ESTADÍSTICA I
Tema 3: Estimación puntual paramétrica

- Modelos paramétricos e inferencia paramétrica
- Estimadores. Concepto, error cuadrático medio y propiedades deseables
- Construcción de estimadores: el método de máxima verosimilitud
- Información de Fisher y cota de Frechet-Cramer-Rao. Estimadores eficientes.
- Comportamiento asintótico de los e.m.v.
- Construcción de estimadores: el método de los momentos
- Construcción de estimadores: metodología bayesiana

Modelos paramétricos e inferencia paramétrica

Formular un modelo estadístico para una v.a. X consiste en especificar cuál es la familia de posibles distribuciones de probabilidad de X.

Un modelo es paramétrico si cada distribución F_θ de la familia es totalmente conocida salvo por el valor de un parámetro $\theta \in \mathbb{R}^k$, es decir, la familia de posibles distribuciones es

$$\{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}.$$

El conjunto Θ de posibles valores del parámetro θ se denomina espacio paramétrico.

X	$f(\cdot\,;\theta)$
absolutamente continua	función de densidad
discreta	función de masa o de probabilidad

En los modelos que manejamos en Estadística I habitualmente la dimensión k del parámetro es 1 o 2.

Supondremos que se cumple la condición de identificabilidad: si $\theta \neq \theta'$, entonces $F_\theta \neq F_{\theta'}$.

El objetivo general de la inferencia estadística paramétrica consiste en, supuesto que X sigue un modelo paramétrico F_θ, inferir o extraer información sobre el parámetro θ a partir de una muestra de observaciones de X.

En particular, el objetivo de la estimación paramétrica puntual es estimar (o aproximar) el valor desconocido de θ. Esto se hace mediante un estimador, que es una función medible $T_n(X_1, \ldots, X_n)$ (o, más bien, una sucesión $\{T_n\}$ de funciones medibles) de los valores muestrales.

Cuestiones de interés en el Tema 3:

- determinar qué se entiende por una "buena" estimación;
- estudiar las propiedades de estos estimadores;
- ofrecer algún procedimiento general para construir estimadores;
- mostrar algunos ejemplos de especial relevancia.

Estimadores

Sean $X_1, \ldots, X_n$ v.a.i.i.d. con función de densidad o de masa $f(\cdot; \theta)$, donde θ es un parámetro desconocido, del que sólo se sabe que pertenece al espacio paramétrico $\Theta \subseteq \mathbb{R}$.

Un estimador de θ es una función medible $T_n(X_1, \ldots, X_n)$ que se utiliza para estimar o aproximar el valor de θ (es frecuente utilizar también la notación $\hat{\theta}_n$ para denotar un estimador de θ).

Es habitual definir los estimadores para todos los posibles valores de n. En este caso, la sucesión $\{T_n\}$ se suele llamar también, con abuso de notación, "estimador".

Por tanto, los estimadores son v.a. y tiene sentido hablar de su media, su varianza, su distribución, etc. La realización del estimador $T_n = T_n(x_1, \ldots, x_n)$ en una muestra concreta observada $x_1, \ldots, x_n$ se denomina estimación y es un valor numérico (no aleatorio).

La calidad de un estimador se puede evaluar mediante su error cuadrático medio

$$\text{ECM}(T_n) = \mathbb{E}[(T_n - \theta)^2]$$

Si suponemos que X sigue un modelo paramétrico de distribución de probabilidad, el ECM puede depender de θ, por lo que a veces se denota $\text{ECM}_\theta(T_n) = \mathbb{E}_\theta[(T_n - \theta)^2]$.

Sumando y restando $\mathbb{E}(T_n)$ es inmediato ver que

$$\text{ECM}(T_n) = \mathbb{E}[(T_n - \mathbb{E}(T_n))^2] + (\mathbb{E}(T_n) - \theta)^2$$

es decir,

$$\text{ECM}(T_n) = \text{Varianza de } T_n + (\text{Sesgo de } T_n)^2$$

El ECM es un caso particular de una función de riesgo $R(\theta, T_n) = \mathbb{E}_\theta(\ell(\theta, T_n))$, siendo $\ell(\theta, a) = (\theta - a)^2$ la función de pérdida cuadrática.

Propiedades interesantes de los estimadores:

➡ Ausencia de sesgo: Se dice que un estimador T_n es insesgado si, siempre que $X_i \sim f(.;\theta)$, se tiene que

$$\mathbb{E}_\theta(T_n) = \theta, \ \forall \theta \in \Theta.$$

Ejemplo: $\bar{X}$ es estimador insesgado de $\mu = \mathbb{E}(X)$.

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

es estimador insesgado de $\sigma^2 = \mathbb{V}(X)$. En cambio,

$$S^2_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

es estimador sesgado.

➡ Consistencia: Se dice que $\{T_n\} := \{T_n(X_1, \ldots, X_n)\}$ es consistente en probabilidad si, siempre que $X_i \sim f(.;\theta)$,

$$T_n \xrightarrow{P} \theta, \quad \forall \theta \in \Theta$$

Si $\xrightarrow{P}$ se reemplaza por $\xrightarrow{\text{c.s.}}$, se obtiene la consistencia fuerte (o casi segura).

TEOREMA DE LA APLICACIÓN CONTINUA.- *Sea $g : \mathbb{R} \to \mathbb{R}$ continua en todo punto de un conjunto C tal que $\mathbb{P}\{X \in C\} = 1$.*

(i) Si $X_n \xrightarrow[n\to\infty]{d} X$, entonces $g(X_n) \xrightarrow[n\to\infty]{d} g(X)$.

(ii) Si $X_n \xrightarrow[n\to\infty]{P} X$, entonces $g(X_n) \xrightarrow[n\to\infty]{P} g(X)$.

(iii) Si $X_n \xrightarrow[n\to\infty]{c.s.} X$, entonces $g(X_n) \xrightarrow[n\to\infty]{c.s.} g(X)$.

¿Cómo demostrar fácilmente la consistencia?

