Repaso de algunos conceptos de estadística y probabilidad

José R. Berrendero

- Modelos paramétricos y no paramétricos
- Distintos métodos de estimación
- Criterios para valorar un estimador
- Convergencias estocásticas
- Ley de los grandes números y teorema central del límite

Modelos paramétricos y no paramétricos

Sea $X_1, \ldots, X_n$ un conjunto de n variables (o vectores, o funciones) aleatorias y $x_1, \ldots, x_n$ los n valores observados o realizaciones correspondientes. La distribución conjunta de $X_1, \ldots, X_n$ no es conocida pero se supone que pertenece a una familia de posibles distribuciones. Formular un modelo estadístico es simplemente especificar cuál es la familia de posibles distribuciones. Un modelo es **paramétrico** si cada distribución de la familia es totalmente conocida salvo por el valor de un parámetro $\theta \in \mathrm{R}^d$. Es decir, la familia de posibles distribuciones es $\{F_\theta\colon \theta \in \Theta \subset \mathrm{R}^d\}$. El conjunto Θ de posibles valores del parámetro se llama **espacio paramétrico**. Siempre vamos a suponer que se da la siguiente **condición de identificabilidad**: si $\theta \neq \theta'$, entonces $F_\theta \neq F_{\theta'}$.

Pueden considerarse también modelos no paramétricos. Por ejemplo, $\{F\colon F \text{ tiene función de densidad } f\}$. En este caso, el espacio paramétrico es el conjunto de todas las funciones de densidad $\{f\colon f \geq 0,\ \int f = 1\}$, que tiene dimensión infinita. Una posible clase de estimadores en este contexto es la de estimadores del núcleo, que veremos más adelante.

Distintos métodos de estimación

Cuando se trabaja con un modelo paramétrico, el objetivo general de la inferencia estadística es extraer información sobre el parámetro θ a partir de las observaciones $x_1, \ldots, x_n$. Veamos tres enfoques bastante diferentes para obtener estimadores (aproximaciones al verdadero valor de θ) a partir de los datos muestrales:

- El método de momentos
- El método de máxima verosimilitud
- Los estimadores bayesianos

El método de momentos

Este es el método conceptualmente más sencillo de construcción de estimadores, aunque en general las propiedades de los estimadores obtenidos no suelen ser óptimas. No obstante, en algunos casos particulares importantes, este método proporciona los mismos resultados que otros procedimientos más sofisticados (como máxima verosimilitud).

Sea $X_1, \ldots, X_n$ vaiid con distribución determinada por $f(\cdot\,;\theta)$, donde $\theta = (\theta_1, \ldots, \theta_d)$ es un parámetro d-dimensional. Si los momentos de la distribución de X, $\alpha_k(\theta) := \mathrm{E}_\theta(X^k)$, $k = 1, \ldots, d$, son funciones sencillas de los los parámetros θ_i, un procedimiento natural para obtener un estimador de θ, es resolver en $\theta_1, \ldots, \theta_d$ el sistema de ecuaciones

$$m_1 = \alpha_1(\theta)\ ,\ \ldots\ ,\ m_d = \alpha_d(\theta),$$

donde m_k es el momento muestral de orden k, es decir, $m_k = \dfrac{\sum_{i=1}^n X_i^k}{n}$.

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 es de esperar que $m_k \approx \alpha_k(\theta_0)$ para tamaños muestrales grandes y, también, $\hat{\theta} \approx \theta_0$.

La principal ventaja del método de los momentos es su sencillez. Basta calcular los momentos muestrales y resolver una ecuación.

Ejemplo: distribución uniforme en $(0, \theta)$

Supongamos que $X_1, \ldots, X_n$ es una muestra de vaiid de una distribución uniforme en $(0, \theta)$. Entonces, $\mu = \theta/2$. El estimador de momentos debe verificar $\bar{x} = \hat{\theta}/2$, es decir, $\hat{\theta} = 2\bar{x}$.

