theta^n} \left(\prod_{j=1}^n x_j \right) e^{-n\overline{x^2} / (2\theta)}.$$

Nótese que $\prod_{j=1}^n x_j > 0$ y que $\overline{x^2} > 0$. La función $\text{VERO}(\theta)$ es derivable, positiva, y tiene límite 0 cuando $\theta \downarrow 0$ y cuando $\theta \uparrow \infty$. Así que tiene máximo en $(0, +\infty)$. Tomando logaritmos,

$$\text{LOGVERO}(\theta) = \ln \left(\prod_{j=1}^n x_j \right) - n \ln(\theta) - \frac{n\overline{x^2}}{2\theta},$$

y derivando (con respecto a θ),

$$\partial_\theta \text{LOGVERO}(\theta) = \frac{-n}{\theta} + \frac{n\overline{x^2}}{2\theta^2},$$

se deduce que el único punto crítico es

$$\hat{\theta} = \frac{\overline{x^2}}{2},$$

y ha de ser donde se alcanza el máximo. Si la muestra de la Rayleigh es (x_1, \dots, x_n) , calculamos primero la media de los cuadrados, luego dividimos por 2, y ésa sería la estimación por máxima verosimilitud del parámetro $\theta = \sigma^2$.

Hemos obtenido que

$$\mathbf{env}_\theta(X_1, \dots, X_n) = \frac{1}{2n} \sum_{j=1}^n X_j^2 = \frac{1}{2} \overline{X^2},$$

que coincide con el estimador asociado al segundo momento que se obtuvo en el ejemplo 5.2.4. ♣

D2. Funciones de densidad/masa con varios parámetros. Si la función de densidad/masa depende de más de un parámetro, la función de verosimilitud es una función de varias variables. Si además es derivable, buscamos sus puntos críticos planteando un sistema de ecuaciones con las derivadas parciales iguales a 0. Lo ilustramos con un par de ejemplos.

EJEMPLO 5.2.11. *Máxima verosimilitud para $X \sim \mathcal{N}(\mu, \sigma^2)$.*

Hay dos parámetros, μ y σ^2 (recalcamos: no σ , sino σ^2), que queremos estimar: $\mu \in \mathbb{R}$ y $\sigma^2 > 0$. La región de parámetros es

$$\Theta = \{(\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty)\},$$

que es el semiplano superior de \mathbb{R}^2 .

La función de verosimilitud con una muestra (x_1, \dots, x_n) fijada es

$$\text{VERO}(\mu, \sigma^2) = \prod_{i=1}^n \frac{e^{-(x_i - \mu)^2 / (2\sigma^2)}}{\sqrt{2\pi\sigma^2}} = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right).$$



Nota 5.2.2. Antes de continuar, analizamos (así, en formato nota y en pequeño) una situación un tanto especial. Imagine, lector, que todos los datos de la muestra de tamaño n coincidieran (con todos los decimales que permita la precisión del instrumento de medición). Esto sugiere, por un lado, que en realidad la muestra no fue producida con un mecanismo aleatorio; y segundo, invita a revisar urgentemente el instrumento de medida.

En cualquier caso, veamos cómo analizar esta situación desde el punto de vista de la verosimilitud para el caso de la normal. Si todos los x_i son iguales, e iguales a \bar{x} , claro, entonces en la expresión de la verosimilitud de arriba hay que sustituir x_i por \bar{x} . Veamos. Para $\mu = \bar{x}$, la función de verosimilitud es simplemente

$$\text{VERO}(\bar{x}, \sigma^2) = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}},$$

que tiende a $+\infty$ cuando $\sigma^2 \downarrow 0$. ¡Hum! La función de verosimilitud no tiene máximo, pero sí supremo, cuando σ^2 se hace 0. Y para cualquier otro posible valor de μ , la función de verosimilitud tiende a 0 cuando $\sigma^2 \rightarrow 0$ y cuando $\sigma^2 \rightarrow \infty$. En cualquier caso, es finita. La conclusión es que la estimación máximo-verosímil ha de ser $\hat{\mu} = \bar{x}$ y $\widehat{\sigma^2} = 0$; sin aleatoriedad, como ya anticipamos.

La función $\text{VERO}(\mu, \sigma^2)$ es diferenciable y positiva en todo Θ .

Siguiendo el plan de ataque general del caso de una variable el primer paso es ver que

$$(†) \quad \text{VERO}(\mu, \sigma^2) \rightarrow 0, \quad \text{cuando } (\mu, \sigma^2) \rightarrow \partial\Theta,$$

donde por $\partial\Theta$ denotamos el “borde” de la región $\Theta \subset \mathbb{R}^2$ que, recuerde el lector, era el semiplano superior. Éste es en realidad el caso, como comprobaremos con detalle más adelante; asumámoslo, por ahora.

Si es así, entonces la función $\text{VERO}(\mu, \sigma^2)$ tiene máximo absoluto en Θ , que ha de ser un punto crítico. Para localizar los posibles puntos críticos, tomamos logaritmos,

$$\text{LOGVERO}(\mu, \sigma^2) = \ln\left(\frac{1}{(2\pi)^{n/2}}\right) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,$$

y planteamos el sistema

$$\begin{cases} 0 = \partial_\mu \text{LOGVERO}(\mu, \sigma^2) = \frac{1}{2\sigma^2} \sum_{i=1}^n 2(x_i - \mu) = \frac{n}{\sigma^2} (\bar{x} - \mu), \\ 0 = \partial_{\sigma^2} \text{LOGVERO}(\mu, \sigma^2) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2} \frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu)^2, \end{cases}$$

cuya solución (única) es

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

Así que $\mathbf{emv}_\mu(X_1, \dots, X_n) = \bar{X}$ y $\mathbf{emv}_{\sigma^2}(X_1, \dots, X_n) = D^2$, que son los mismos estimadores que se obtenían por el método de momentos. ♣



Nota 5.2.3. Para el lector interesado: verificación de la condición (h).

Observe, lector que, en el semiplano $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$,

- Para μ fijo, tanto si $\sigma^2 \rightarrow +\infty$ como si $\sigma^2 \rightarrow 0$, se tiene que $\text{VERO}(\mu, \sigma^2) \rightarrow 0$;
- Para σ^2 fijo, si $\mu \rightarrow \pm\infty$, también se tiene que $\text{VERO}(\mu, \sigma^2) \rightarrow 0$.

Estas dos observaciones pintan bien cara a la verificación de (h), pero no bastan para nuestro objetivo, pues, por ejemplo, no nos garantizan que sobre la recta $\mu = \sigma^2$ no se tenga que $\text{VERO}(\mu, \mu) \rightarrow +\infty$, cuando $\mu \rightarrow \infty$.

Para la comprobación completa, usaremos el siguiente:

Lema 5.2 Para entero $n \geq 1$, sea $H_n(y)$ la función dada en $(0, +\infty)$ por

$$H_n(y) = \frac{1}{y^{n/2}} e^{-n/y}, \quad \text{para } y > 0.$$

La función $H_n(y)$ es continua y positiva, se anula en $y = 0$ e $y = +\infty$, es decir,

$$\lim_{y \downarrow 0} H_n(y) = 0 \quad y \quad \lim_{y \uparrow 0} H_n(y) = 0,$$

y además esta acotada; de hecho, $H_n(y) \leq 1$ para $y > 0$ y $n \geq 1$.

De hecho, el máximo de $H_n(y)$ se alcanza en $y = 2$ y el valor máximo es $(1/\sqrt{2e})^n$.

Llamemos, como es habitual, \bar{x} a la media de los datos y σ_x^2 a su varianza. Obsérvese primero que

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n ((x_i - \bar{x}) + (\bar{x} - \mu))^2 = n(\sigma_x^2 + (\bar{x} - \mu)^2).$$

Esto nos permite reescribir la función de verosimilitud en la forma

$$\text{VERO}(\mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{n(\sigma_x^2 + (\bar{x} - \mu)^2)}{2\sigma^2}\right),$$

de donde, usando la notación del lema anterior, resulta que

$$\text{VERO}(\mu, \sigma^2) \leq \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left(-n \frac{\sigma_x^2}{2\sigma^2}\right) = \frac{1}{\pi^{n/2}} \frac{1}{(\sigma_x^2)^{n/2}} H_n(2\sigma^2/\sigma_x^2).$$

Esta acotación (que no depende de μ) tiende a 0 tanto cuando $\sigma^2 \downarrow 0$ como cuando $\sigma^2 \uparrow +\infty$, por el lema 5.2.

Pero también se tiene que

$$\begin{aligned} \text{VERO}(\mu, \sigma^2) &\leq \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left(-n \frac{(\bar{x} - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\pi^{n/2}} \frac{1}{((\bar{x} - \mu)^2)^{n/2}} H_n(2\sigma^2/(\bar{x} - \mu)^2) \leq \frac{1}{\pi^{n/2}} \frac{1}{((\bar{x} - \mu)^2)^{n/2}}, \end{aligned}$$

usando en el último paso que $H_n(y) \leq 1$ para todo $y > 0$, según el lema 5.2. Esta acotación (que no depende de σ^2) de la función de verosimilitud tiende a 0 cuando $|\mu| \rightarrow \infty$.

La combinación de estas dos acotaciones nos da (h). └──────────┘

EJEMPLO 5.2.12. *Máxima verosimilitud para $X \sim \text{GAMMA}(\lambda, t)$.*

La función de densidad es, en este caso,

$$f_{\lambda, t}(x) = \frac{1}{\Gamma(t)} \lambda^t x^{t-1} e^{-\lambda x} \quad \text{para } x > 0.$$

La región de parámetros es

$$\Theta = \{(\lambda, t) \in (0, +\infty) \times (0, +\infty)\},$$

que es el cuadrante positivo de \mathbb{R}^2 .

Dada una muestra (x_1, \dots, x_n) , la función de verosimilitud es

$$\text{VERO}(\lambda, t; x_1, \dots, x_n) = \prod_{i=1}^n f_{\lambda, t}(x_i) = \frac{\lambda^{nt}}{\Gamma(t)^n} e^{-\lambda n \bar{x}} \left(\prod_{i=1}^n x_i \right)^{t-1}.$$

Dejamos como ejercicio para el lector la comprobación de que esta función tiene un único máximo global en Θ , y nos ponemos con su cálculo. Tomamos logaritmos,

$$\text{LOGVERO}(\lambda, t) = nt \ln(\lambda) - n \ln(\Gamma(t)) - \lambda n \bar{x} + n(t-1) \overline{\ln(x)},$$

donde hemos abreviado con el símbolo $\overline{\ln(x)}$ al promedio $\frac{1}{n} \sum_{i=1}^n \ln(x_i)$ de los logaritmos de la muestra, y planteamos el sistema de ecuaciones

$$\begin{cases} 0 = \partial_\lambda \text{LOGVERO}(\lambda, t) = \frac{nt}{\lambda} - n \bar{x}, \\ 0 = \partial_t \text{LOGVERO}(\lambda, t) = n \ln(\lambda) - n \frac{\Gamma'(t)}{\Gamma(t)} + n \overline{\ln(x)}. \end{cases}$$


Veamos. De la primera ecuación se obtiene que $\lambda = t/\bar{x}$, que llevado a la segunda nos da la ecuación (en la variable t) siguiente:

$$\ln\left(\frac{t}{\bar{x}}\right) - \frac{\Gamma'(t)}{\Gamma(t)} + \overline{\ln(x)} = 0,$$

o equivalentemente,

$$\ln(t) - \frac{\Gamma'(t)}{\Gamma(t)} = \ln(\bar{x}) - \overline{\ln(x)}.$$

Para empezar, no sabemos despejar t en esta ecuación. Así que, de tener solución, deberá ser calculada numéricamente. Solución tiene, y única. La comprobación pasa, por un lado, por verificar que la función $t \mapsto \ln(t) - \Gamma'(t)/\Gamma(t)$ decrece desde $+\infty$ hasta 0 cuando t varía de 0 a $+\infty$ y por otro, cerciorarse de que el miembro de la derecha es un número positivo. ♣

 **Nota 5.2.4.** Sean x_1, \dots, x_n unos números positivos. Consideremos los promedios $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ y $\overline{\ln(x)} = \frac{1}{n} \sum_{i=1}^n \ln(x_i)$. El que $\ln(\bar{x}) \geq \overline{\ln(x)}$ se sigue de aplicar la desigualdad de Jensen a la función logaritmo, véase (2.4), tomando allí como variable X la que toma cada uno de los valores x_1, \dots, x_n con probabilidad $1/n$.