- Mediante la ley de los grandes números:
  Si $g : \mathbb{R} \longrightarrow \mathbb{R}$ es continua entonces, $T_n = g(\bar{X})$ es estimador consistente c.s. de $\theta = g(\mu)$.
- La consistencia en probabilidad se puede intentar probar usando la desigualdad de Markov:

$$\mathbb{P}\{|T_n - \theta| > \epsilon\} \leq \frac{\mathbb{E}|T_n - \theta|}{\epsilon}$$

$$\mathbb{P}\{|T_n - \theta| > \epsilon\} \leq \frac{\mathbb{E}[(T_n - \theta)^2]}{\epsilon^2}$$

y, por tanto,

$$\mathbb{E}|T_n - \theta| \to 0 \Rightarrow T_n \xrightarrow{P} \theta$$

$$\mathbb{E}[(T_n - \theta)^2] \to 0 \Rightarrow T_n \xrightarrow{P} \theta.$$

- Respecto a la consistencia c.s recordemos que $T_n \xrightarrow{c.s} \theta$ si y solo si

$$\mathbb{P}\{\omega \in \Omega : \lim_{n\to\infty} T_n(X_1(\omega), \ldots, X_n(\omega)) = \theta\} = 1. \qquad (1)$$

Esta condición es, en general, difícil de comprobar directamente. Por ello, es habitual utilizar condiciones suficientes. Por ejemplo, una condición suficiente para que $T_n \xrightarrow[n\to\infty]{\text{c.s.}} \theta$ es la condición de Borel-Cantelli

$$\sum_{n=1}^{\infty} \mathbb{P}\{|T_n - \theta| > \epsilon\} < \infty, \ \forall \epsilon > 0. \qquad (2)$$

Por la desigualdad de Markov, (2) puede establecerse a su vez usando alguna de las condiciones suficientes

$$\sum_{n=1}^{\infty} \mathbb{E}|T_n - \theta| < \infty \text{ o bien } \sum_{n=1}^{\infty} \mathbb{E}[(T_n - \theta)^2] < \infty.$$

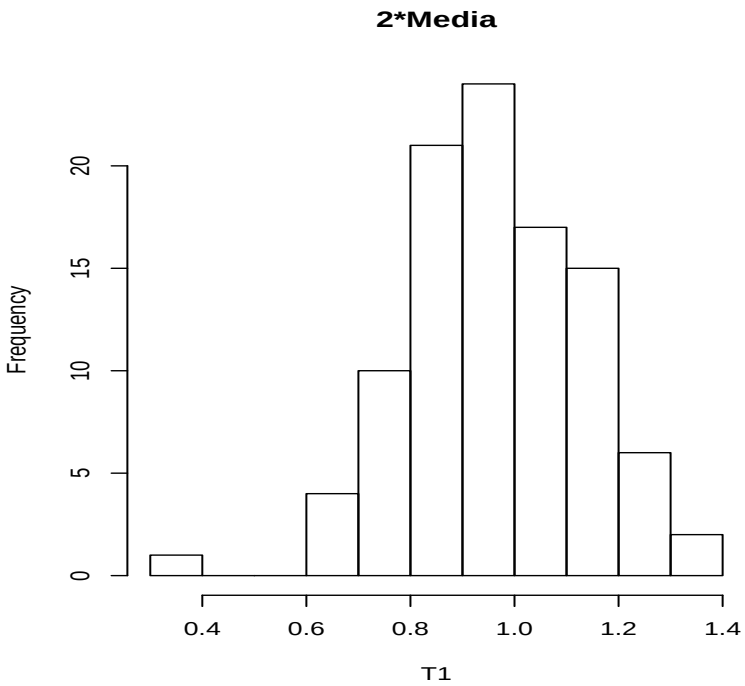
Ejercicio: Sean $X_1, \ldots, X_n$ v.a.i.i.d. con distribución uniforme en el intervalo $[0, \theta]$, $\theta > 0$. Estudiar la consistencia de los siguientes estimadores de θ:

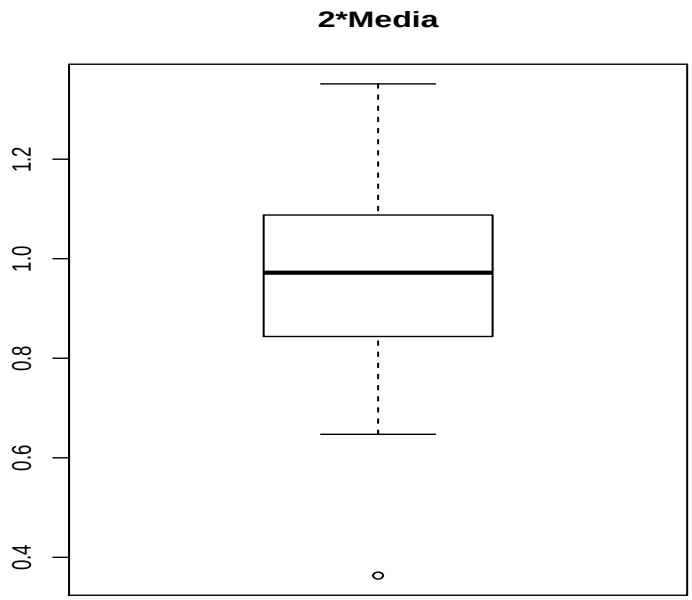
a) $T_n = 2\bar{X}$

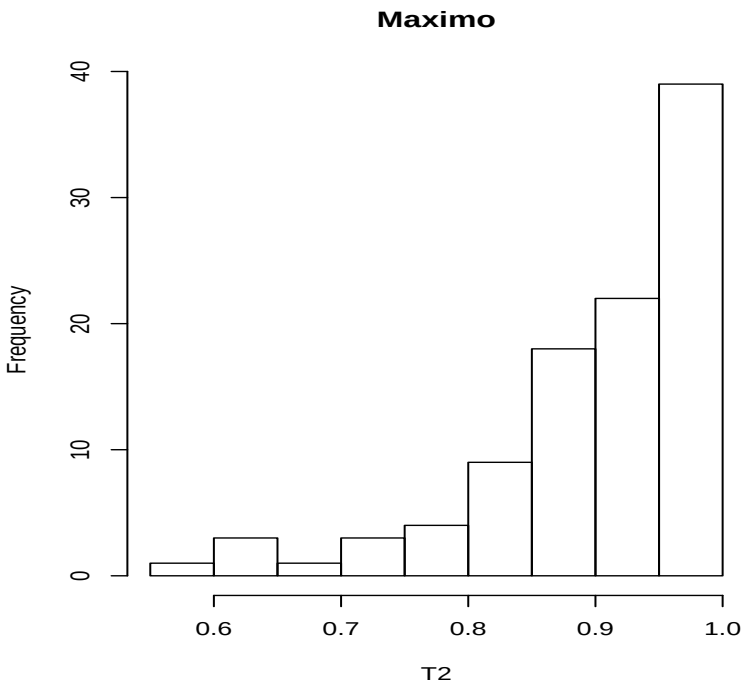
b) $T_n^* = X_{(n)} = \max\{X_1, \ldots, X_n\}$.