Ejemplo: un caso particular de la distribución beta

Supongamos que $X_1, \ldots, X_n$ es una muestra de vaiid de una distribución con densidad $f(x; \theta) = (\theta + 1)x^\theta$, con $x \in [0, 1], \theta > -1$. ¿Cuál es el estimador de momentos de θ?

El método de máxima verosimilitud

Las realizaciones muestrales $x_1, \ldots, x_n$ son más o menos probables en función de lo que valga el parámetro θ. Por ejemplo, si sabemos que los datos vienen de una distribución $\mathrm{N}(\theta, 1)$, entonces si fuese $\theta = 4$ la probabilidad de que la mitad de los datos sea mayor que 4 es 1/2. Sin embargo, si fuese $\theta = 0$, es prácticamente imposible que algún dato sea mayor que 4. Por lo tanto, si nos encontráramos con una muestra para la que varias observaciones son mayores que 4 y tenemos que apostar entre los valores $\theta = 0$ y $\theta = 4$, ¿cuál sería la mejor opción?

El método de máxima verosimilitud consiste en estimar el parámetro θ mediante el valor que hace más verosímiles los datos que realmente hemos observado. Veamos un ejemplo adicional para insistir en la idea antes de dar una definición más formal.

Ejemplo

En una urna cerrada hay 4 bolas, θ de ellas son blancas y $4 - \theta$ son negras (θ desconocido). Se llevan a cabo dos extracciones de bolas con reemplazamiento. ¿Cuál será el estimador de máxima verosimilitud si una de las bolas extraídas es blanca y la otra es negra? Dicho más coloquialmente, si tuviéramos que apostar por el número de bolas blancas, ¿por qué valor apostaríamos dados los resultados obtenidos? La respuesta se deduce de la siguiente tabla, en la que aparecen las probabilidades de observar una bola blanca y otra negra en función de θ:

θ	$\mathbf{P}\{x_1 = B, x_2 = N\}$
0	0
1	3/16
2	4/16
3	3/16
4	0

Está claro que nadie apostaría por $\theta = 0$ o $\theta = 4$ ya que para estos dos valores es imposible obtener una bola negra y otra blanca. Observando la tabla, vemos que si $\theta = 2$ la probabilidad de obtener una bola negra y otra blanca es máxima y por lo tanto $\hat{\theta} = 2$ es el estimador de máxima verosimilitud. Supongamos que se extrae una tercera bola y resulta ser negra. ¿Cuál es el estimador de máxima verosimilitud en este caso?

La función de verosimilitud

Sea $x_1, \ldots, x_n$ una realización de una muestra $X_1, \ldots, X_n$ con función de densidad o de probabilidad conjunta $g(x_1, \ldots, x_n; \theta)$ en $(x_1, \ldots, x_n)$, donde $\theta \in \Theta \subset \mathrm{R}^d$. Para esta muestra, la **función de verosimilitud** $L : \Theta \to \mathrm{R}$ se define como

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = g(x_1, \ldots, x_n; \theta).$$

Si $X_1, \ldots, X_n$ son vaiid con densidad o probabilidad $f(x; \theta)$, entonces $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$

Hay una función de verosimilitud diferente para cada muestra, pero lo que interesa es cómo varía la verosimilitud al variar θ puesto que la muestra ya ha sido observada y, por lo tanto, está fija. A partir de $L(\theta)$ ya podemos definir el estimador de máxima verosimilitud.

Sea $x_1, \ldots, x_n$ una muestra y sea $L(\theta)$ la correspondiente función de verosimilitud. Un **estimador de máxima verosimilitud (EMV)** de θ es un valor $\hat{\theta} \in \Theta$ tal que

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

En el caso en el que Θ es un conjunto abierto y $L(\theta)$ es derivable y cóncava en Θ, para calcular el EMV basta derivar e igualar a cero. Sin embargo, como es más fácil derivar sumas que productos, en vez de maximizar directamente $L(\theta)$ resulta más conveniente maximizar $\ell(\theta; x) = \ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i, \theta)$. Como el logaritmo es una función creciente, $L(\theta)$ y $\ell(\theta)$ alcanzan el máximo en el mismo punto.