La otra comprobación requiere conocer ciertas propiedades de la función $\Gamma'(t)/\Gamma(t)$, la derivada logarítmica de la función Gamma, que se conoce en ambientes selectos como la *función digamma*, a veces denotada por $\Psi(t)$. No entraremos aquí en los detalles.

D3. Matizaciones sobre el cálculo del estimador de máxima verosimilitud.

Exhibimos a continuación unos cuantos ejemplos en los que este marco habitual de (único) punto crítico no se puede aplicar, y en los que el cálculo del estimador de máxima verosimilitud requiere técnicas y análisis específicos.

EJEMPLO 5.2.13. *Máxima verosimilitud para $X \sim \text{UNIF}[0, a]$, con $a > 0$.*

La función de densidad es

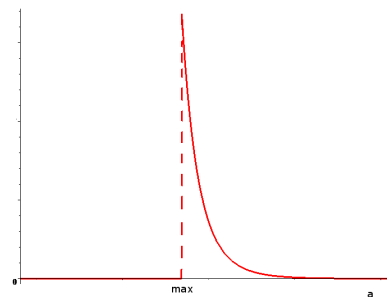
$$f(x; a) = \begin{cases} 1/a & \text{si } 0 \leq x \leq a, \\ 0 & \text{en caso contrario.} \end{cases}$$

Supongamos que tenemos una muestra observada (x_1, \dots, x_n) de la variable X . Fijamos a . Si en esa lista de números hay *alguno* que sea mayor que a , entonces es imposible (inverosímil) que la muestra provenga de una uniforme con *ese* parámetro a . En otras palabras, la verosimilitud de la muestra sería cero si el *máximo* de los x_i estuviera por encima de a .

La conclusión es que, dada una muestra observada (x_1, \dots, x_n) ,

$$\text{VERO}(a) = \begin{cases} \frac{1}{a^n}, & \text{si } \max\{x_j\} \leq a, \\ 0, & \text{si } \max\{x_j\} > a. \end{cases}$$

Aquí no derivamos, sino que observamos directamente que el máximo de la función de verosimilitud se alcanza justamente en el valor $\hat{a} = \max\{x_1, \dots, x_n\}$, como se aprecia en la figura, pues $1/a^n$ es decreciente en a .



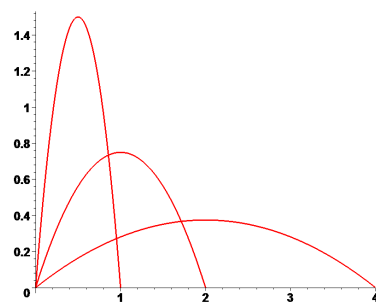
Así que el estimador resulta ser $\mathbf{emv}_a(X_1, \dots, X_n) = \max\{X_1, \dots, X_n\}$. Compárese con el estimador obtenido por momentos, que es $2\bar{X}$. Véanse también los ejercicios 5.15 y 5.16. ♣

EJEMPLO 5.2.14. *Estimador por máxima verosimilitud para el parámetro $\theta > 0$ de una variable X con función de densidad*

$$f(x; \theta) = \frac{6}{\theta^2} x \left(1 - \frac{x}{\theta}\right) \quad \text{si } x \in [0, \theta],$$

y $f(x; \theta) = 0$ en caso contrario.

Observe el lector que, como en el ejemplo anterior, el soporte de la distribución depende del parámetro. A la derecha dibujamos el aspecto de la función de densidad para los valores $\theta = 1$, $\theta = 2$ y $\theta = 4$.

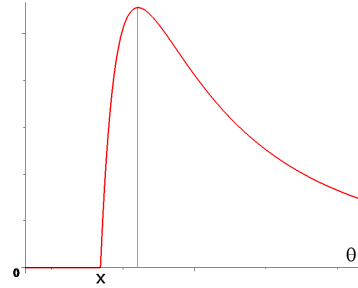


Empezamos con el (modesto) caso en el que solo disponemos de una muestra, digamos el número x_1 .

En este caso, la función de verosimilitud es, argumentando como en el ejemplo anterior,

$$\text{VERO}(\theta) = \begin{cases} \frac{6}{\theta^2} x_1 \left(1 - \frac{x_1}{\theta}\right), & \text{si } x_1 < \theta, \\ 0, & \text{si } x_1 \geq \theta, \end{cases}$$

cuyo aspecto representamos a la derecha. El máximo de la función, marcado con una línea azul en la figura, se alcanza en $\hat{\theta} = 3x_1/2$, como podrá comprobar el lector que se entretenga calculando derivadas con respecto a θ (de la función de verosimilitud o de su logaritmo) e igualando a 0.



En el caso general, para una muestra observada (x_1, \dots, x_n) de tamaño n , su función de verosimilitud resulta ser

$$\text{VERO}(\theta) = \begin{cases} 6^n \left(\prod_{i=1}^n x_i \right) \frac{1}{\theta^{2n}} \prod_{i=1}^n \left(1 - \frac{x_i}{\theta}\right), & \text{si } \max(x_1, \dots, x_n) < \theta, \\ 0, & \text{si } \max(x_1, \dots, x_n) \geq \theta. \end{cases}$$

La función $\text{VERO}(\theta)$, que es no negativa, vale 0 en $\theta = \max(x_1, \dots, x_n)$ y tiende a 0 cuando $\theta \rightarrow \infty$, de manera que tendrá (al menos) un máximo en la región de interés, $\theta > \max(x_1, \dots, x_n)$. Como veremos, hay un *único* máximo en esa región, y de hecho el aspecto de la función $\text{VERO}(\theta)$ es muy similar al de la figura del caso de una muestra, salvo que ahora la función vale 0 hasta el valor $\max(x_1, \dots, x_n)$.

Busquemos los puntos críticos de la función de verosimilitud, o mejor, los de la de log-verosimilitud. Tomamos logaritmos (allá donde la función de verosimilitud no es 0) para obtener

$$\text{LOGVERO}(\theta) = n \ln(6) + \ln \left(\prod_{i=1}^n x_i \right) - 2n \ln(\theta) + \sum_{i=1}^n \ln \left(1 - \frac{x_i}{\theta} \right),$$

y derivando alegremente,

$$\partial_{\theta} \text{LOGVERO}(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta} \sum_{i=1}^n \frac{x_i}{\theta - x_i}.$$

Finalmente, igualamos a 0 para obtener que los puntos críticos de la función de verosimilitud verifican la ecuación

$$(\star) \quad 2n = \sum_{i=1}^n \frac{x_i}{\theta - x_i}.$$

Ecuación que tiene una *única* solución en la región $\theta > \max(x_1, \dots, x_n)$. Para verlo, observamos que el miembro de la derecha es una función decreciente en θ (lo que se comprueba por simple inspección de la función, o viendo que su derivada es negativa);

que tiende a $+\infty$ cuando $\theta \rightarrow \max(x_1, \dots, x_n)$; y que tiende a 0 cuando $\theta \rightarrow \infty$. Por lo tanto, solo pasará una vez por la altura $2n$. Ese único punto crítico será el máximo global de la función de verosimilitud, y por tanto la estimación por máxima verosimilitud de θ . Llamémosle, como corresponde, $\hat{\theta}$.

Llamemos, por abreviar, $M = \max(x_1, \dots, x_n)$. Por un lado sabemos que $\hat{\theta} > M$, claro. Pero en realidad

$$(\star\star) \quad M < \hat{\theta} \leq \frac{3}{2} M.$$

Para verlo, supongamos que $\hat{\theta}$ fuera mayor que $3M/2$, y por tanto $\hat{\theta} > 3x_i/2$ para cada $i = 1, \dots, n$. Entonces se tendría que

$$\sum_{i=1}^n \frac{x_i}{\hat{\theta} - x_i} < \sum_{i=1}^n \frac{x_i}{3x_i/2 - x_i} = \sum_{i=1}^n 2 = 2n,$$

lo que contradice que $\hat{\theta}$ es solución de (\star) .

En cuanto al cálculo en sí del valor de $\hat{\theta}$, en general deberemos resolver la ecuación (\star) con algún procedimiento numérico, para el que, por cierto, la estimación a priori $(\star\star)$ de la ubicación de la solución puede venir de perlas. Observe el lector que, en el caso de una única muestra, la ecuación (\star) es simplemente $2 = x/(\theta - x)$, de solución $\hat{\theta} = 3x/2$, como ya vimos.

Por cierto, como en este caso

$$\mathbf{E}_\theta(X) = \int_0^\infty \frac{6}{\theta^2} x^2 \left(1 - \frac{x}{\theta}\right) dx = \frac{\theta}{2},$$

resulta que la estimación por momentos de θ para una muestra (x_1, \dots, x_n) es, simplemente, $\hat{\theta} = 2\bar{x}$. ¡Vaya! ♣

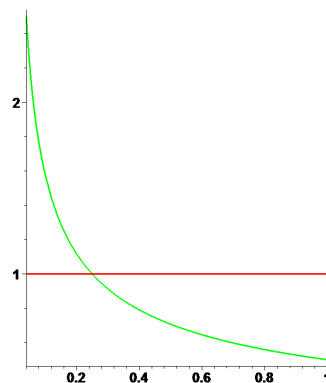
EJEMPLO 5.2.15. La variable X tiene como función de densidad $f(x; \theta)$, donde $\theta \in \Theta = \{0, 1\}$. Las dos funciones de densidad alternativas vienen dadas, para $x \in (0, 1)$, por

$$f(x; 0) = 1, \quad f(x; 1) = 1/(2\sqrt{x}).$$

Se pide determinar el estimador de máxima verosimilitud del valor de $\theta \in \{0, 1\}$.

En la figura de la derecha representamos las dos posibles funciones de densidad. Sea $(x_1, \dots, x_n) \in (0, 1)^n$ una muestra de X . Nótese que el espacio de parámetros es discreto, y por tanto no tiene sentido derivar la función de verosimilitud de la muestra, que toma únicamente dos valores:

$$V(\theta; x_1, \dots, x_n) = \begin{cases} 1, & \text{si } \theta = 0; \\ \frac{1}{2^n \sqrt{x_1 \cdots x_n}}, & \text{si } \theta = 1. \end{cases}$$



Basta comparar estas dos cantidades para decidir cuál es la estimación: cuando $2^n \sqrt{x_1 \cdots x_n}$ sea menor que 1 (de manera que su recíproco es mayor que 1), optaremos por $\hat{\theta} = 1$, y en caso contrario tomaremos $\hat{\theta} = 0$. En términos más simplificados, tendremos la estimación $\hat{\theta} = 1$ cuando $x_1 \cdots x_n < 1/4^n$, y $\hat{\theta} = 0$ en caso contrario.

Obsérvese que el que el producto $x_1 \cdots x_n$ sea menor que $1/4^n$ requiere que muchos de los x_i sean relativamente pequeños, lo que apunta a que, efectivamente, fueron sorteados con el modelo $f(x; 1)$. Por ejemplo, si disponemos de una única muestra x_1 , nos decidiremos por el modelo $f(x; 1)$ cuando $x_1 < 1/4$. ♣

En el siguiente ejemplo, tenemos un modelo en el que la función de máxima verosimilitud puede tener, según sean los datos de la muestra, varios máximos global, un único máximo global, o ninguno. ¡Vaya, vaya!