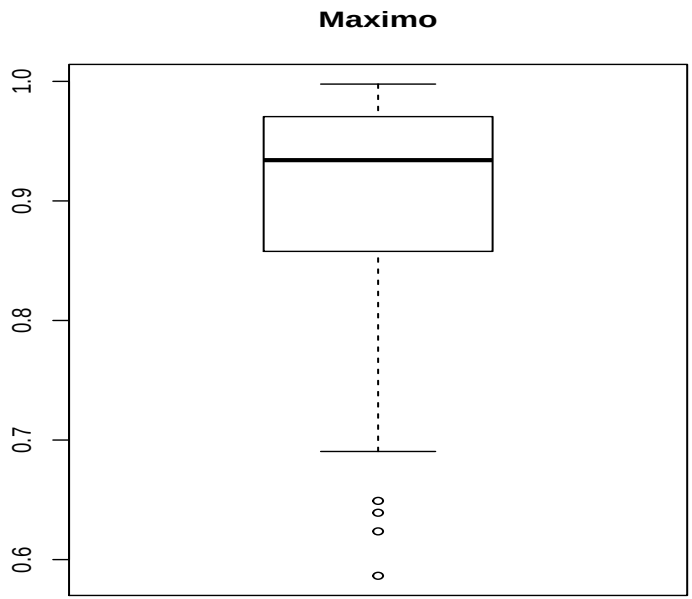
¿Cuál es el ECM de ambos estimadores? ¿Cuál de los dos estimadores preferiríamos?

```
# Comparación de dos estimadores del extremo derecho del
    soporte en una uniforme [0,1].

nMC = 100 # Número de muestras Monte Carlo (simulaciones)
n = 10 # Tamaño de cada muestra
T1 = rep(0,nMC)
T2 = rep(0,nMC)
for (i in 1:nMC){
            X = runif(n)
            T1[i] = 2*mean(X)
            T2[i] = max(X)
}

layout(matrix(1:4, 2, 2, byrow = TRUE))
hist(T1,main="2*Media")
boxplot(T1,main="2*Media")
hist(T2,main="Maximo")
boxplot(T2,main="Maximo")
par(def.par)  # vuelta al gráfico por defecto
```


2*Media
Frequency
0
5
10
15
20
0.4
0.6
0.8
1.0
1.2
1.4
T1


2*Media
0.4
0.6
0.8
1.0
1.2


Maximo
Frequency
0
10
20
30
40
0.6
0.7
0.8
0.9
1.0
T2


Maximo
0.6
0.7
0.8
0.9
1.0

➡ Normalidad asintótica: Se dice que una sucesión de estimadores $\{T_n\}$ del parámetro θ es asintóticamente normal (con tasa $\sqrt{n}$) si se satisface

$$\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{d} N(0, \sigma(\theta)), \tag{3}$$

donde $\sigma^2(\theta)$ es una cierta función del parámetro θ denominada varianza asintótica del estimador.

La propiedad (3) es muy útil para obtener una idea más precisa del error cometido en la estimación ya que permite hacer cálculos aproximados del tipo

$$\mathbb{P}\{|T_n - \theta| < c\} = \mathbb{P}\{\sqrt{n}|T_n - \theta| < \sqrt{n}c\} \simeq$$
$$\simeq \Phi\left(\frac{c}{\sigma(\theta)/\sqrt{n}}\right) - \Phi\left(-\frac{c}{\sigma(\theta)/\sqrt{n}}\right)$$

siendo Φ la función de distribución de la N(0,1).

¿Cómo se puede probar en la práctica la normalidad asintótica?

Si denotamos $\mu := \mathbb{E}(X)$ y $\sigma^2 := \mathbb{V}(X)$, por el TCL sabemos que

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow[n\to\infty]{d} N(0, \sigma).$$

Supongamos que $T_n = g(\bar{X})$, siendo $g : \mathbb{R} \longrightarrow \mathbb{R}$ función derivable con derivada continua. Se tiene entonces

$$\sqrt{n}(g(\bar{X}) - g(\mu)) \xrightarrow[n\to\infty]{d} N(0, |g'(\mu)|\sigma).$$

Este resultado se llama "método delta" y es consecuencia casi inmediata del teorema del valor medio y de las propiedades elementales de la convergencia en distribución.

Una observación sobre la normalidad asintótica:

En general, también se habla de normalidad asintótica (con tasa $\{a_n\}$) cuando se verifica un resultado del tipo

$$a_n(T_n - \theta) \xrightarrow[n\to\infty]{d} N(0, \sigma(\theta)), \tag{4}$$

donde $0 < a_n \nearrow \infty$ es una sucesión de constantes.

Tiene especial importancia el caso $a_n = \sqrt{n}$ porque es el que aparece en el Teorema Central del Límite. Sin embargo, en muy diversos problemas estadísticos se obtiene normalidad asintótica con sucesiones a_n diferentes.

Construcción del estimador de máxima verosimilitud

Planteamiento: Suponemos que la muestra $X_1, \ldots, X_n$ está formada por v.a.i.i.d. cuya distribución tiene una función de densidad o de masa $f(\cdot; \theta_0)$ perteneciente a una familia $\{f(\cdot; \theta) : \theta \in \Theta\}$. El verdadero valor del parámetro se denota por θ_0 y la letra θ designa un valor genérico de este parámetro.

Habitualmente θ_0 es desconocido. El primer objetivo de la inferencia paramétrica es definir un método general para estimarlo. En este tema la expresión "función de densidad" deberá entenderse en sentido genérico para referirse también a las funciones de masa de las variables discretas.