Ejemplo: datos censurados

En este ejemplo algunos de los factores de la función de verosimilitud son probabilidades y otros son funciones de densidad: el tiempo de vida de los ratones con cierta enfermedad sometidos a un tratamiento es una v.a. con distribución exponencial de parámetro θ. Se lleva a cabo un experimento con n ratones, se observa el tiempo de vida de m de ellos $(x_1, \ldots, x_m)$ pero se interrumpe el experimento transcurrido un tiempo T de manera que de los $n - m$ restantes solo se sabe que su tiempo de vida es superior a T. Calcula el EMV de θ. Se supone que todos los tiempos son independientes.

Si $f(x_i; \theta)$ es la función de densidad exponencial y $F(x_i; \theta)$ la función de distribución, observa que la función de verosimilitud en este caso es

$$L(\theta) = (1 - F(T; \theta))^{n-m} \prod_{i=1}^{m} f(x_i; \theta).$$

Ejemplo: distribución uniforme en $(0, \theta)$

Supongamos que $X_1, \ldots, X_n$ es una muestra de vaiid de una distribución uniforme en $(0, \theta)$. Veamos cuál es el EMV de θ. En este caso, la función de verosimilitud es

$$L(\theta) = \begin{cases} \dfrac{1}{\theta^n}, & \text{si } \theta \geq X_{(n)}, \\ 0, & \text{si } \theta < X_{(n)}. \end{cases}$$

Esta función de verosimilitud no es derivable (ni siquiera es continua). Si la representamos gráficamente es claro que $\hat{\theta} = X_{(n)}$, el máximo de las observaciones. Observa que en este ejemplo, el método de momentos y el de máxima verosimilitud dan estimadores muy diferentes. ¿Cuál crees que es mejor?

Sobre el cálculo de los EMV

Salvo casos muy sencillos, como los anteriores, normalmente hay que usar algún algoritmo de optimización para calcular los EMV. Puede usarse cualquier algoritmo estándar pero algunos métodos se han diseñado sobre la base de ideas probabilísticas o estadísticas. Tal vez el ejemplo más conocido dentro de estos sea el del algoritmo EM (expectation-maximization) (https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm).

Estimadores bayesianos

Hay dos interpretaciones clásicas de la probabilidad de un suceso: la primera, que podemos llamar **frecuentista**, consiste en interpretar la probabilidad de un suceso como el límite de la frecuencia relativa de veces que ocurre este suceso cuando un experimento aleatorio se va repitiendo más y más veces. La segunda interpretación, la **bayesiana**, consiste en interpretar la probabilidad de un suceso como el grado de creencia subjetiva en que tal suceso ocurra.

Hasta ahora hemos interpretado el parámetro θ como una cantidad fija pero desconocida que tratamos de estimar a partir de los datos $x_1, \ldots, x_n$. Sin embargo, si adoptamos un enfoque bayesiano, tiene sentido describir la incertidumbre sobre el parámetro mediante una distribución de probabilidad definida en el espacio paramétrico Θ, es decir, tratar a θ como si fuera una variable aleatoria en lugar de una constante desconocida. Las probabilidades que asigna esta distribución a los diferentes subconjuntos de Θ reflejan la creencia que tenemos de que θ pertenezca a cada uno de estos subconjuntos.

El método bayesiano opera entonces de la forma siguiente: en un principio, como hemos dicho, tenemos que establecer una distribución sobre Θ que no dependa de los datos del experimento, que sea previa a la observación de la muestra y que refleje la opinión de un experto sobre los valores del parámetro. Esta distribución podría estar basada en otros experimentos o muestras recogidas anteriormente. Llamaremos a esta distribución la **distribución a priori** y denotaremos por $\pi(\theta)$ su función de densidad o probabilidad. La información sobre θ contenida en $\pi(\theta)$ se puede combinar con el modelo estadístico para los datos $f(x;\theta)$ (que en el contexto bayesiano se suele denotar como $f(x\,|\,\theta)$, ya que se interpreta como una distribución condicionada) mediante el teorema de Bayes para calcular la llamada **distribución a posteriori** $\pi(\theta\,|\,x)$, que representa la incertidumbre sobre el parámetro, una vez que los datos ya han sido observados:

$$\pi(\theta\,|\,x) = \frac{f(x\,|\,\theta)\pi(\theta)}{\int_\Theta f(x\,|\,\theta)\pi(\theta)d\theta}$$

Finalmente, se toma como estimador de θ alguna medida numérica de posición que resuma la distribución a posteriori. Lo más habitual es considerar la esperanza, es decir, $\hat\theta = \mathrm{E}(\theta\,|\,x)$. Otra posibilidad podría ser usar la mediana o la moda de la distribución a posteriori.