EJEMPLO 5.2.16. *Estimación del coeficiente de correlación ρ de un vector aleatorio (X, Y) que sigue una normal bidimensional de parámetros $\mathbf{m} = \mathbf{0}$ y matriz de covarianzas $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.*

Ambas X e Y son normales estándar. La distribución conjunta viene determinada por $\rho \in \Theta = (-1, 1)$. Recordemos que la función de densidad conjunta es, en este caso,

$$f(x, y; \rho) = \frac{1}{2\pi} \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{1}{2} \frac{1}{1-\rho^2} (x^2 - 2xy\rho + y^2)}$$

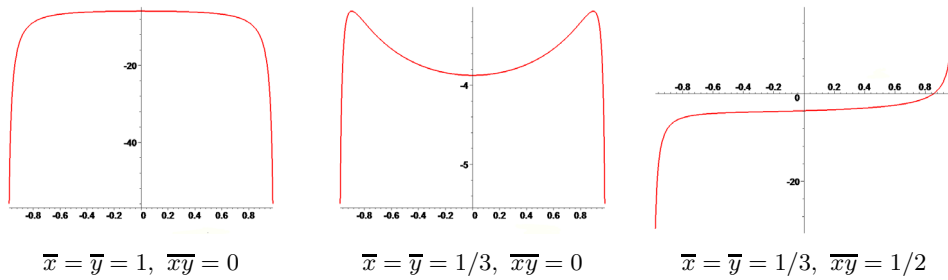
Supongamos que tenemos una muestra $((x_1, y_1), \dots, (x_n, y_n))$. La función de logverosimilitud es

$$\begin{aligned} \text{LOGVERO}(\rho; ((x_1, y_1), \dots, (x_n, y_n))) \\ &= - \sum_{j=1}^n \left[\ln \left(\frac{1}{2\pi} \right) - \frac{1}{2} \ln(1 - \rho^2) - \frac{1}{2} \frac{1}{1 - \rho^2} (x_j^2 - 2x_j y_j \rho + y_j^2) \right] \\ &= n \ln \left(\frac{1}{2\pi} \right) - \frac{n}{2} \ln(1 - \rho^2) - \frac{n}{2} \frac{\overline{x^2} + \overline{y^2} - 2\rho \overline{xy}}{1 - \rho^2}, \end{aligned}$$

esto es,

$$\frac{2}{n} \text{LOGVERO}(\rho) = 2 \ln \left(\frac{1}{2\pi} \right) - \ln(1 - \rho^2) - \frac{\overline{x^2} + \overline{y^2} - 2\rho \overline{xy}}{1 - \rho^2}.$$

Vista como función de ρ (en el intervalo $(-1, 1)$), esta logverosimilitud puede tener comportamientos diversos, en función de los valores $\overline{x^2}$, $\overline{y^2}$ y \overline{xy} provenientes de la muestra. Obsérvese que, cuando $\rho^2 \rightarrow 1$, el segundo término tiende a $-\infty$, pero el tercero puede tender a $+\infty$ o a $-\infty$ en función del signo del numerador. Este balance puede dar lugar a diversas situaciones, algunas de las cuales representamos en las siguientes figuras:



En la primera situación habría un máximo global, en la segunda dos, y en la tercera no habría tal máximo.

La ecuación de los puntos críticos

$$\partial_{\rho} \left(\frac{2}{n} \text{LOGVERO}(\rho) \right) = 0$$

deviene en

$$\rho(1 - \rho^2) + \bar{xy}(1 - \rho^2) - \rho(\bar{x}^2 + \bar{y}^2 - 2\rho\bar{xy}) = 0,$$

que es una ecuación cúbica en ρ .

A la vista de cómo son las variables X e Y , uno “espera” que $\bar{x} = \bar{y} \approx 0$, y que $\bar{x}^2 = \bar{y}^2 \approx 1$. Si fueran exactamente esos valores, entonces tendríamos

$$\rho(1 - \rho^2) + \bar{xy}(1 - \rho^2) - 2\rho(1 - \rho\bar{xy}) = 0,$$

una de cuyas soluciones es $\rho = \bar{xy}$. Es decir, habría un punto crítico justo en la correlación muestral. Pero, ¿es único?, ¿es un máximo global?

Pongámonos en otro caso, en el que $\bar{xy} = 0$ y $\bar{x}^2 = \bar{y}^2 = u > 0$. La ecuación es entonces

$$\rho(1 - \rho^2) - 2u\rho = 0.$$

Una raíz es $\rho = 0$. Las otras dos posibles raíces serían solución de $\rho^2 = 1 - 2u$.

Si $u > 1/2$ no hay más puntos críticos, y $\rho = 0$ es un máximo de la función de verosimilitud.

Pero si $u < 1/2$, hay dos raíces más: $\rho = \pm\sqrt{1 - 2u}$. En este caso el valor en $\rho = 0$ de $(2/n)\text{LOGVERO}(\rho)$ (quitándole el término constante inicial) es $-2u$, mientras que en $\rho = \pm\sqrt{1 - 2u}$ es $\ln(1/2u) - 1$. Como para $z \in (0, 1)$ se tiene que $\ln(1/z) > 1 - z$, se deduce que, en este caso, que $\rho = \pm\sqrt{1 - 2u}$ son máximos globales, mientras que $\rho = 0$ es un mínimo local. Estaríamos en una situación como la representada en la figura intermedia de más arriba. ♣

5.3. Información y cota de Cramér–Rao

Es deseable disponer de estimadores con el menor error cuadrático medio posible, o alternativamente, de estimadores insesgados con la menor varianza posible.

Como vamos a ver en esta sección, existe una cota, conocida como la *cota de Cramér–Rao* (teorema 5.8), que depende sólo de la “forma” de (o mejor, de cuánto varía al moverse el parámetro de interés la forma de) la función de masa/densidad, y que restringe cuán pequeño puede llegar a ser ese error cuadrático medio, poniendo así un límite a la ambición en la búsqueda de estimadores insesgados de varianza pequeña.

Antes de iniciar la discusión, vamos a analizar una situación especialmente sencilla, que justificará algunas de las nociones que vamos a introducir para tratar la situación general. Supongamos que la variable X es una variable aleatoria continua con función de densidad $f(x; \theta)$, y que sólo hay (o sólo nos interesan) dos posibles valores distintos del parámetro: θ_0 y θ_1 . Suponemos, también para simplificar, que en ambos casos las funciones de densidad tienen como soporte todo \mathbb{R} .

El objetivo, como se ha explicado en discusiones anteriores, es obtener información sobre el valor del parámetro a partir de muestras de la variable, lo que en este caso supone simplemente distinguir si las muestras se produjeron con el valor θ_0 o con el valor θ_1 .

La función

$$(5.10) \quad \frac{f(x; \theta_1) - f(x; \theta_0)}{f(x; \theta_0)}$$

mide la diferencia (relativa) entre las dos funciones de densidad en x .

Tomamos como referencia el valor θ_0 , y promediamos ahora sobre *todas las posibles* muestras. Es decir, consideramos que la muestra se genera con θ_0 , y consideramos la expresión (5.10) como una variable aleatoria. Primero, comprobamos que su media vale 0:

$$(5.11) \quad \begin{aligned} \mathbf{E}_{\theta_0} \left(\frac{f(X; \theta_1) - f(X; \theta_0)}{f(X; \theta_0)} \right) &= \int_{\mathbb{R}} \frac{f(x; \theta_1) - f(x; \theta_0)}{f(x; \theta_0)} f(x; \theta_0) dx \\ &= \int_{\mathbb{R}} f(x; \theta_1) dx - \int_{\mathbb{R}} f(x; \theta_0) dx = 1 - 1 = 0, \end{aligned}$$

pues tanto $f(x; \theta_0)$ como $f(x; \theta_1)$ son funciones de densidad.

Como la media es 0, para medir la magnitud (promediada sobre muestras) de esa diferencia relativa recurrimos a la varianza, que en este caso es

$$(5.12) \quad \begin{aligned} \mathbf{V}_{\theta} \left(\frac{f(X; \theta_1) - f(X; \theta_0)}{f(X; \theta_0)} \right) &= \mathbf{E}_{\theta_0} \left[\left(\frac{f(X; \theta_1) - f(X; \theta_0)}{f(X; \theta_0)} \right)^2 \right] \\ &= \int_{\mathbb{R}} \left(\frac{f(x; \theta_1) - f(x; \theta_0)}{f(x; \theta_0)} \right)^2 f(x; \theta_0) dx = \int_{\mathbb{R}} \left(\frac{f(x; \theta_1)}{f(x; \theta_0)} - 1 \right)^2 f(x; \theta_0) dx. \end{aligned}$$

Veamos. Supongamos que esta varianza fuera pequeña, prácticamente 0. Eso querría decir que el integrando sería muy pequeño en todo su soporte (que es \mathbb{R}), y por

tanto $f(x; \theta_0)$ habría de ser muy similar a $f(x; \theta_1)$ en todo el soporte. La conclusión sería que las dos funciones de densidad serían prácticamente indistinguibles, y que la muestra en sí no nos será de gran utilidad para discernir entre θ_0 y θ_1 .

Digamos ahora que tenemos una muestra $\mathbf{x} = (x_1, \dots, x_n)$. Procedemos de manera análoga: consideramos la función

$$(5.13) \quad \frac{f(\mathbf{x}; \theta_1) - f(\mathbf{x}; \theta_0)}{f(\mathbf{x}; \theta_0)},$$

la interpretamos como una variable aleatoria sobre muestras \mathbb{X} de X , y comprobamos que su media, suponiendo que las muestras se producen con θ_0 , es 0:

$$(5.14) \quad \begin{aligned} \mathbf{E}_{\theta_0} \left(\frac{f(\mathbb{X}; \theta_1) - f(\mathbb{X}; \theta_0)}{f(\mathbb{X}; \theta_0)} \right) &= \int_{\mathbb{R}^n} \frac{f(\mathbf{x}; \theta_1) - f(\mathbf{x}; \theta_0)}{f(\mathbf{x}; \theta_0)} f(\mathbf{x}; \theta_0) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} f(\mathbf{x}; \theta_1) d\mathbf{x} - \int_{\mathbb{R}^n} f(\mathbf{x}; \theta_0) d\mathbf{x} = 1 - 1 = 0, \end{aligned}$$

pues $f(\mathbf{x}; \theta_0)$ y $f(\mathbf{x}; \theta_1)$ son funciones de densidad. Finalmente, interpretamos de nuevo que un valor pequeño de la varianza

$$(5.15) \quad \mathbf{V}_{\theta} \left(\frac{f(\mathbb{X}; \theta_1) - f(\mathbb{X}; \theta_0)}{f(\mathbb{X}; \theta_0)} \right) = \int_{\mathbb{R}^n} \left(\frac{f(\mathbf{x}; \theta_1)}{f(\mathbf{x}; \theta_0)} - 1 \right)^2 f(\mathbf{x}; \theta_0) d\mathbf{x}$$

se corresponde con la situación en la que las dos funciones de densidad son casi indistinguibles, y será extremadamente complicado discernir si las muestras fueron generadas con θ_0 o con θ_1 .

Para extender el argumento a una situación más general, podríamos considerar, en lugar de la cantidad (5.13), la siguiente variación:

$$\frac{f(\mathbf{x}; \theta_1) - f(\mathbf{x}; \theta_0)}{\theta_1 - \theta_0} \frac{1}{f(\mathbf{x}; \theta_0)},$$

que sugiere que, pasando al límite, la cantidad de interés será

$$\frac{\partial_{\theta} f(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)}, \quad \text{es decir,} \quad \partial_{\theta} \ln f(\mathbf{x}; \theta).$$

Recordamos, lector, que ∂_{θ} indica derivada con respecto a θ . Y que $f(\mathbf{x}; \theta) = \text{VERO}(\theta; \mathbf{x})$, la función de verosimilitud en \mathbf{x} .

La función $\partial_{\theta} \ln f(\mathbf{x}; \theta)$ mide cuán sensible es la función de verosimilitud a variaciones del parámetro θ : el cociente de la derivada entre el valor. En inglés, se conoce también como *score*.

Por analogía con la situación descrita antes, una varianza pequeña de la variable aleatoria $\partial_{\theta} \ln f(\mathbb{X}; \theta)$ se corresponderá con la situación en que la función de densidad (de muestras) para un cierto θ será casi indistinguible de la función de densidad (de muestras) para valores del parámetro muy cercanos a ese θ .

Es hora ya de formalizar.

Derivadas de la función de densidad/masa

Escribimos ahora una expresión explícita de $\partial_\theta \ln f(\mathbf{x}; \theta)$, usando, claro, que la función de masa/densidad conjunta $f(\mathbf{x}; \theta)$ se factoriza. Para $\mathbf{x} \in \mathbf{sop}^n$, con la regla de Leibniz se tiene que

$$(5.16) \quad \partial_\theta(f(\mathbf{x}; \theta)) = \partial_\theta \left(\prod_{j=1}^n f(x_j; \theta) \right) = \sum_{j=1}^n \partial_\theta f(x_j; \theta) \frac{f(\mathbf{x}; \theta)}{f(x_j; \theta)},$$

es decir,

$$\frac{\partial_\theta f(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} = \sum_{j=1}^n \frac{\partial_\theta f(x_j; \theta)}{f(x_j; \theta)}$$

(expresión que se podía haber obtenido igualmente tomado logaritmos en $f(\mathbf{x}; \theta) = \prod_{j=1}^n f(x_j; \theta)$ y luego derivando). Podemos reescribir la expresión anterior en la forma

$$(5.17) \quad \partial_\theta \ln f(\mathbf{x}; \theta) = \sum_{j=1}^n \partial_\theta \ln(f(x_j; \theta)).$$

Obsérvese que, para $\theta \in \Theta$, tanto $\partial_\theta \ln(f(x; \theta))$ como $\partial_\theta(\ln f(\mathbf{x}; \theta))$ están bien definidas para $x \in \mathbf{sop}$ y $\mathbf{x} \in \mathbf{sop}^n$, respectivamente.