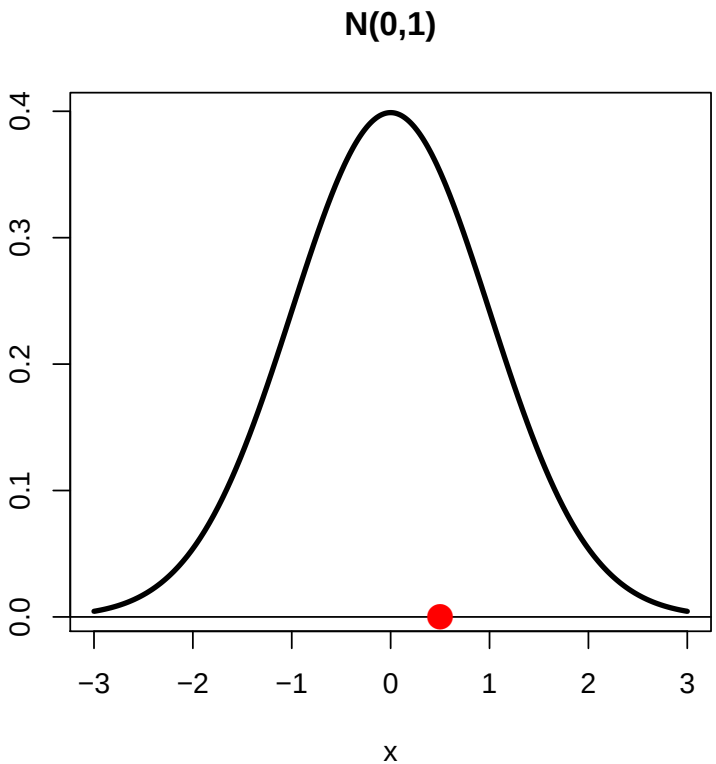
Condición de identificabilidad: Suponemos que se cumple

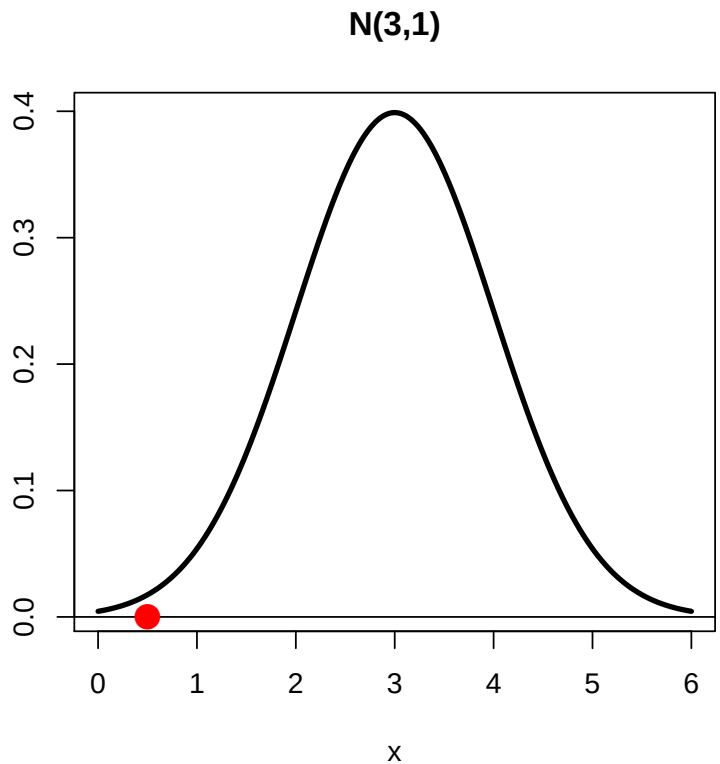
(MV0) Las distribuciones correspondientes a $f(\cdot; \theta)$ son distintas para diferentes valores de θ

El método de máxima verosimilitud

La idea es que la propia densidad $f(\ ;\theta)$ mide lo verosímil que es el valor de θ en base a la muestra observada.

¿Cuál de las dos densidades siguientes es más creíble que haya generado la observación $x = 0.5$ (en rojo)?





Dada una muestra fija $x_1, \ldots, x_n$, llamamos función de verosimilitud (*likelihood function*) a

$$L_n(\theta; x_1, \ldots, x_n) := L_n(\theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Denominamos estimador de máxima verosimilitud (EMV, *maximum likelihood estimator*, *MLE*) a

$$\hat{\theta}_n = \hat{\theta}_n(x_1, \ldots, x_n) = \arg\max_{\theta \in \Theta} L_n(\theta; x_1, \ldots, x_n), \tag{5}$$

cuando este máximo está bien definido.

También se puede utilizar $\log L_n$ en lugar de L_n en la expresión (5), ya que el logaritmo es una función creciente.

Cálculo efectivo del EMV:

Típicamente $\hat{\theta}_n$ se obtiene como solución de la ecuación de verosimilitud

estadístico gradiente (score statistic)

$$\frac{\partial}{\partial\theta}\log L_n = \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f(x_i;\theta) = 0.$$

Ejemplos:

- Para la distribución de Poisson de parámetro λ, $\hat{\lambda}_n = \bar{x}$.
- Para la exponencial de parámetro θ, $\hat{\theta}_n = \bar{x}$, si $f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}$, $x > 0$.
- Para la normal $N(\mu,\sigma)$, $\hat{\mu}_n = \bar{x}$, $\hat{\sigma}_n^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$.
- Para otras distribuciones, como la gamma o la Weibull, los e.m.v. de los parámetros existen pero no hay expresiones "cerradas" para ellos (aparecen como soluciones únicas de ciertas ecuaciones).

Propiedad de invariancia del e.m.v.: Si g es una función biyectiva, entonces el e.m.v. de $g(\theta)$ es $g(\hat{\theta})$, siendo $\hat{\theta}$ el e.m.v. de θ.

Información de Fisher y cota de Frechet-Cramer-Rao

Una comparación entre estimadores de un mismo parámetro θ en base a su ECM puede no dar un único óptimo porque la clase de todos los posibles estimadores es muy amplia. Una manera de buscar el "mejor estimador" es restringir la búsqueda en la clase de los estimadores insesgados.