Ejemplos

Estimación de una proporción con distribución a priori beta. Sea $X_1, \ldots, X_n$ una muestra de vaiid de una distribución $\mathrm{B}(1, \theta)$. Se supone que la distribución a priori de θ es una distribución beta (https://en.wikipedia.org/wiki/Beta_distribution) de parámetros α y β adecuados. A los parámetros de la distribución a priori se les suele llamar **hiperparámetros**. Observa que (despreciando las constantes que no dependen de θ)

$$\pi(\theta\,|\,x) \propto \pi(\theta)f(x\,|\,\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Por lo tanto, $\theta\,|\,x \equiv \mathrm{Beta}(\alpha + n\bar x, n + \beta - n\bar x)$ (¿Por qué?).

Como consecuencia,

$$\hat\theta = \mathrm{E}(\theta\,|\,x) = \frac{\alpha + n\bar x}{\alpha + \beta + n} = \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\bar x = w_n\frac{\alpha}{\alpha+\beta} + (1-w_n)\bar x.$$

El estimador bayesiano $\mathrm{E}(\theta\,|\,x)$ es una media ponderada entre el valor esperado a priori del parámetro $\mathrm{E}(\theta)$ y el estimador de máxima verosimilitud $\bar x$.

Media de una normal con distribución a priori normal (varianza conocida). Sea $X_1, \ldots, X_n$ una muestra de vaiid de una distribución $\mathrm{N}(\mu, \sigma^2)$. Se supone que la distribución a priori de μ es también una distribución normal, $\mathrm{N}(\mu_0, \sigma_0^2)$. Puede comprobarse que la distribución a posteriori es de nuevo normal, $\mathrm{N}(\mu_1, \sigma_1^2)$, donde

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

y

$$\mu_1 = \mu_0 \frac{1/\sigma_0^2}{1/\sigma_0^2 + n/\sigma^2} + \bar{x} \frac{n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}.$$

Sobre el cálculo de los estimadores bayesianos

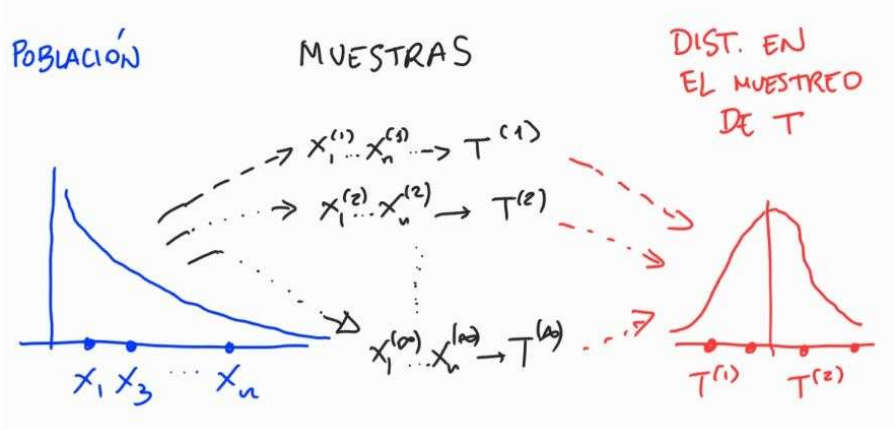
Desde el punto de vista teórico, el enfoque bayesiano es muy simple. Sin embargo, en función de cuál sea la distribución a priori especificada y el modelo estadístico utilizado, los problemas computacionales que presenta el cálculo de la distribución a posteriori y su esperanza pueden ser muy difíciles, 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. Las **familias conjugadas** para un modelo son familias de distribuciones a priori tales que, al combinarse con ese modelo, dan lugar a una distribución a posteriori que pertenece a la misma familia paramétrica. Por ejemplo, la distribución beta es conjugada para el modelo binomial en la estimación de una proporción. También, la distribución normal es conjugada para el modelo normal en la estimación de la media con varianza conocida.