Variables plácidas

Para poner en práctica el plan de análisis que hemos esbozado más arriba y poder comparar funciones de densidad para valores próximos de θ , usando derivadas respecto de θ , vamos a restringirnos a variables aleatorias que denominaremos *plácidas*, y que se definen como sigue:

Definición 5.3 (Variables plácidas) Decimos que una variable X es **plácida** si su función de densidad/masa $f(x; \theta)$ cumple los siguientes requisitos:

- a) el espacio de parámetros es un intervalo (abierto) $\Theta = (a, b)$, $-\infty \leq a < b \leq \infty$;
- b) el soporte \mathbf{sop} es fijo y no depende de $\theta \in \Theta$;
- c) para cada $x \in \mathbf{sop}$, la función $\theta \in \Theta \mapsto f(x; \theta)$ es C^2 , es decir, tiene segundas derivadas continuas;
- d) por último, para cada $\theta \in \Theta$,

$$\int_{\mathbf{sop}} |\partial_\theta \ln f(x; \theta)|^2 f(x; \theta) dx < +\infty, \quad \text{o} \quad \sum_{x \in \mathbf{sop}} |\partial_\theta \ln f(x; \theta)|^2 f(x; \theta) < +\infty,$$

en función de que X sea continua o discreta, respectivamente.

En lo sucesivo nombraremos como plácidas tanto a las variables como a sus correspondientes funciones de masa/densidad, siempre que cumplan las exigencias anteriores.

La hipótesis a) es natural, pues pretendemos derivar con respecto a θ . Y dada a), la hipótesis c) también es natural. Para ciertos resultados, en c) nos bastará con que $f(x; \theta)$ sea función continua de θ o que tenga primeras derivadas continuas, pero con dos derivadas continuas abarcamos todas las aplicaciones de interés.

Sobre la hipótesis b). Imagine, lector: derivamos $f(x; \theta)$ con respecto a θ . Veamos: cocientes incrementales. Fijamos θ y x , que habrá de estar en \mathbf{sop}_θ ; ahora variamos θ , pero, ¡hum!, el soporte se mueve, a su vez, y ya no estamos seguro de si $x \dots$ La hipótesis b) evita este círculo de dependencias. Más adelante discutiremos la relevancia de esta hipótesis de soporte fijo.

Digamos que X es una variable aleatoria continua. La condición integral d) dice que para cada $\theta \in \Theta$, la variable $Y = \partial_\theta \ln f(X; \theta)$, que va a desempeñar un papel central en lo que sigue, satisface

$$\mathbf{E}_\theta(Y^2) = \int_{\mathbf{sop}} |\partial_\theta \ln f(x, \theta)|^2 f(x; \theta) dx < +\infty$$

y, nos permitirá tomar esperanzas y varianzas de Y sin grave riesgo para la salud.

Las familias de funciones de masa/densidad habituales son todas plácidas; las (notables) excepciones, como la uniforme $\text{UNIF}[0, a]$, lo son por incumplir la cláusula b): el espacio de parámetros es $a \in \Theta = (0, +\infty)$, pero $\mathbf{sop}_a = [0, a]$, que depende del parámetro.



Nota 5.3.1. Comprobemos que dos modelos particulares, la Poisson y la exponencial, son plácidos.

Si $X \sim \text{POISS}(\lambda)$, entonces el espacio de parámetros es el intervalo $(0 + \infty)$, y el soporte (fijo) son los enteros no negativos. Esto nos da a) y b). La función de masa viene dada por

$$f(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \text{ entero}, k \geq 0,$$

Claramente, para cada k , $f(k; \lambda)$ es una función de λ con cuantas derivadas se precise. En cuanto a la condición d), se tiene que $\partial(\ln f(k; \lambda))/\partial \lambda = -1 + k/\lambda$, y para cualquier $\lambda > 0$, la serie

$$\sum_{x \in \mathbf{sop}} |\partial_\lambda \ln f(x; \lambda)|^2 f(x; \lambda) = \sum_{k=0}^{\infty} \left(\frac{k}{\lambda} - 1\right)^2 e^{-\lambda} \frac{\lambda^k}{k!} < +\infty;$$

compruébelo, lector, usando su método de comprobación de convergencia de series favorito.

Si $X \sim \text{EXP}(\lambda)$, entonces el espacio de parámetros es de nuevo el intervalo $(0 + \infty)$, y el soporte (fijo) es $(0, +\infty)$. La función de densidad es

$$f(x; \lambda) = \lambda e^{-\lambda x}, \quad \text{para } x \geq 0.$$

Para cada $x > 0$, la función $f(x; \lambda)$ es otra vez una función de λ con cuantas derivadas se requieran. Por último, $\partial \ln f(x; \lambda)/\partial \lambda = -x + 1/\lambda$, de manera que

$$\int_{\mathbf{sop}} |\partial_\lambda \ln f(x; \lambda)|^2 f(x; \lambda) dx = \int_0^\infty \left(\frac{1}{\lambda} - x\right)^2 \lambda e^{-\lambda x} dx,$$

que es una integral convergente para cualquier valor $\lambda > 0$. (De hecho, en estas dos ilustraciones, los valores de la serie y de la integral se pueden calcular explícitamente).

En el siguiente apartado introduciremos los conceptos de “variable de información” y “cantidad de información” de X (o de una muestra aleatoria de X); luego obtendremos uno de los resultados centrales del capítulo: la cota de Cramér–Rao.

El manejo de estos conceptos requiere prestar atención a ciertas cuestiones analíticas, fundamentalmente relacionadas con la derivación bajo el signo integral y con la derivación de series, que como el lector recuerda no se deben tratar a la ligera.

Si le parece, lector, para estudiar estos conceptos y sus propiedades, argumentaremos *primero* en la situación de variables plácidas y finitas (con **sop** común finito), en el que estas sutilezas analíticas no desempeñan papel alguno, para poder así presentar lo esencial de los argumentos.

Luego, por supuesto, estudiaremos la situación general (en realidad, una situación bastante general), centrándonos en cómo incorporar cabalmente la gestión de estos detalles de convergencia.

5.3.1. Información y cantidad de información

Con el objetivo de promediar la cantidad anterior, $\partial_\theta \ln f(\mathbf{x}; \theta)$, sobre el universo de posibles muestras \mathbf{x} , consideramos primero la variable aleatoria Y dada por

$$(5.18) \quad Y = \frac{\partial_\theta f(X; \theta)}{f(X; \theta)} = \partial_\theta \ln f(X; \theta),$$

a la que nos referiremos como la **información de la variable** X .

Para cada $\theta \in \Theta$, la variable Y está definida por la expresión anterior si $x \in \mathbf{sop}$; fuera de \mathbf{sop}_θ entendemos que $Y \equiv 0$. Obsérvese que Y es una función de X .

La varianza de la variable de información Y es conocida como **cantidad de información** o **información de Fisher** de X , y se denota por $I_X(\theta)$:

$$(5.19) \quad I_X(\theta) = \mathbf{V}_\theta(Y) = \mathbf{V}_\theta(\partial_\theta \ln(f(X; \theta))).$$

Obsérvese que $I_X(\theta)$ es una función definida para $\theta \in \Theta$.

Para variables plácidas y finitas, como vamos a comprobar seguidamente, se tiene siempre que $\mathbf{E}_\theta(Y) = 0$. Para variables plácidas pero no finitas, hacen falta hipótesis adicionales, que discutiremos más adelante en el apartado 5.3.3 y que, le anticipamos, lector, son bastante generales, si es que queremos asegurar que $\mathbf{E}_\theta(Y) = 0$.

Lema 5.4 (de Diotivede⁵, caso de soporte finito). *Si X es variable aleatoria plácida con soporte finito, entonces*

$$\mathbf{E}_\theta(Y) = 0, \quad \text{para cada } \theta \in \Theta$$

e

$$I_X(\theta) = \mathbf{V}_\theta(Y) = \mathbf{E}_\theta\left(\left(\frac{\partial_\theta f(X; \theta)}{f(X; \theta)}\right)^2\right).$$

DEMOSTRACIÓN. Partimos de que

$$(5.20) \quad 1 = \sum_{x \in \mathbf{sop}} f(x; \theta), \quad \text{para todo } \theta \in \Theta.$$

⁵Diotivede, Teodoro Diotivede.

Se trata de una suma finita, pues \mathbf{sop} es, por hipótesis, finito. Derivando (5.20) respecto de θ en el intervalo Θ obtenemos que

$$0 = \sum_{x \in \mathbf{sop}} \partial_{\theta} f(x; \theta) = \sum_{x \in \mathbf{sop}} \left(\frac{\partial_{\theta} f(x; \theta)}{f(x; \theta)} \right) f(x; \theta) = \mathbf{E}_{\theta}(Y).$$

Para la segunda parte, como $\mathbf{E}_{\theta}(Y) = 0$, se tiene que $\mathbf{V}_{\theta}(Y) = \mathbf{E}_{\theta}(Y^2)$. ■



Nota 5.3.2. Una expresión para $I_X(\theta)$, alternativa a la dada en el lema 5.4, y que en ocasiones pudiera simplificar algún cálculo, es la siguiente:

$$(5.21) \quad I_X(\theta) = -\mathbf{E}_{\theta}(\partial_{\theta\theta} \ln(f(X; \theta))),$$

donde con el símbolo $\partial_{\theta\theta}$ indicamos segunda derivada con respecto a θ . Veámoslo. Observamos primero que

$$\partial_{\theta\theta} \ln(f(X; \theta)) = \partial_{\theta} \left(\frac{\partial_{\theta} f(x; \theta)}{f(x; \theta)} \right) = \frac{\partial_{\theta\theta} f(x; \theta)}{f(x; \theta)} - \left(\frac{\partial_{\theta} f(x; \theta)}{f(x; \theta)} \right)^2.$$

Suponemos que X es plácida y finita. Usando la identidad anterior en la igualdad (\star) ,

$$\begin{aligned} I_X(\theta) &= \mathbf{E}_{\theta} \left(\left(\frac{\partial_{\theta} f(X; \theta)}{f(X; \theta)} \right)^2 \right) = \sum_{x \in \mathbf{sop}} \left(\frac{\partial_{\theta} f(x; \theta)}{f(x; \theta)} \right)^2 f(x; \theta) \\ &\stackrel{(\star)}{=} \sum_{x \in \mathbf{sop}} \partial_{\theta\theta} f(x; \theta) - \sum_{x \in \mathbf{sop}} \partial_{\theta\theta} \ln(f(X; \theta)) f(x; \theta) = \partial_{\theta\theta} \left(\sum_{x \in \mathbf{sop}} f(x; \theta) \right) - \mathbf{E}_{\theta}(\partial_{\theta\theta} \ln(f(X; \theta))). \end{aligned}$$

El intercambio de derivadas y suma es legal, pues se trata de una suma finita. El argumento concluye observando que $\sum_{x \in \mathbf{sop}} f(x; \theta) = 1$.

La expresión (5.21) es válida para variables plácidas muy generales, siempre que, además de verificarse el lema 5.4, el intercambio de derivadas y sumas infinitas/integrales del argumento que acabamos de exponer esté justificado.

A. Ejemplos de cálculo de cantidad de información

Veamos cómo se obtiene la variable de información y cómo se calcula la cantidad de información en los modelos más habituales. En todos los ejemplos que siguen se tiene que $\mathbf{E}_{\theta}(Y) = 0$, aunque, de los considerados, sólo el caso de $X \sim \text{BER}(p)$ es de una variable finita y se puede aplicar el lema 5.4 de Diotivede, caso de soporte finito. En todos ellos se tiene, por tanto, que $I_X(\theta) = \mathbf{V}_{\theta}(Y) = \mathbf{E}_{\theta}(Y^2)$.

En este apartado nos limitamos a calcular $I_X(\theta)$ en cada caso; dejamos para el siguiente apartado B la discusión sobre el significado de los resultados obtenidos.