Diremos que un estimador T_n^* de θ es el estimador insesgado uniformemente de mínima varianza (ECUMV) si $\mathbb{E}(T_n^*) = \theta$ para todo $\theta \in \Theta$ y $\mathbb{V}_\theta(T_n^*) \leq \mathbb{V}_\theta(T_n)$ para todo $\theta \in \Theta$ y para todo estimador insesgado T_n de θ.

Encontrar el ECUMV no es fácil. Una posibilidad es buscar una cota inferior para la varianza $\mathbb{V}_\theta(T_n)$ de cualquier estimador insesgado T_n de θ y después encontrar un estimador insesgado T_n^* cuya varianza alcance esa cota. Entonces habremos encontrado el ECUMV.

En la discusión que sigue suponemos que, para $k = 1$ y 2,

$$\frac{\partial^k}{\partial \theta^k} \int f(x;\theta) dx = \int \frac{\partial^k}{\partial \theta^k} f(x;\theta)\, dx.$$

Condiciones para permutar la derivada con la integral:

Sea una función $g = g(x;\theta)$, donde $x \in \mathbb{R}$ y $\theta \in \Theta$, un intervalo abierto de $\mathbb{R}$. Supongamos que

a) Para cada $\theta \in \Theta$, g es integrable como función de x (esto se cumple cuando g es una densidad de probabilidad).

b) Para casi todo x y para todo θ, existe $\dfrac{\partial}{\partial \theta} g(x;\theta)$.

c) Existe una función integrable $G : \mathbb{R} \to \mathbb{R}$, tal que, para todo θ,

$$\left| \frac{\partial}{\partial \theta} g(x;\theta) \right| \leq G(x).$$

Entonces, para todo θ, se cumple que

$$\frac{\partial}{\partial \theta} \int g(x;\theta) dx = \int \frac{\partial}{\partial \theta} g(x;\theta)\, dx.$$

Entonces, para cada $\theta \in \Theta$ fijo,

$$\int f(x;\theta)dx = 1 \Rightarrow \int \frac{\partial}{\partial\theta} f(x;\theta)dx = 0.$$

Por tanto,

$$\int \frac{\partial}{\partial\theta}(\log f(x;\theta))f(x;\theta)dx = \mathbb{E}_\theta[\frac{\partial}{\partial\theta}\log f(X;\theta)] = 0$$

y también (comprobarlo)

$$\mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right].$$

Si esta última cantidad es finita se denota por $I(\theta)$ y se denomina información de Fisher. Representa intuitivamente la "cantidad de información" acerca del valor del parámetro θ contenida en una observación de la v.a. X.

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] \\ &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] \\ &= \mathbb{E}_\theta\left[\left(\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\right)^2\right] \end{aligned}$$

Entonces, la v.a. $Z = \dfrac{\partial}{\partial\theta} \log L_n = \displaystyle\sum_{i=1}^{n} \dfrac{\partial}{\partial\theta} \log f(X_i;\theta)$ satisface

$$\mathbb{E}_\theta(Z) = 0,\ \mathbb{V}_\theta(Z) = n\, I(\theta).$$

Sea ahora $T_n = T_n(X_1, \ldots, X_n)$ un estimador insesgado de θ. Comprobemos que $\mathrm{Cov}_\theta(Z, T_n) = 1$. En efecto,

$$\begin{aligned}
&\mathbb{E}_\theta(ZT_n) \\
&\quad = \int \ldots \int T_n(x_1, \ldots, x_n) \left[\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i;\theta)}{f(x_i;\theta)} \right] \prod_{j=1}^{n} f(x_j,\theta) dx_1 \ldots dx_n
\end{aligned}$$

Pero

$$\left[\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i;\theta)}{f(x_i;\theta)} \right] \prod_{j=1}^{n} f(x_j,\theta) = \frac{\partial}{\partial\theta} \prod_{j=1}^{n} f(x_j,\theta)$$

Por tanto,

$$\begin{aligned}
\mathrm{Cov}_\theta(Z, T_n) &= \int \ldots \int T_n(x_1, \ldots, x_n) \frac{\partial}{\partial \theta} \prod_{j=1}^{n} f(x_j, \theta) dx_1 \ldots dx_n \\
&= \frac{\partial}{\partial \theta} \int \ldots \int T_n(x_1, \ldots, x_n) \prod_{j=1}^{n} f(x_j, \theta) dx_1 \ldots dx_n \\
&= \frac{\partial}{\partial \theta} \mathbb{E}_\theta(T_n) = \frac{\partial}{\partial \theta} \theta = 1.
\end{aligned}$$

Por otra parte, la desigualdad de Cauchy-Schwartz establece que

$$\mathrm{Cov}_\theta^2(Z, T_n) \leq \mathbb{V}_\theta(Z) \mathbb{V}_\theta(T_n)$$

es decir,

$$\mathbb{V}_\theta(T_n) \geq \frac{1}{n\, I(\theta)}.$$

En definitiva, hemos probado que

Supuesto que todas las integrales anteriores son finitas, que existen las derivadas utilizadas en los cálculos y que se dan las condiciones que permiten intercambiar las derivadas y las integrales, se verifica que para cualquier estimador insesgado de θ, T_n,

$$\mathbb{V}_\theta(T_n) \geq \frac{1}{n\, I(\theta)}$$

La expresión $\frac{1}{n\, I(\theta)}$ se denomina cota de Fréchet-Cramér-Rao. Los estimadores (insesgados) cuya varianza coincide con el valor de esta cota se denominan eficientes.

En general, si T_n no es insesgado se tiene que

$$\mathbb{V}_\theta(T_n) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(T_n)\right)^2}{n\, I(\theta)}.$$

Comportamiento asintótico de los e.m.v.

Evaluar las propiedades asintóticas de un estimador, T_n, de θ es una manera de estudiar su optimalidad mucho menos restrictiva que la teoría de "pequeñas muestras" (insesgadez, eficiencia,...).

Los estimadores de interés típicamente son consistentes y asintóticamente normales cuando $n \to \infty$. La varianza asintótica, $\sigma^2(\theta)$, proporciona una posible medida de la precisión del estimador. Si el estimador T_n es consistente y asintóticamente normal con

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

entonces decimos que T_n es *asintóticamente eficiente*.