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)** (https://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo)) que permiten extender la aplicación de los métodos bayesianos a modelos muy complejos.

Criterios para valorar un estimador

La calidad de un estimador depende esencialmente de su distribución en el muestreo. Dado que un estimador $\hat{\theta} = T(X_1, \ldots, X_n)$ es función de variables aleatorias, él mismo es también una variable aleatoria cuya distribución se denomina **distribución en el muestreo** del estimador. Esta distribución determina los valores que podemos esperar que tome $\hat{\theta}$ si dispusiéramos de muchas muestras de la misma población. Tiene sentido preguntarnos por el valor esperado de estos valores, $\mathrm{E}(\hat{\theta})$, o su dispersión medida a través de, por ejemplo, $\mathrm{Var}(\hat{\theta})$.



Sesgo y varianza

Una buena propiedad que podemos pedir a un estimador es que no tenga tendencia sistemática a infraestimar o sobreestimar el parámetro. Se dice que un estimador $\hat{\theta}$ es **insesgado** si $\mathrm{E}(\hat{\theta}) = \theta$, para todo $\theta \in \Theta$.

Si el estimador es insesgado, su valor esperado coincide con el parámetro **para cualquier valor de este**. En el caso de que esto no ocurra el **sesgo** se define como $\mathrm{Sesgo}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$. Si el sesgo es positivo, hay una tendencia sistemática a sobreestimar el parámetro, y lo contrario si es negativo.

Por otra parte, otra buena propiedad que debe tener un estimador es no dar resultados muy diferentes para las distintas posibles muestras. No queremos que al replicar el mismo experimento muchas veces, los resultados sean muy distintos entre sí. Matemáticamente, esto significa que queremos que la varianza del estimador, $\mathrm{Var}(\hat{\theta})$, sea lo menor posible.

Es muy habitual en muchos procedimientos estadísticos que el sesgo y la varianza sean objetivos contrapuestos de manera que al reducirse el primero aumenta la segunda y viceversa. Normalmente, los métodos dan buenos resultados si el sesgo y la varianza están equilibrados adecuadamente.

Una cantidad que tiene en cuenta tanto el sesgo como la varianza simultáneamente es el **error cuadrático medio** del estimador:

$$\mathrm{ECM}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2]$$

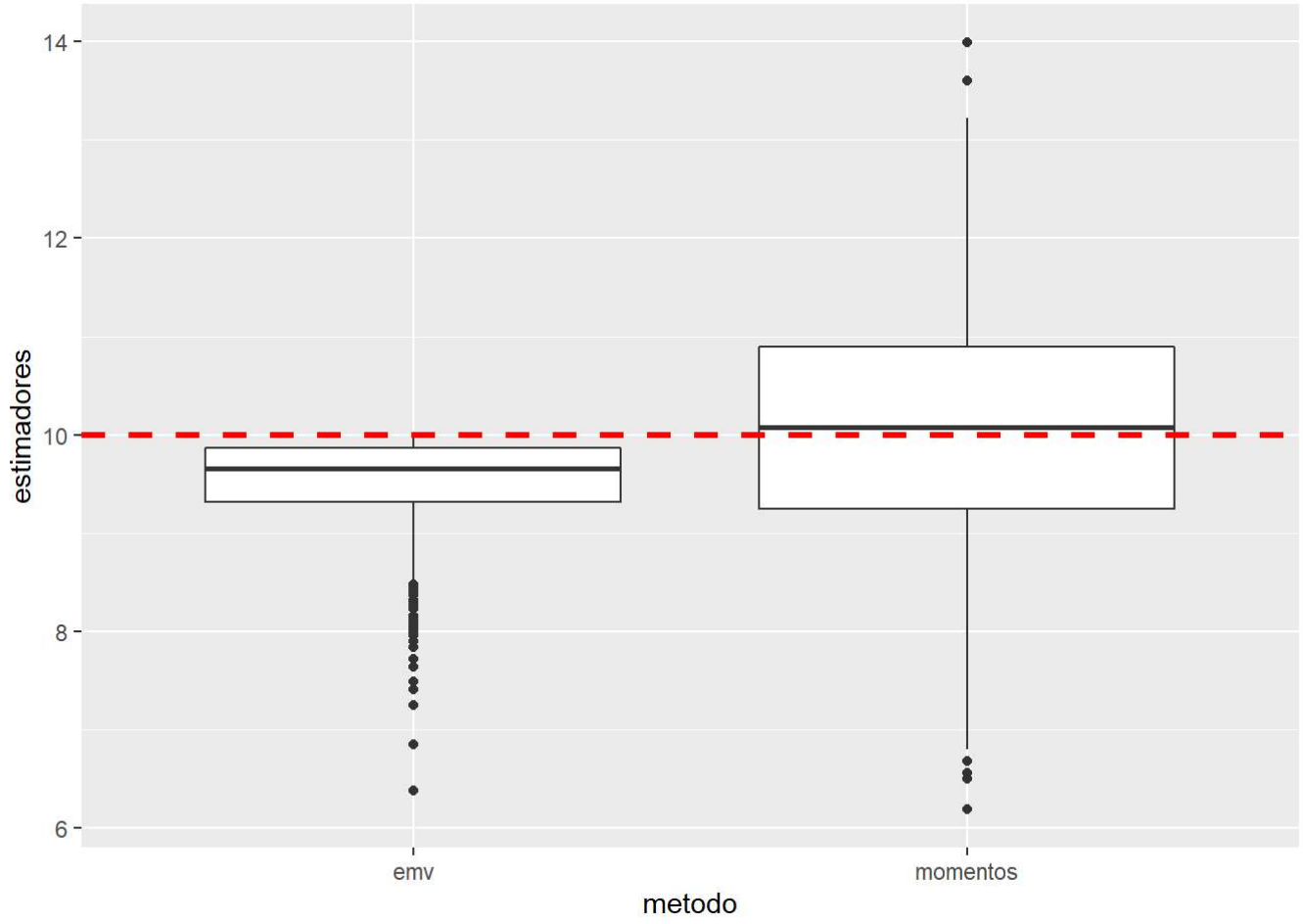
Es muy fácil ver que el ECM es igual al sesgo al cuadrado más la varianza:

$$\mathrm{ECM}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \mathrm{E}(\hat{\theta}) + \mathrm{E}(\hat{\theta}) - \theta)^2] = \mathrm{Var}(\hat{\theta}) + \mathrm{Sesgo}(\hat{\theta})^2 + 2\mathrm{E}[(\hat{\theta} - \mathrm{E}(\hat{\theta}))(\mathrm{E}(\hat{\theta}) - \theta)],$$