EJEMPLO 5.3.1. *Cantidad de información para $X \sim \text{BER}(p)$.*

Aquí $\mathbf{sop} = \{0, 1\}$ y $\Theta = (0, 1)$. Tenemos que

$$\begin{aligned} f(1; p) = p &\implies \ln(f(1; p)) = \ln(p) \implies \partial_p \ln(f(1; p)) = \frac{1}{p}, \\ f(0; p) = 1 - p &\implies \ln(f(0; p)) = \ln(1 - p) \implies \partial_p \ln(f(0; p)) = -\frac{1}{1 - p} \end{aligned}$$

En otras palabras, cuando $X = 1$ se tiene $Y = 1/p$ y cuando $X = 0$ se tiene que $Y = -1/(1 - p)$. Una alternativa para compactar esta relación entre X e Y es escribir

$$Y = \frac{X - p}{p(1 - p)}.$$

Como $\mathbf{E}_p(X) = p$, se tiene que $\mathbf{E}_p(Y) = 0$. Además, como $\mathbf{V}_p(X) = p(1 - p)$,

$$I_X(p) = \mathbf{V}_p(Y) = \frac{1}{p^2(1 - p)^2} \mathbf{V}_p(X - p) = \frac{1}{p^2(1 - p)^2} \mathbf{V}_p(X) = \frac{1}{p(1 - p)}.$$



EJEMPLO 5.3.2. *Cantidad de información para una POISSON(λ), con $\lambda > 0$.*

Aquí, $\mathbf{sop} = \{0, 1, \dots\}$ y $\Theta = (0, +\infty)$. Como

$$\ln f(k; \lambda) = -\lambda + k \ln(\lambda) - \ln(k!), \quad \text{para } k \geq 0 \text{ y } \lambda > 0,$$

(donde $0! = 1$, como de costumbre), se tiene que

$$\partial_\lambda \ln f(k; \lambda) = -1 + \frac{k}{\lambda} = \frac{1}{\lambda}(k - \lambda), \quad \text{para } k \geq 0 \text{ y } \lambda > 0,$$

En otros términos, si X toma el valor k , entonces Y toma el valor $\frac{1}{\lambda}(k - \lambda)$. Podemos registrar compactamente esta relación entre X e Y mediante

$$Y = \frac{1}{\lambda}(X - \lambda).$$

El soporte de la variable Y es el conjunto $\{(k - \lambda)/\lambda; k = 0, 1, \dots\}$.

Obsérvese que $\mathbf{E}_\lambda(Y) = 0$, pues $\mathbf{E}_\lambda(X) = \lambda$, y que

$$\mathbf{V}_\lambda(Y) = \frac{1}{\lambda^2} \mathbf{V}_\lambda(X - \lambda) = \frac{1}{\lambda^2} \mathbf{V}_\lambda(X) = \frac{1}{\lambda},$$

usando que $\mathbf{V}_\lambda(X) = \lambda$. Así que

$$I_X(\lambda) = \frac{1}{\lambda}, \quad \text{para cada } \lambda \in (0, +\infty).$$



EJEMPLO 5.3.3. *Cantidad de información para $X \sim \text{EXP}(\lambda)$.*

La función de densidad, para cada $\lambda > 0$, es

$$f(x; \lambda) = \lambda e^{-\lambda x}, \quad \text{para } x > 0.$$

Es decir, $\Theta = (0, +\infty)$ y $\mathbf{sop} = (0, +\infty)$.

Como $\ln f(x; \lambda) = \ln(\lambda) - \lambda x$ y $\partial_\lambda \ln f(x; \lambda) = 1/\lambda - x$, resulta que la variable de información se puede escribir como

$$Y = \frac{1}{\lambda} - X,$$

de manera que $\mathbf{E}_\lambda(Y) = 0$ y $\mathbf{V}_\lambda(Y) = I_X(\lambda) = 1/\lambda^2$.

Pero supongamos ahora que el parámetro de interés es $\theta = 1/\lambda$, la media de la distribución. Escribimos la función de densidad de la exponencial como

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}, \quad \text{para } x > 0.$$

Aquí, $\Theta = (0, +\infty)$ y $\text{sop} = (0, +\infty)$. Esto nos da que

$$(\partial_\theta \ln f(x; \theta)) = \frac{(x - \theta)}{\theta^2}, \quad \text{y por tanto,} \quad Y = \frac{X - \theta}{\theta^2}.$$

Como $\mathbf{E}_\theta(X) = (1/\lambda) = \theta$ y que $\mathbf{V}_\theta(X) = (1/\lambda^2) = \theta^2$, se deduce que $\mathbf{E}_\theta(Y) = 0$ y

$$I_X(\theta) = \mathbf{V}_\theta(Y) = \frac{1}{\theta^4} \mathbf{V}_\theta(X - \theta) = \frac{1}{\theta^4} \mathbf{V}_\theta(X) = \frac{1}{\theta^2}.$$



EJEMPLO 5.3.4. *Cantidad de información para $X \sim \mathcal{N}(\mu_0, \sigma^2)$. Aquí, μ_0 es un dato conocido.*

Observe, lector, que estamos suponiendo de partida que se sabe que la media de esta distribución normal es μ_0 . Queremos estimar $\theta = \sigma^2 \in \Theta = (0, +\infty)$. Aquí, $\text{sop} = \mathbb{R}$. Tenemos

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}\sqrt{\theta}} e^{-\frac{1}{2}(x-\mu_0)^2/\theta} \quad \text{para todo } x \in \mathbb{R}.$$

De manera que

$$\ln(f(x; \theta)) = \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2} \ln(\theta) - \frac{1}{2} \frac{(x-\mu_0)^2}{\theta} \implies \partial_\theta \ln(f(x; \theta)) = \frac{(x-\mu_0)^2 - \theta}{2\theta^2},$$


y por tanto

$$Y = \frac{(X - \mu_0)^2 - \theta}{2\theta^2}.$$

Como $X \sim \mathcal{N}(\mu_0, \theta)$, podemos escribir $X = \mu_0 + \sqrt{\theta}Z$, donde $Z \sim \mathcal{N}(0, 1)$. Así que $\mathbf{E}_\theta((X - \mu_0)^2) = \theta$, lo que nos dice que $\mathbf{E}_\theta(Y) = 0$.

Finalmente,

$$I_X(\theta) = \mathbf{V}_\theta(Y) = \frac{1}{4\theta^4} \mathbf{V}_\theta((X - \mu_0)^2 - \theta) = \frac{1}{4\theta^4} \mathbf{V}_\theta((X - \mu_0)^2) = \frac{1}{4\theta^2} \mathbf{V}_\theta(Z^2) = \frac{1}{2\theta^2},$$

usando que $\mathbf{V}_\theta(Z^2) = \mathbf{E}_\theta(Z^4) - \mathbf{E}_\theta(Z^2)^2 = 2$, pues $\mathbf{E}(Z^2) = 1$ y $\mathbf{E}(Z^4) = 3$ si Z es normal estándar (nota 2.3.4). 

EJEMPLO 5.3.5. *Cantidad de información para $X \sim \mathcal{N}(\mu, \sigma_0^2)$, con σ_0^2 conocida.*

El parámetro de interés es $\mu \in \mathbb{R}$. Se tiene

$$f(x; \mu) = \frac{1}{\sigma_0 \sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma_0^2}, \quad \text{para todo } x \in \mathbb{R}.$$

Por tanto,

$$\partial_\mu \ln(f(x; \mu)) = \frac{x - \mu}{\sigma_0^2} \implies Y = \frac{X - \mu}{\sigma_0^2}.$$

De manera que $\mathbf{E}_\mu(Y) = 0$, puesto que $\mathbf{E}_\mu(X) = \mu$, e

$$I_X(\mu) = \frac{1}{\sigma_0^4} \mathbf{E}_\mu((X - \mu)^2) = \frac{\mathbf{V}_\mu(X)}{\sigma_0^4} = \frac{\sigma_0^2}{\sigma_0^4} = \frac{1}{\sigma_0^2},$$

que, obsérvese, es una constante (no depende de μ). ♣

EJEMPLO 5.3.6. *Cantidad de información para la distribución de Rayleigh.*

La función de densidad de $X \sim \text{RAY}(\sigma^2)$, viene dada por

$$f(x; \sigma^2) = \begin{cases} \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} & \text{si } x \geq 0, \\ 0 & \text{si } x < 0. \end{cases}$$

Como el parámetro de la distribución de Rayleigh es σ^2 , para los cálculos con derivadas respecto de parámetros que siguen conviene poner $\theta = \sigma^2$:

$$f(x; \theta) = \begin{cases} \frac{x}{\theta} e^{-x^2/2\theta} & \text{si } x \geq 0, \\ 0 & \text{si } x < 0. \end{cases}$$

Recuerde, lector, de (3.16), que

$$\mathbf{E}_\theta(X^2) = 2\theta \quad \text{y} \quad \mathbf{V}_\theta(X^2) = 4\theta^2.$$

Calculamos

$$\partial_\theta \ln(f(\theta; x)) = -\frac{1}{\theta} + \frac{x^2}{2\theta^2} = \frac{1}{2\theta^2} (x^2 - 2\theta),$$

así que

$$Y = \frac{1}{2\theta^2} (X^2 - 2\theta).$$

Obsérvese que $\mathbf{E}_\theta(Y) = 0$, y que

$$I_X(\theta) = \mathbf{V}_\theta(Y) = \frac{1}{4\theta^4} \mathbf{V}_\theta(X^2 - 2\theta) = \frac{1}{4\theta^4} \mathbf{V}_\theta(X^2) = \frac{1}{\theta^2}. \quad \clubsuit$$

Veamos a continuación un ejemplo adicional, más allá de los sospechosos habituales.

EJEMPLO 5.3.7. *Sea X una variable con función de densidad $f(x; \alpha) = \alpha x^{\alpha-1}$ para $0 < x < 1$, donde α es un parámetro positivo, $\alpha > 0$.*

Nos interesa estimar el parámetro $\theta = 1/\alpha$, así que escribimos:

$$f(x; \theta) = \frac{1}{\theta} x^{1/\theta-1}.$$

Aquí $\mathbf{sop} = (0, 1)$ y $\Theta = (0, +\infty)$.

Como

$$\partial_{\theta} \ln(f(x; \theta)) = \frac{1}{\theta^2} \left(\ln \left(\frac{1}{x} \right) - \theta \right),$$

se tiene que

$$Y = \frac{1}{\theta^2} \left(\ln \left(\frac{1}{X} \right) - \theta \right).$$

Obsérvese que, para cada entero $k \geq 0$,

$$\begin{aligned} \mathbf{E}_{\theta}((\ln(1/X))^k) &= \int_0^1 (\ln(1/x))^k \frac{1}{\theta} x^{1/\theta-1} dx \stackrel{[x=e^{-\theta y}]}{=} \theta^k \int_0^{\infty} y^k e^{-y} dy \\ &= \theta^k \Gamma(k+1) = \theta^k k!, \end{aligned}$$

apelando a la función Gamma de Euler. En particular,

$$\mathbf{E}_{\theta}(\ln(1/X)) = \theta \quad \text{y} \quad \mathbf{V}_{\theta}(\ln(1/X)) = \mathbf{E}_{\theta}(\ln(1/X)^2) - \mathbf{E}_{\theta}(\ln(1/X))^2 = 2\theta^2 - \theta^2 = \theta^2.$$

Esto nos da, por un lado, que $\mathbf{E}_{\theta}(Y) = 0$, como ya es habitual, y por otro que

$$I_X(\theta) = \mathbf{V}_{\theta}(Y) = \frac{1}{\theta^4} \mathbf{V}_{\theta}(\ln(1/X) - \theta) = \frac{1}{\theta^4} \mathbf{V}_{\theta}(\ln(1/X)) = \frac{1}{\theta^2}.$$

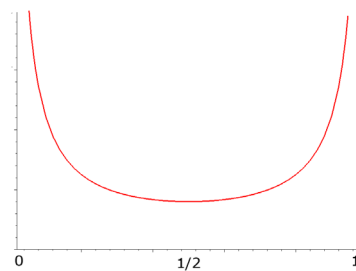


B. Significado de la cantidad de información

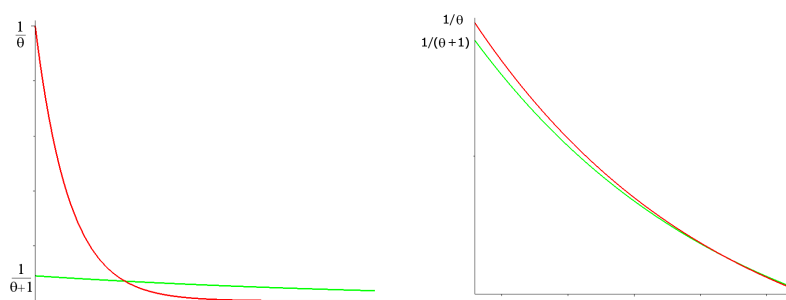
Antes de estudiar cómo interviene la cantidad de información $I_X(\theta)$ en la cota de Cramér–Rao, que es el objeto de la próxima sección 5.3.2, nos detenemos un momento en el análisis de los resultados de tres de los ejemplos anteriores, para percibir mejor el significado de la cantidad de información.