Un estimador asintóticamente eficiente minimiza la varianza asintótica uniformemente en θ dentro de la clase de los estimadores asintóticamente insesgados.

TEOREMA 3.- *Supongamos que se cumplen las condiciones (MV0)*
(MV1) Θ *es abierto*
(MV2) Las distribuciones $f(\cdot;\theta)$ *tienen un soporte común.*
(MV3) Para cada x *la densidad* $f(x;\theta)$ *es tres veces diferenciable respecto a* θ*, con la tercera derivada continua en* θ*.*
(MV4) $0 < I(\theta_0) < \infty$
(MV5) $\Psi(\theta) := \mathbb{E}_{\theta_0}(\log f(X;\theta)) < \infty$ *para todo* $\theta \in \Theta$*.*
(MV6) Se puede permutar dos veces la integral $\int f(x;\theta)dx$ *con la derivada respecto a* θ*.*
(MV7) Para cada $\theta_0 \in \Theta$ *existen un número* $c > 0$ *y una función* $M(x)$ *(que pueden depender de* θ_0*) tales que* $\mathbb{E}_{\theta_0}[M(X)] < \infty$ *y*
$$\left|\frac{\partial^3 \log f}{\partial \theta^3}(x;\theta)\right| \leq M(x),\ \forall x,\ \theta \in (\theta_0 - c, \theta_0 + c).$$
(MV8) $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$
Entonces
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{\sqrt{I(\theta_0)}}\right).$$

Demostración del Teorema 3: Denotemos

$$\log L_n(\theta; X_1, \ldots, X_n) = \tilde{\Psi}_n(\theta) := \sum_{i=1}^{n} \log f(X_i; \theta)$$

$$\tilde{\Psi}'_n = (\partial/\partial\theta)\tilde{\Psi}_n$$

$$f' = (\partial/\partial\theta) f.$$

Recordemos que $\tilde{\Psi}_n(\theta)$ depende de la muestra. Para cada muestra fija se tiene

$$\begin{aligned} 0 = \tilde{\Psi}'_n(\hat{\theta}_n) &= \tilde{\Psi}'_n(\theta_0) + (\hat{\theta}_n - \theta_0)\tilde{\Psi}''_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\tilde{\Psi}'''_n(\theta_n^*) \\ &= \tilde{\Psi}'_n(\theta_0) + (\hat{\theta}_n - \theta_0)\left[\tilde{\Psi}''_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\tilde{\Psi}'''_n(\theta_n^*)\right] \end{aligned}$$

para algún θ_n^* entre $\hat{\theta}_n$ y θ_0. Como el primer miembro es 0, resulta

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{(1/\sqrt{n})\tilde{\Psi}'_n(\theta_0)}{-(1/n)\tilde{\Psi}''_n(\theta_0) - (1/2n)(\hat{\theta}_n - \theta_0)\tilde{\Psi}'''_n(\theta_n^*)} \qquad (6)$$

Demostración del Teorema 3 (cont.):

La idea clave de la demostración es:

Paso 1 Probar que el numerador de (6) tiende en distribución a una $N(0, \sqrt{I(\theta_0)})$ (recordemos que, en nuestra notación, el segundo parámetro de la normal denota la desv. típica, no la varianza).

Paso 2 Probar que el primer término del denominador de (6) tiende en probabilidad a $I(\theta_0)$.

Paso 3 Probar que el segundo término del denominador de (6) tiende en probabilidad a 0.

Paso 4 A partir de aquí el resultado es inmediato ya que, como consecuencia del Teorema de Slutsky,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \frac{1}{I(\theta_0)} N(0, \sqrt{I(\theta_0)}) = N\left(0, \frac{1}{\sqrt{I(\theta_0)}}\right).$$

Demostración del Teorema 3 (cont.):

Paso 1:

$$\frac{1}{\sqrt{n}}\tilde{\Psi}'_n(\theta_0) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta_0) - \mathbb{E}_{\theta_0}\left(\frac{\partial}{\partial\theta}\log f(X;\theta_0)\right)\right],$$

Como $\mathbb{E}_{\theta_0}\left(\frac{f'(X;\theta_0)}{f(X;\theta_0)}\right) = 0$, la aplicación del TCL (a las variables $Y_i = \frac{\partial}{\partial\theta}\log f(X_i;\theta_0)$) y la definición de $I(\theta_0)$ proporcionan directamente

$$(1/\sqrt{n})\tilde{\Psi}'_n(\theta_0) \xrightarrow{d} N(0, \sqrt{I(\theta_0)}).$$

Paso 2:

$$-\frac{1}{n}\tilde{\Psi}''_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{f'^2(X_i;\theta_0) - f(X_i;\theta_0)f''(X_i;\theta_0)}{f^2(X_i;\theta_0)}.$$

Por la LGN esto tiende en probabilidad a

$$I(\theta_0) - \mathbb{E}_{\theta_0}\left(\frac{f''(X;\theta_0)}{f(X;\theta_0)}\right) = I(\theta_0)$$

Demostración del Teorema 3 (cont.):

Paso 3: Por último,

$$\frac{1}{n}\tilde{\Psi}_n'''(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\log f(X_i;\theta),$$

de modo que

$$\left|\frac{1}{n}\tilde{\Psi}_n'''(\theta_n^*)\right| < \frac{1}{n}\sum_{i=1}^{n} M(X_i)$$

con probabilidad tendiendo a 1. Por tanto $\left|\frac{1}{n}\tilde{\Psi}_n'''(\theta)\right|$ está acotado en probabilidad ya que el segundo miembro de la desigualdad anterior tiende a $\mathbb{E}_{\theta_0}[M(X)]$. En definitiva, como $\hat{\theta}_n \xrightarrow{P} \theta_0$, se deduce

$$(1/2n)(\hat{\theta}_n - \theta_0)\tilde{\Psi}_n'''(\theta_n^*) \xrightarrow{P} 0,$$

lo que concluye la prueba del Paso 3 y del teorema. □

Para n suficientemente grande el resultado anterior concluye que

$$\hat{\theta} \overset{aprox}{\sim} N\left(\theta, \frac{1}{nI(\theta)}\right).$$

Es decir, para tamaños muestrales grandes esperamos que el emv sea aproximadamente insesgado y casi alcance la cota de F-C-R. Observemos que la variabilidad asintótica del emv es inversamente proporcional a la información de Fisher.