Para $X \sim \mathcal{N}(\mu, 1)$, hemos visto (ejemplo 5.3.5) que $I_X(\mu) = 1$, una constante que no depende de μ . Lo que, siguiendo el razonamiento con el que abríamos esta sección 5.3, nos dice que distinguir, a partir de muestras, digamos, un valor $\mu = 0$ de un valor $\mu = 0.1$ será tan fácil/difícil como discernir entre, por ejemplo, $\mu = 7$ de un valor $\mu = 7.1$. Lo que es bien natural si observamos que trasladar el parámetro μ en una normal no cambia la forma de la distribución.

Cuando $X \sim \text{BER}(p)$, tenemos (ejemplo 5.3.1) que $I_X(p) = 1/(p(1-p))$. Esta función tiene mínimo en $p = 1/2$, y tiende a ∞ cuando $p \rightarrow 0$ ó $p \rightarrow 1$. Así que será mucho más difícil distinguir, a partir de muestras, entre por ejemplo $p = 50\%$ y $p = 51\%$, que entre p y $p + 1\%$ si p es muy pequeño. Para ilustrarlo con un ejemplo extremo (¡muy extremo!), supongamos que pretendemos discernir entre $p = 0\%$ y $p = 1\%$ a partir de una muestra de tamaño 100. Obsérvese que, en cuanto en esa muestra aparezca un 1, nos decantaremos por $p = 1\%$. Por el contrario, observemos cuán difícil (y aventurado) sería decantarnos por $p = 50\%$ o por $p = 51\%$ si en esa muestra tuviéramos, por ejemplo, 50 ceros y 50 unos.



Veamos, por último, el caso en el que $X \sim \text{EXP}(\lambda)$, pero donde el parámetro de interés es la esperanza $\theta = 1/\lambda$. Ahora tenemos $I_X(\theta) = 1/\theta^2$ (ejemplo 5.3.3), que es pequeño cuando θ es grande. La función de densidad es $f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$. Si θ es muy grande, entonces las funciones de densidad del caso θ y, por ejemplo, $\theta + 1$, son prácticamente indistinguibles. Pero si θ es por ejemplo próximo a 0, entonces $\theta + 1$ es casi un 1, lo que es un mundo de diferencia. Véanse las figuras, que corresponden a las funciones de densidad $f(x; \theta)$ y $f(x; \theta + 1)$ para los casos $\theta = 0.1$ (izquierda) y $\theta = 10$ (derecha).



En el primer caso será relativamente sencillo distinguir los dos parámetros a partir de las muestras, mientras que en el segundo será extremadamente complicado.

C. Información y cantidad de información de una muestra aleatoria

Para una muestra aleatoria (X_1, \dots, X_n) de X , consideramos las variables de información de cada X_j , es decir,

$$Y_j = \frac{\partial_\theta f(X_j; \theta)}{f(X_j; \theta)}, \quad \text{para } 1 \leq j \leq n.$$

Estas variables de información son clones de Y y son independientes entre sí. La suma Z_n de estas Y_j ,

$$(5.22) \quad Z_n = \sum_{j=1}^n Y_j,$$

es función de la muestra (X_1, \dots, X_n) , y por tanto un estadístico.

La variable Z_n es conocida como la variable de **información (total) de la muestra**.

La varianza $\mathbf{V}_\theta(Z_n)$ de la información (total) Z_n es la **cantidad de información (total) de la muestra** (o información de Fisher de la muestra).

Lema 5.5 Para variables plácidas y finitas se tiene, para cada $\theta \in \Theta$, que

$$\mathbf{E}_\theta(Z_n) = 0 \quad \text{y} \quad \mathbf{V}_\theta(Z_n) = n I_X(\theta).$$

DEMOSTRACIÓN. Recuérdese que, para cada $j = 1, \dots, n$, $\mathbf{E}_\theta(Y_j) = 0$ y $\mathbf{V}_\theta(Y_j) = I_X(\theta)$. Esto nos da directamente que $\mathbf{E}_\theta(Z_n) = \sum_{j=1}^n \mathbf{E}_\theta(Y_j) = 0$; y por la independencia de las Y_j , que $\mathbf{V}_\theta(Z_n) = \sum_{j=1}^n \mathbf{V}_\theta(Y_j) = n I_X(\theta)$. ■

5.3.2. La cota de Cramér–Rao

La cota de Cramér–Rao, que nos ocupa ahora, es una cota inferior para la varianza de los estimadores insesgados del parámetro θ , que sólo depende de la cantidad de información de X .

De nuevo presentamos el argumento que nos conduce a esa cota primero para variables finitas (plácidas, por supuesto).

En el apartado 5.3.3 veremos condiciones bastante generales sobre la familia $f(x; \theta)$ y sobre el estimador bajo las que se tiene la cota de Cramér–Rao en el caso no finito.

Supongamos que la variable X es plácida y tiene soporte **sop** finito. Su función de masa se denota por $f(x; \theta)$. El espacio de parámetros es Θ .

Sea T un estadístico *insesgado* del parámetro θ para muestras de tamaño n . Esto es, $T = h(X_1, \dots, X_n)$ para una cierta función h , y además

$$\theta = \mathbf{E}_\theta(T), \quad \text{para todo } \theta \in \Theta.$$

Reescribimos que T es insesgado en la forma

$$(5.23) \quad \theta = \mathbf{E}_\theta(h(X_1, \dots, X_n)) = \sum_{\mathbf{x} \in \text{sop}^n} h(\mathbf{x}) f(\mathbf{x}; \theta).$$

Derivando (5.23) respecto de θ , y usando (5.16), tenemos, pues se trata de una suma finita, que

$$1 = \sum_{\mathbf{x} \in \text{sop}^n} h(\mathbf{x}) \left(\sum_{j=1}^n \frac{\partial_\theta f(x_j; \theta)}{f(x_j; \theta)} \right) f(\mathbf{x}; \theta)$$

que volvemos a escribir, inasequibles al desaliento, como esperanza en la forma

$$1 = \mathbf{E}_\theta(T \cdot Z_n),$$

utilizando la variable Z_n de información de la muestra, definida en (5.22).

Ahora, como $\mathbf{E}_\theta(Z_n) = 0$ (lema 5.5),

$$1 = \mathbf{E}_\theta(T \cdot Z_n) - \mathbf{E}_\theta(T) \mathbf{E}_\theta(Z_n) = \text{cov}_\theta(T, Z_n).$$

De la desigualdad de Cauchy–Schwarz, en su versión del corolario 2.4, que dice que

$$|\text{cov}_\theta(T, Z_n)|^2 \leq \mathbf{V}_\theta(T) \mathbf{V}_\theta(Z_n),$$

se deduce que

$$1 \leq \mathbf{V}_\theta(T) \mathbf{V}_\theta(Z_n) = \mathbf{V}_\theta(T) n I_X(\theta),$$

usando de nuevo el lema 5.5 en la última identidad.

En otros términos, hemos probado:

Teorema 5.6 (Cota de Cramér–Rao, caso de soporte finito) *Si X es una variable aleatoria plácida y con soporte finito, entonces para todo estadístico insesgado T de θ para muestras de tamaño n se cumple que*

$$(5.24) \quad \mathbf{V}_\theta(T) \geq \frac{1}{n I_X(\theta)}, \quad \text{para todo } \theta \in \Theta$$

La relevancia de este resultado estriba en que esta cota inferior para la varianza de T como estimador depende de la distribución de X directamente y de θ , y no del estimador. La cota es válida para todos los estimadores insesgados: siempre hay una cierta dispersión (varianza), al menos la que viene dada por la cota de Cramér–Rao. Parafraseando,

todo estimador insesgado de θ tiene una varianza no inferior a $1/(nI_X(\theta))$.

La cota de Cramér–Rao del teorema 5.6 recoge, cuantitativa y precisamente, la idea/intuición antes descrita de que si la información $I_X(\theta)$ es pequeña, entonces resultará difícil estimar θ a partir de muestras: si $I_X(\theta)$ es pequeña, entonces, por (5.24), cualquier estimador (insesgado) de θ tiene varianza grande, y las estimaciones que produce se dispersarán bastante en torno al verdadero valor del parámetro.

A un estimador insesgado T del parámetro θ cuya varianza es justamente la cota de Cramér–Rao, es decir, tal que

$$\mathbf{V}_\theta(X) \cdot I_X(\theta) = \frac{1}{n}, \quad \text{para todo } \theta \in \Theta,$$

se le dice **estimador eficiente** o **insesgado de mínima varianza**. Tal estimador es más eficiente que cualquier otro estimador de θ .

A. Ejemplos de cotas de Cramér–Rao y de estimadores eficientes

Sigue ahora una lista de ejemplos, las familias habituales, donde exhibimos la cota de Cramér–Rao (apoyándonos en los cálculos de las cantidades de información del apartado anterior), y donde investigamos posibles estimadores eficientes.

Sólo el primero de los ejemplos, $\text{BER}(p)$ es finito. Los demás ejemplos cumplen las condiciones que detallaremos en el apartado 5.3.3 siguiente, y por tanto para ellos se cumple también la cota de Cramér–Rao.

EJEMPLO 5.3.8. *Cota de Cramér–Rao para $X \sim \text{BER}(p)$.*

Como ya sabemos del ejemplo 5.3.1, $I_X(p) = 1/(p(1-p))$, y por tanto, la cota de Cramér–Rao es

$$\frac{p(1-p)}{n}.$$

Como \bar{X} es insesgado y $\mathbf{V}_p(\bar{X}) = \mathbf{V}_p(X)/n = p(1-p)/n$, tenemos que \bar{X} es estimador insesgado de mínima varianza. ♣

EJEMPLO 5.3.9. *Cota de Cramér–Rao para $X \sim \text{POISSON}(\lambda)$.*

Como ya sabemos del ejemplo 5.3.2, la cantidad de información es $I_X(\lambda) = 1/\lambda$. Así que la cota de Cramér–Rao es, en este caso, λ/n .

La media muestral \bar{X} es un estimador insesgado de λ . Como $\mathbf{V}_\lambda(\bar{X}) = \mathbf{V}_\lambda(X)/n = \lambda/n$, tenemos que \bar{X} es un estimador eficiente, con la mínima varianza posible. ♣

EJEMPLO 5.3.10. *Cota de Cramér–Rao para $X \sim \text{EXP}(\lambda)$.*

Si queremos estimar el parámetro λ , usamos que (ejemplo 5.3.3)

$$I_X(\lambda) = \frac{1}{\lambda^2},$$

para concluir que si T es un estimador insesgado de λ , entonces

$$\mathbf{V}_\lambda(T) \geq \frac{\lambda^2}{n}.$$

Pero si lo que queremos es estimar el parámetro $\theta = 1/\lambda$ (la media de la distribución), tendríamos, de nuevo por el ejemplo 5.3.3, que

$$I_X(\theta) = \frac{1}{\theta^2}.$$

De manera que si T es estimador insesgado de $1/\lambda$, entonces

$$\mathbf{V}_\theta(T) \geq \frac{\theta^2}{n}, \quad \text{o en términos del parámetro oficial,} \quad \mathbf{V}_\lambda(T) \geq \frac{1}{n\lambda^2}.$$

El estimador \bar{X} cumple que $\mathbf{E}_\lambda(\bar{X}) = 1/\lambda$ y que $\mathbf{V}_\lambda(\bar{X}) = \mathbf{V}_\lambda(X)/n = 1/(n\lambda^2)$. Así que \bar{X} es un estimador (de $1/\lambda$) insesgado y de mínima varianza, esto es, un estimador eficiente de $1/\lambda$. ♣

EJEMPLO 5.3.11. *Cota de Cramér–Rao para $X \sim \mathcal{N}(\mu_0, \sigma^2)$.*

Queremos estimar $\theta = \sigma^2$. El valor μ_0 es conocido. Como ya sabemos del ejemplo 5.3.4,

$$I_X(\theta) = \frac{1}{2\theta^2}.$$

Así que la cota de Cramér–Rao es $2\theta^2/n$.

Como ya vimos en el ejemplo 5.1.10, para la cuasivarianza muestral S^2 , que es un estimador insesgado de θ , se tiene que

$$\mathbf{V}_\theta(S^2) = \frac{2\theta^2}{n-1},$$

que no alcanza (por poco) la cota de Cramér–Rao.

Recuérdese que $X = \mu_0 + \sqrt{\theta} Z$, con $Z \sim \mathcal{N}(0, 1)$. Así que $\mathbf{E}_\theta((X - \mu_0)^2) = \theta$ y $\mathbf{E}_\theta((X - \mu_0)^4) = \theta^2 \mathbf{E}_\theta(Z^4) = 3\theta^2$.

Consideremos, por otro lado, el estadístico

$$T = \frac{1}{n} \sum_{j=1}^n (X_j - \mu_0)^2,$$

que es insesgado, pues

$$\mathbf{E}_\theta(T) = \mathbf{E}_\theta((X - \mu_0)^2) = \theta,$$

y además, es, sí, de mínima varianza, porque

$$\begin{aligned} \mathbf{E}_\theta(T^2) &= \frac{1}{n^2} \left(\sum_{j=1}^n \mathbf{E}_\theta((X_j - \mu_0)^4) + \sum_{1 \leq i \neq j \leq n} \mathbf{E}((X_i - \mu_0)^2) \mathbf{E}((X_j - \mu_0)^2) \right) \\ &= \frac{\theta^2}{n^2} (3n + n(n-1)) = \left(1 + \frac{2}{n}\right) \theta^2 \end{aligned}$$

y, por tanto

$$\mathbf{V}_\theta(T) = \mathbf{E}_\theta(T^2) - \mathbf{E}_\theta(T)^2 = \left(1 + \frac{2}{n}\right) \theta^2 - \theta^2 = \frac{2\theta^2}{n},$$

que es justo la cota de Cramér–Rao. ♣

EJEMPLO 5.3.12. *Cota de Cramér–Rao para $X \sim \mathcal{N}(\mu, \sigma_0^2)$, con σ_0^2 conocida.*

Queremos estimar $\theta = \mu$. Como ya sabemos del ejemplo 5.3.5,

$$I_X(\mu) = \frac{1}{\sigma_0^2},$$

de donde la cota de Cramér–Rao es σ_0^2/n .

La media muestral \bar{X} es estimador insesgado de μ , y $\mathbf{V}(\bar{X}) = \mathbf{V}(X)/n = \sigma_0^2/n$. Así que la cota se alcanza y \bar{X} es estimador insesgado de mínima varianza. ♣

EJEMPLO 5.3.13. *Sea X una variable con función de densidad $f(x; \alpha) = \alpha x^{\alpha-1}$ para $0 < x < 1$, donde α es un parámetro, $\alpha > 0$.*

Nos interesa estimar el parámetro $\theta = 1/\alpha$. Como ya sabemos del ejemplo 5.3.7,

$$I_X(\theta) = \frac{1}{\theta^2},$$

y, por tanto, que la cota de Cramér–Rao es θ^2/n .

Usando que $\mathbf{E}_\theta(\ln(1/X)) = \theta$ y $\mathbf{V}_\theta(\ln(1/X)) = \theta^2$, tal y como vimos en el citado ejemplo 5.3.7, resulta que el estadístico

$$T(X_1, \dots, X_n) = \frac{1}{n} \sum_{j=1}^n \ln \left(\frac{1}{X_j} \right)$$

es un estimador insesgado de θ , y además $\mathbf{V}_\theta(T) = \mathbf{V}_\theta(\ln(1/X))/n = \theta^2/n$. Así que el estadístico T es estimador insesgado de mínima varianza de θ . ♣

EJEMPLO 5.3.14. *Cota de Cramér–Rao para la distribución de Rayleigh.*

Para $X \sim \text{RAY}(\sigma^2)$, nombramos el parámetro que queremos estimar como $\theta = \sigma^2$. Como ya sabemos del ejemplo 5.3.6, la cantidad de información viene dada por

$$I_X(\theta) = \frac{1}{\theta^2},$$

de manera que la cota de Cramér–Rao es θ^2/n .

Recuerde, lector, de (3.15), que $\mathbf{E}_\theta(X^2) = 2\theta$ y que $\mathbf{E}_\theta(X^4) = 8\theta^2$.

El estimador T de máxima verosimilitud de θ (y también uno de los estimadores por momentos), véase el ejemplo 5.2.10, viene dado por $T = \frac{1}{2}\overline{X^2}$. Como $\mathbf{E}_\theta(X^2) = 2\theta$, se tiene que $\mathbf{E}_\theta(T) = \frac{1}{2}\mathbf{E}_\theta(\overline{X^2}) = \frac{1}{2}\mathbf{E}_\theta(X^2) = \theta$, y por tanto el estimador T es insesgado.

Además,

$$\begin{aligned} \mathbf{E}_\theta(T^2) &= \frac{1}{4} \mathbf{E}_\theta \left[\left(\frac{1}{n} \sum_{j=1}^n X_j^2 \right)^2 \right] = \frac{1}{4} \frac{1}{n^2} \left(\sum_{j=1}^n \mathbf{E}_\theta(X_j^4) + \sum_{1 \leq i \neq j \leq n} \mathbf{E}_\theta(X_i^2) \mathbf{E}_\theta(X_j^2) \right) \\ &= \frac{\theta^2}{4n^2} (8n + 4n(n-1)) = \left(1 + \frac{1}{n} \right) \theta^2, \end{aligned}$$

de manera que

$$\mathbf{V}_\theta(T) = \mathbf{E}_\theta(T^2) - \mathbf{E}_\theta(T)^2 = \left(1 + \frac{1}{n} \right) \theta^2 - \theta^2 = \frac{\theta^2}{n},$$

y, por tanto, T es estimador eficiente. ♣

5.3.3. Complementos sobre la cota de Cramér–Rao

Recogemos en este apartado unas cuantas observaciones sobre la cota de Cramér–Rao, a saber,

- las versiones generales del lema 5.4 de Diotivede y de la cota de Cramér–Rao (teorema 5.6) para variables continuas (o discretas con soporte numerable);
- la expresión de la cota de Cramér–Rao para el caso de estimadores sesgados;
- la unicidad del estimador eficiente (cuando éste exista) y una “expresión” general para el estimador eficiente;
- y, finalmente, una reflexión sobre el caso de las variables cuyo soporte depende del parámetro.

A. Condiciones para el lema de Diotivede y el teorema de Cramér–Rao

En el argumento que nos condujo al lema 5.4 de Diotivede en el caso en que la variable X , plácida ella, tiene soporte finito, se ha derivado la ecuación (5.20) y se ha usado que la derivada de una suma (finita) es la suma de las derivadas.

Cuando la variable X es continua, las esperanzas son integrales. Queremos derivar bajo el signo integral la expresión

$$1 = \int_{\text{sop}} f(x; \theta) dx$$

para obtener que

$$0 = \int_{\text{sop}} \partial_{\theta} f(x; \theta) dx = \int_{\text{sop}} \frac{\partial_{\theta} f(x; \theta)}{f(x; \theta)} f(x; \theta) dx,$$

que ya es la conclusión del lema de Diotivede: $\mathbf{E}_{\theta}(Y) = 0$.

La siguiente condición:

para cada intervalo cerrado $I \subset \Theta$ se tiene que

$$[\text{DIO}] \quad \int_{x \in \text{sop}} \sup_{\theta \in I} |\partial_{\theta} f(x, \theta)| dx < +\infty,$$


permite el intercambio de derivación e integración que hemos señalado más arriba.

Lema 5.7 (Diotivede) *Si X es una variable aleatoria continua, plácida y que cumple [DIO], entonces*

$$\mathbf{E}_{\theta}(Y) = 0, \quad \text{para cada } \theta \in \Theta,$$

Además,

$$I_X(\theta) = \mathbf{V}_{\theta}(Y) = \mathbf{E}_{\theta}\left(\left(\frac{\partial_{\theta} f(X; \theta)}{f(X; \theta)}\right)^2\right) = \int_{\text{sop}} \left(\frac{\partial_{\theta} f(x; \theta)}{f(x; \theta)}\right)^2 f(x; \theta) dx.$$

 **Nota 5.3.3.** Va aquí el detalle de la justificación del intercambio entre derivación a integración. Denotemos

$$A_I(x) \triangleq \sup_{\theta \in I} |\partial_{\theta} f(x, \theta)|, \quad \text{para } x \in \text{sop}.$$

Fijemos $\theta \in \Theta$. Sea $I \subset \Theta$ un intervalo cerrado que contiene a θ en su interior. Sea $\delta > 0$, tal que $\theta + \delta \in I$. Como la variable es plácida, y en particular por la propiedad c) de la definición 5.3, tenemos que, para $x \in \text{sop}$,

$$f(x; \theta + \delta) - f(x; \theta) - \delta \partial_{\theta} f(x; \theta) = \int_{\theta}^{\theta + \delta} \int_{\theta}^{\eta} \partial_{\theta\theta} f(x; \phi) d\phi d\eta;$$

(esta expresión no es más que el polinomio de Taylor de grado 1, con término de error), y por tanto,

$$(\star) \quad |f(x; \theta + \delta) - f(x; \theta) - \delta \partial_{\theta} f(x; \theta)| \leq \frac{1}{2} \delta^2 A_I(x).$$

Como $f(x; \theta + \delta)$ y $f(x; \theta)$ son funciones de densidad,

$$(b) \quad \int_{\mathbf{sop}} f(x; \theta + \delta) dx = 1 = \int_{\mathbf{sop}} f(x; \theta) dx.$$

Por (\star) y (b) (división por δ mediante) se tiene que

$$\left| \int_{x \in \mathbf{sop}} f_{\theta}(x; \theta) dx \right| \leq \frac{1}{2} \delta \int_{x \in \mathbf{sop}} A_I(x) dx.$$

Obsérvese que el lado izquierdo de la identidad anterior no depende de δ . Como, por la hipótesis [DIO], la integral $\int_{x \in \mathbf{sop}} A_I(x) dx$ es finita, se deduce haciendo $\delta \rightarrow 0$ que

$$\int_{\mathbf{sop}} f_{\theta}(x; \theta) dx = 0,$$

y, por tanto,

$$\mathbf{E}_{\theta}(Y) = \int_{\mathbf{sop}} \partial_{\theta} \ln f(x; \theta) f_{\theta}(x; \theta) dx = \int_{\mathbf{sop}} f_{\theta}(x; \theta) dx = 0.$$

Para el caso en que X es variable aleatoria discreta (infinita), plácida, el resultado es el mismo sustituyendo la condición [DIO] del enunciado por

para cada intervalo cerrado $I \subset \Theta$ se tiene

$$[\text{DIO}]^{\star} \quad \sum_{x \in \mathbf{sop}} \sup_{\theta \in I} |\partial_{\theta\theta} f(x, \theta)| dx < +\infty$$



Nota 5.3.4. La demostración es análoga a la de la nota 5.3.3: basta reemplazar $\int_{x \in \mathbf{sop}}$ por $\sum_{x \in \mathbf{sop}}$ en cada ocurrencia.

Estas condiciones [DIO] o [DIO] * , se cumplen para toda la batería de modelos para X que se usan en la práctica.

EJEMPLO 5.3.15. Condiciones del lema 5.7 para la familia $\text{EXP}(\lambda)$.

Para la familia exponencial tenemos que $\mathbf{sop} = (0, +\infty)$ y que $\Theta = (0, +\infty)$. Para $x \in \mathbf{sop}$ y $\lambda > 0$ se tiene $f(x; \lambda) = \lambda e^{-\lambda x}$.

Tenemos

$$\partial_{\lambda\lambda} f(x; \lambda) = -(2x + \lambda x^2) e^{-\lambda x}, \quad \text{para } \lambda > 0 \text{ y } x > 0,$$

y $f(x; \lambda)$ es plácida. Si $I = [\alpha, \beta] \subset (0, +\infty) = \Theta$, (así que $\alpha > 0$ y $\beta < +\infty$) se tiene

$$A_I(x) \leq (2x + \beta x^2) e^{-\alpha x}, \quad \text{para } x > 0,$$

De manera que

$$\int_0^{\infty} A_I(x) dx \leq \int_0^{\infty} (2x + \beta x^2) e^{-\alpha x} dx = \frac{2}{\alpha^2} + \frac{2\beta}{\alpha^3} < +\infty.$$



En la demostración del teorema 5.6 de Cramér–Rao en el caso de soporte finito, además de derivar la función de densidad como en el lema de Diotivede, se deriva la identidad

$$\theta = \mathbf{E}_\theta(T) = \mathbf{E}_\theta(h(X_1, \dots, X_n)) = \sum_{\mathbf{x} \in \text{sop}^n} h(\mathbf{x}) f(\mathbf{x}; \theta),$$

respecto de $\theta \in \Theta$, donde T es el estimador insesgado dado por $T = h(X_1, \dots, X_n)$.