Los EMV son consistentes bajo las condiciones de regularidad. La consistencia resulta bastante plausible: por la L.G.N.,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i,\theta) \overset{c.s.}{\to} \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}f(X;\theta)\right).$$

Se cumple que $0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i,\hat{\theta})$ y también que $\mathbb{E}_{\theta_0}\left(\frac{\partial}{\partial\theta}f(X;\theta)\right) = 0$.

Es de esperar que para valores grandes de n, ambas ecuaciones se parezcan y por lo tanto sus respectivas soluciones, $\hat{\theta}_n$ y θ_0 también.

Sin embargo, la convergencia puntual de una sucesión de funciones no implica la convergencia de sus ceros a los de la función límite. Hace falta convergencia uniforme. Además, esta por sí misma no es suficiente ya que también hay que garantizar que la sucesión de soluciones esté eventualmente incluida en un conjunto cerrado y acotado. Por todo ello, demostrar en general la consistencia de los EMV es un resultado bastante técnico y que queda fuera del alcance de la asignatura. A menudo es más fácil probar directamente la consistencia para cada caso particular en el que estemos interesados.

El método de los momentos

Sea $X \sim f(\cdot;\theta)$, donde $\theta = (\theta_1, \ldots, \theta_p)$ es un parámetro p-dimensional ($p \geq 1$).
Si los momentos $\alpha_k(\theta) := \mathbb{E}_\theta(X^k)$, $k = 1, \ldots, p$, son funciones sencillas de los θ_i, un procedimiento natural para obtener un estimador de θ, es resolver en $\theta_1, \ldots, \theta_p$ el sistema de ecuaciones

$$m_1 = \alpha_1(\theta) \ , \ \ldots \ , \ m_p = \alpha_p(\theta),$$

donde m_k es el momento muestral de orden k, $m_k = \frac{\sum_{i=1}^n X_i^k}{n}$.
La idea es estimar el parámetro como aquel valor de θ que hace que los momentos poblacionales (tantos como componentes tenga θ) coincidan con los correspondientes momentos muestrales. En general, si θ_0 es el verdadero valor del parámetro, NO sucederá que $m_k = \alpha_k(\theta_0)$ (de hecho, m_k es aleatorio) pero, por la ley de los grandes números, $m_k \to \alpha_k(\theta_0)$ cuando $n \to \infty$.

- La principal ventaja del método de los momentos es que proporciona estimadores con expresión sencilla en algunos casos en los que el e.m.v. no se puede obtener en forma "cerrada" (porque aparece la solución de una ecuación complicada). Esto sucede, por ejemplo, en la estimación de los parámetros a y p en la distribución gamma.
- En general, el método de m.v. proporciona estimadores mejores (con menor error) que el de los momentos, aunque en algunos casos importantes ambos métodos llevan al mismo estimador.
- También puede plantearse el método de los momentos (en el caso $k = 2$) definiendo el estimador como solución de las ecuaciones $\mathbb{E}_\theta(X) = \bar{X}$, $\mathbb{V}_\theta(X) = s^2$, es decir, usando la varianza en lugar del momento de orden 2, $\mathbb{E}_\theta(X^2)$.

Ejemplos del método de los momentos:

Si se tiene el modelo

$$f(x;\theta) = \frac{1+\theta x}{2}\ \mathbb{1}_{[-1,1]}(x),\ \theta \in [-1,1]$$

no es sencillo calcular el e.m.v., pero sí obtener el estimador por el método de los momentos:

$$\mathbb{E}_\theta(X) = \int_{-1}^{1} x f(x;\theta) dx = \frac{\theta}{3}$$

Por tanto, la solución de $\bar{X} = \mathbb{E}_\theta(X)$ es $\tilde{\theta}_n = 3\bar{X}$, cuya varianza es

$$\mathbb{V}_\theta(\tilde{\theta}_n) = \mathbb{V}_\theta(3\bar{X}) = 9\frac{\sigma^2}{n} = \frac{3-\theta^2}{n}$$

ya que $\sigma^2 = \mathbb{V}_\theta(X) = \mathbb{E}_\theta(X^2) - (\mathbb{E}_\theta(X))^2 = \frac{1}{3} - \frac{\theta^2}{9}$. Este estimador es consistente ya que, por la LGN, $3\bar{X} \xrightarrow{c.s.} 3\mathbb{E}_\theta(X) = \theta$.

Si $X \sim \text{Beta}(a, b)$, $a, b > 0$

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x).$$

$$\mathbb{E}_\theta(X) = \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^a (1-x)^{b-1} dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+1)}{\Gamma(a+b+1)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} x^a (1-x)^{b-1} dx = \frac{a}{a+b}$$

$$\mathbb{V}_\theta(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Los estimadores por el método de los momentos son:

$$\hat{a} = \bar{X}\left(\frac{\bar{X}(1-\bar{X})}{s^2} - 1\right), \quad \hat{b} = (1-\bar{X})\left(\frac{\bar{X}(1-\bar{X})}{s^2} - 1\right).$$

Estimadores bayesianos

La distribución a priori

En muchos casos se tiene cierta información a priori (es decir, antes de extraer la muestra) sobre la probabilidad (entendida como "grado de creencia subjetiva") de los diferentes valores del parámetro θ. En estos casos se sabe, o se supone, que ciertos intervalos de valores de θ son "más probables que otros" y se concreta esta información en una distribución a priori sobre θ cuya función de densidad se denota $\pi(\theta)$.

Formalmente, este planteamiento equivale a considerar el parámetro θ como una variable aleatoria con distribución conocida. En consecuencia, la densidad $f(x;\theta)$ con la que se generan los datos puede considerarse como la densidad condicionada de la v.a. X dado el valor θ de la v.a. "parámetro". Por este motivo se suele emplear la notación $f(x|\theta)$ en lugar de $f(x;\theta)$.