En el caso continuo habremos de derivar

$$\theta = \mathbf{E}_\theta(T) = \mathbf{E}_\theta(h(X_1, \dots, X_n)) = \int_{\mathbf{x} \in \text{sop}^n} h(\mathbf{x}) f(\mathbf{x}; \theta) d\mathbf{x}$$

La condición

para cada intervalo cerrado $I \subset \Theta$ se tiene que

$$[\text{CR}] \quad \int_{\mathbf{x} \in \text{sop}} h(\mathbf{x}) \sup_{\theta \in I} |\partial_{\theta\theta} f(\mathbf{x}, \theta)| d\mathbf{x} < +\infty,$$

que involucra a la variable X y al estadístico T , permite intercambiar derivada e integral y probar:

Teorema 5.8 (Cota de Cramér–Rao) *Sea X es una variable aleatoria continua, plácida y que cumple [DIO], y sea T un estadístico insesgado que se cumple [CR]. Entonces,*

$$\mathbf{V}_\theta(T) \geq \frac{1}{n I_X(\theta)}, \quad \text{para todo } \theta \in \Theta.$$

La condición [DIO] nos da que $\mathbf{E}_\theta(Y) = 0$ y que $\mathbf{E}_\theta(Z_n) = 0$, mientras que la condición [CR] nos permite derivar bajo el signo integral de manera análoga a como se hizo en la demostración del lema 5.7 de Diotivede.

Para variables discretas (de soporte infinito) la condición [CR] se traslada en

para cada intervalo cerrado $I \subset \Theta$ se tiene que

$$[\text{CR}]^* \quad \sum_{\mathbf{x} \in \text{sop}} h(\mathbf{x}) \sup_{\theta \in I} |\partial_{\theta\theta} f(\mathbf{x}, \theta)| < +\infty.$$

y bajo [DIO]^{*} y [CR]^{*} la conclusión de la cota de Cramér–Rao sigue siendo válida.

B. Cota de Cramér–Rao, caso sesgado

El siguiente resultado exhibe una suerte de cota de Cramér–Rao para estimadores generales, no necesariamente insesgados.

Sea T un estadístico para el que

$$(\star) \quad \mathbf{E}_\theta(T) = m_T(\theta) \quad \text{para cada } \theta \in \Theta,$$

donde $m_T(\theta)$ es una cierta función de θ .

Si fuera $m_T(\theta) = \theta$, entonces diríamos que T es un estadístico estimador de θ *insesgado*. En caso contrario, podemos interpretar (\star) de dos maneras, alternativas (y sugerentes):

- T es un estadístico estimador de θ *sesgado*, donde el sesgo es $m_T(\theta) - \theta$;
- o bien T es un estimador *insesgado* de $m_T(\theta)$, que es una cierta función de parámetro θ .

En cualquier caso, y siguiendo fielmente los pasos de la demostración del teorema 5.6, obtendríamos primero que

$$m'_T(\theta) = \mathbf{E}_\theta(T \cdot Z_n) = \text{cov}(T, Z_n),$$

y luego, usando la desigualdad de Cauchy–Schwarz del corolario 2.4, tendríamos que

$$|m'_T(\theta)|^2 = |\text{cov}(T, Z_n)|^2 \leq \mathbf{V}_\theta(T) \mathbf{V}_\theta(Z_n) = \mathbf{V}_\theta(T) n I_X(\theta).$$

Es decir,

Teorema 5.9 *Con las notaciones anteriores, para todo $\theta \in \Theta$, se tiene que*

$$\mathbf{V}_\theta(T) \geq \frac{m'_T(\theta)^2}{n I_X(\theta)}.$$

Observe, lector, que ahora la cota para la varianza de T es algo menos interesante que en el caso insesgado, pues no depende únicamente de la función de densidad $f(x; \theta)$, sino que incorpora el factor $m_T(\theta)$, que depende del estimador T utilizado.

EJEMPLO 5.3.16. *La exponencial y sus parámetros.*

Del ejemplo 5.3.10, recordamos que la información de Fisher (con respeto del parámetro λ) es $I_X(\lambda) = 1/\lambda^2$.

Supongamos que tenemos un estadístico T tal que

$$\mathbf{E}_\lambda(T) = \frac{1}{\lambda}.$$

Es decir, T es un estimador insesgado de $1/\lambda$. Apelando al teorema 5.9, y tomando allí $m(\lambda) = 1/\lambda$, para el que $m'(\lambda) = -1/\lambda^2$, concluimos que

$$\mathbf{V}_\lambda(T) \geq \frac{1/\lambda^4}{n/\lambda^2} = \frac{1}{n\lambda^2},$$

que es justamente la cota de Cramér–Rao que obtuvimos en el ejemplo 5.3.10 con el (algo más tortuoso) camino consistente en reescribir la función de densidad de la exponencial en términos de $\theta = 1/\lambda$. ♣

C. Unicidad y existencia de estimadores eficientes

Lema 5.10 (Unicidad del estimador insesgado de mínima varianza) Sean T_1 y T_2 estimadores insesgados de θ , ambos de mínima varianza. Entonces $T_1 \equiv T_2$.

Aquí, $T_1 \equiv T_2$ significa que las variables son iguales con probabilidad 1.

En realidad, este resultado no depende de la cota de Cramér–Rao en sí; simplemente dice que, con mínima varianza (sea cual sea ésta), sólo puede haber un estimador insesgado.

DEMOSTRACIÓN. Sea S el estimador

$$S = \frac{1}{2}T_1 + \frac{1}{2}T_2.$$

El estimador S es insesgado. Es decir, $\mathbf{E}_\theta(S) = \theta$, para todo θ . Además,

$$\begin{aligned}\mathbf{V}_\theta(S) &= \frac{1}{4}\mathbf{V}_\theta(T_1) + \frac{1}{4}\mathbf{V}_\theta(T_2) + \frac{1}{2}\text{cov}_\theta(T_1, T_2) \\ &\leq \frac{1}{4}\mathbf{V}_\theta(T_1) + \frac{1}{4}\mathbf{V}_\theta(T_2) + \frac{1}{2}\sqrt{\mathbf{V}_\theta(T_1)}\sqrt{\mathbf{V}_\theta(T_2)} = \mathbf{V}_\theta(T_1) = \mathbf{V}_\theta(T_2).\end{aligned}$$

Como T_1 (ó T_2) son de mínima varianza (estrictamente positiva), se ha de cumplir que $\mathbf{V}_\theta(S) = \mathbf{V}_\theta(T_1) = \mathbf{V}_\theta(T_2)$, para cada θ . Es decir, en la cadena anterior ha de haber igualdades, y por tanto

$$\text{cov}_\theta(T_1, T_2) = \sqrt{\mathbf{V}_\theta(T_1)}\sqrt{\mathbf{V}_\theta(T_2)}.$$

Así que estamos en el caso de igualdad en Cauchy–Schwarz (corolario 2.4), y por tanto, para ciertas funciones $a(\theta)$ y $b(\theta)$, se cumple que

$$(\star) \quad T_2 = a(\theta)T_1 + b(\theta).$$

Como T_1 y T_2 son insesgados y de igual varianza, deducimos⁶ que $a \equiv 1$ y $b \equiv 0$. ■

El siguiente resultado nos da una expresión para el estimador eficiente, en caso de que exista.

Proposición 5.11 Sea X una variable plácida con función de densidad $f(x; \theta)$. Llamamos $Z_n = Y_1 + \cdots + Y_n$ a la variable de información total para muestras de tamaño n . Si T es estimador eficiente de θ , entonces

$$T \equiv \frac{Z_n}{nI_X(\theta)} + \theta.$$

⁶En realidad, de la condición de iguales varianzas se deduce que $a^2(\theta) = 1$, y por tanto $a(\theta) = \pm 1$. Tomando $a(\theta) = 1$ se deduce que $b(\theta) = 0$. La otra “solución”, $a(\theta) = -1$, llevaría a $b(\theta) = 2\theta$; pero esto contradiría la definición de estadístico, que no puede contener referencias al parámetro θ .

Observe, lector, que implícita en el enunciado subyace la hipótesis de que existe un tal estimador eficiente.

Recuerde, lector, que un estimador es una expresión $T = h(X_1, \dots, X_n)$ que depende tan sólo de la muestra y en la que, por supuesto, no puede aparecer el parámetro. En la expresión de la proposición 5.11 aparece el parámetro θ en Z_n , en $I_X(\theta)$, y también en el sumando θ . Así que esta expresión de T sólo será útil si, como por ensalmo, todas estas apariciones de θ se cancelan y T no depende de θ .

Si ése fuera el caso, si

$$T = \frac{Z_n}{nI_X(\theta)} + \theta$$

fuera realmente un estadístico estimador, entonces sería insesgado, pues $\mathbf{E}_\theta(Z_n) = 0$, y de mínima varianza, pues $\mathbf{V}_\theta(Z_n) = nI_X(\theta)$, y por tanto, $\mathbf{V}_\theta(T) = 1/(nI_X(\theta))$.

DEMOSTRACIÓN. Si T es de mínima varianza, entonces

$$1 = \text{cov}_\theta(T, Z_n) \leq \sqrt{\mathbf{V}_\theta(T)} \sqrt{\mathbf{V}_\theta(Z_n)} = \sqrt{\mathbf{V}_\theta(T)} \sqrt{nI_X(\theta)} = 1,$$

y, por tanto, $\text{cov}_\theta(T, Z_n) = \sqrt{\mathbf{V}_\theta(T)} \sqrt{\mathbf{V}_\theta(Z_n)}$. El caso de igualdad en la desigualdad de Cauchy-Schwarz nos dice que

$$T \equiv \alpha(\theta) Z_n + \beta(\theta),$$

para ciertas funciones $\alpha(\theta)$ y $\beta(\theta)$. Como $\mathbf{E}_\theta(T) = \theta$ y $\mathbf{E}_\theta(Z_n) = 0$, ha de ser $\beta = \theta$, y como $\mathbf{V}_\theta(T) = 1/(nI_X(\theta))$ y $\mathbf{V}_\theta(Z_n) = nI_X(\theta)$, ha de ser $\alpha = 1/(nI_X(\theta))$. ■

En el ejemplo 5.3.11 sobre la cota de Cramér–Rao y estimadores eficientes para $\mathcal{N}(\mu_0, \sigma^2)$, con μ_0 conocido, vimos que el estimador natural S^2 no era estimador eficiente de σ^2 . Inopinadamente se propuso allí $(\bar{X} - \mu_0)^2$ como estimador. Veamos.

Pongamos, por simplificar notación, que $\mu_0 = 0$. De los ejemplos 5.3.4 y 5.3.11 obtenemos que para $\mathcal{N}(0, \sigma^2)$, poniendo $\theta = \sigma^2$, tenemos que

$$Y = \frac{X^2 - \theta}{2\theta^2} \quad \text{y} \quad Z_n = n \frac{\bar{X}^2 - \theta}{2\theta^2}.$$

Además, $I_X(\theta) = 1/(2\theta^2)$.

Por tanto, siguiendo la proposición 5.11, el estimador eficiente, de existir, ha de escribirse como

$$T = \frac{Z_n}{nI_X(\theta)} + \theta = \frac{2\theta^2}{n} n \frac{\bar{X}^2 - \theta}{2\theta^2} - \theta = \bar{X}^2,$$

tras sustituir las expresiones de Z_n y de $I_X(\theta)$ y asistir, asombrados, a la fulminante cancelación de todas las apariciones del parámetro.

D. Soporte dependiente del parámetro

Hemos derivado el lema de Diotivede y la cota de Cramér–Rao para variables plácidas con alguna hipótesis adicional en el caso de soporte no finito. Una de las

condiciones para que una variable sea plácida es que su soporte sea independiente del parámetro $\theta \in \Theta$. Esta hipótesis no es únicamente conveniente para los cálculos, sino de hecho, indispensable.

En el ejemplo siguiente, la uniforme $\text{UNIF}[0, a]$, vamos a ver que *no se cumplen* las conclusiones ni del lema de Diotivede ni de la cota de Cramér–Rao, tal y como se han formulado.

EJEMPLO 5.3.17. $X \sim \text{UNIF}[0, a]$. ¿Cota? de Cramér–Rao.

Si procedemos directamente,

$$f(x; a) = \frac{1}{a} \implies \ln(f(x; a)) = -\ln(a) \implies \partial_a \ln(f(x; a)) = -1/a.$$

Así que $Y \equiv -1/a$. Obsérvese que ya no se cumpliría que $\mathbf{E}_a(Y) = 0$, como debería. Además $\mathbf{E}_a(Y^2) = 1/a^2$. Así que la cota de Cramér–Rao debería ser a^2/n .

Sin embargo, en el ejemplo 5.1.9 hemos comprobado que el estimador insesgado $T = \frac{n+1}{n} \max(X_1, \dots, X_n)$ tiene varianza

$$\mathbf{V}_a(T) = \frac{1}{n(n+2)} a^2.$$

Por último, nótese, en cualquier caso, que $\mathbf{V}_a(Y) = 0$.