Del mismo modo, la función de densidad de la muestra $f(x_1,\ldots,x_n;\theta) = \prod_{i=1}^n f(x_i|\theta)$ puede interpretarse (como función de $(x_1,\ldots,x_n)$ para cada θ fijo) como la densidad de la v.a. vectorial $(X_1,\ldots,X_n)$, condicionada al valor θ del parámetro. Por este motivo, suele denotarse (con cierto abuso de notación),

$$f(x_1,\ldots,x_n|\theta) := \prod_{i=1}^n f(x_i|\theta).$$

La distribución a posteriori

Esencialmente, la idea de la metodología bayesiana consiste en combinar la información a priori (recogida en $\pi(\theta)$) con la información muestral y la del modelo (recogidas en $f(x_1, \ldots, x_n|\theta)$). Esto se consigue aplicando la fórmula de Bayes para obtener la "distribución a posteriori", definida por la siguiente función de densidad en Θ

$$\pi(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta)\pi(\theta)}{\int_\Theta f(x_1, \ldots, x_n|\theta)\pi(\theta)d\theta} \tag{7}$$

La densidad a posteriori (7) recoge toda la información disponible sobre el parámetro y es la base de todos los procedimientos de inferencia en la metodología bayesiana.

El cálculo de $\pi(\theta|x_1,\ldots,x_n)$ puede ser complicado a veces, pero en muchos casos se simplifica recordando que $\pi(\theta|x_1,\ldots,x_n)$ debe ser una función de densidad en θ y, por tanto, debe integrar 1. Esto significa que la integral que aparece en el denominador de (7) puede ser considerada como una constante (en el sentido de que no depende de θ) y esta es justamente la constante necesaria para que la integral en θ de $\pi(\theta|x_1,\ldots,x_n)$ sea 1. Por tanto, en algunos ejemplos importantes, no es necesario calcular esa integral, siempre que, a partir del numerador de (7), se pueda identificar el tipo de distribución al que corresponde $\pi(\theta|x_1,\ldots,x_n)$.

Por ejemplo, si el numerador de la fórmula de Bayes (7) es, salvo constantes multiplicativas, de la forma $\theta^{p-1}e^{-a\theta}$, entonces no es necesario calcular la integral del denominador porque ya podemos asegurar que $\pi(\theta|x_1,\ldots,x_n)$ es una densidad gamma de parámetros a y p.

Ejemplo: Sea θ la proporción de votantes de un partido P. Sea X la v.a. Bernoulli que toma valor 1 cuando un votante elige el partido P y 0 en caso contrario:

$$f(x|\theta) = \theta^x(1-\theta)^{1-x}, \qquad x = 0 \text{ ó } 1.$$

Entonces tenemos que

$$f(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Suponemos que la distribución a priori es una Beta(4,10):

$$\pi(\theta) = \frac{\Gamma(14)}{\Gamma(4)\Gamma(10)}\theta^3(1-\theta)^9\mathbb{1}_{[0,1]}(\theta).$$

Aplicando la fórmula de Bayes (7), tenemos (el signo $\propto$ indica "proporcional a"):

$$\pi(\theta|x_1,\ldots,x_n) \propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\theta^3(1-\theta)^9 = \theta^{3+\sum x_i}(1-\theta)^{9+n-\sum x_i},$$

que corresponde a una densidad beta$(a = 4+\sum x_i, b = 10+n-\sum x_i)$.

Estimadores Bayes

El estimador Bayes se define, para cada muestra dada $(x_1, \ldots, x_n)$, como la esperanza de la distribución a posteriori

$$T_n(x_1, \ldots, x_n) = \int_\Theta \theta \pi(\theta | x_1, \ldots, x_n) d\theta.$$

Ejemplo (cont.): Hemos visto que $\pi(\theta | x_1, \ldots, x_n)$ tiene distribución $Beta(a = 4 + \sum x_i, b = 10 + n - \sum x_i)$. Por tanto el estimador Bayes es

$$T_n = \frac{4 + \sum x_i}{14 + n} = \left(\frac{n}{4 + 10 + n} \right) \bar{x} + \left(\frac{4 + 10}{4 + 10 + n} \right) \frac{4}{4 + 10}.$$

Es sencillo comprobar que en este caso el emv coincide con el estimador basado en el método de los momentos, $\hat{\theta}_n = \bar{x}$.
En la muestra concreta que tenemos $\sum x_i = 125$ y $n = 1000$, luego

$$T_n = \frac{129}{1014} = 0.12722, \ \hat{\theta}_n = \frac{125}{1000} = 0.125$$

En función de cuál sea la distribución a priori especificada y el modelo estadístico utilizado, los problemas computacionales que presenta el cálculo del estimador Bayes pueden ser muy difíciles, especialmente si θ es un vector de alta dimensión.

Tradicionalmente, con el fin de simplificar los cálculos, se elegía una distribución a priori de tal forma que la distribución a posteriori se pudiera identificar fácilmente.

Familias conjugadas

Cuando $X \sim \text{Bernoulli}(p)$, no es necesario suponer que $p \sim \text{Beta}$, pero sí tiene ventajas, como obtener una expresión cerrada para el estimador Bayes.

Sea $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$ una clase de densidades, por ejemplo, $\mathcal{F} = \{\text{Bernoulli}(p) : p \in (0,1)\} = \{f(x|p) = p^x(1-p)^{1-x} : p \in (0,1)\}$.

Una clase Π de distribuciones a priori es una *familia conjugada* de $\mathcal{F}$ si la distribución a posteriori $\pi(\theta|x_1, \ldots, x_n)$ también pertenece a Π, para toda $f \in \mathcal{F}$, para toda a priori $\pi \in \Pi$ y para toda muestra $x_1, \ldots, x_n$ con valores en el espacio muestral.

La familia $\Pi = \{\text{Beta}(a, b) : a, b > 0\}$ de las betas es la conjugada de la clase $\mathcal{F} = \{B(n, p) : p \in (0, 1)\}$ de las binomiales.

Mas recientemente se han desarrollado métodos numéricos basados en simulación de cadenas de Markov (Gibbs sampling y, más en general, métodos MCMC (Markov chain Monte Carlo)) que permiten extender la aplicación de los métodos bayesianos a modelos muy complejos.