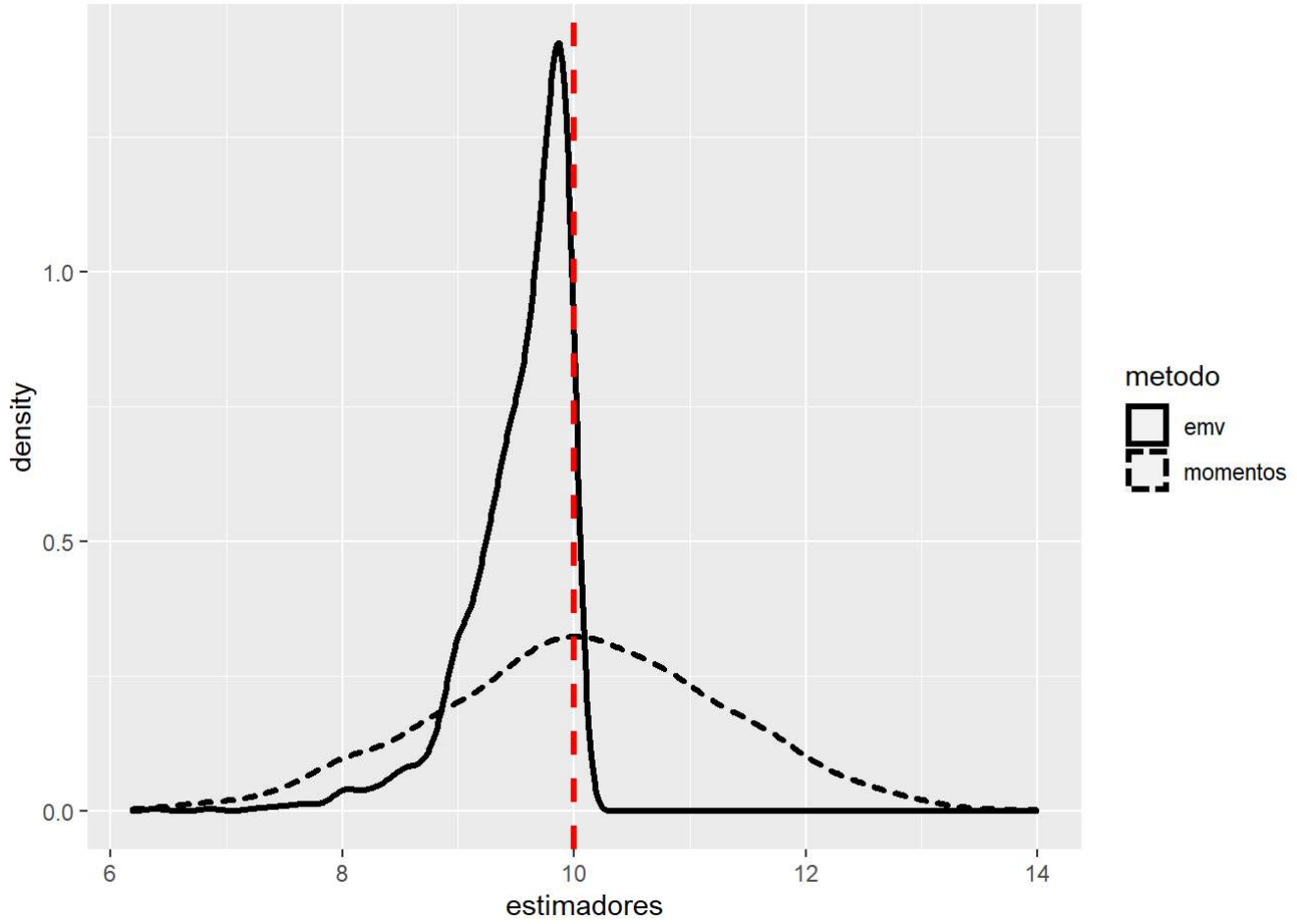
pero el último término se anula (¿por qué?).

Ejemplo: comparación del EMV y del estimador momentos en $U(0, \theta)$

```
# Parámetros ------------------------------------------------------------

theta <- 10  # valor verdadero del parámetro
n <- 20   # tamaño muestral
m <- 1000 # número de muestras

# Genera los datos ------------------------------------------------------


set.seed(1234)   # para reproducir los resultados
muestras <- matrix(runif(n*m, 0, theta), n)

# Calcula estimadores ---------------------------------------------------

emv <- apply(muestras, 2, max)
momentos <- 2*apply(muestras, 2, mean)
metodo <- gl(2, m, labels = c('emv', 'momentos'))
df <- data.frame(estimadores = c(emv, momentos), metodo = metodo)

# Gráficos --------------------------------------------------------------

ggplot(df) +
  geom_boxplot(aes(x = metodo, y = estimadores)) +
  geom_hline(yintercept = theta, col = 'red', size = 1.1, linetype = 2)
```




```
ggplot(df) +
  geom_density(aes(x = estimadores, linetype = metodo), size = 1.1) +
  geom_vline(xintercept = theta, col = 'red', size = 1.1, linetype = 2)
```



Otros criterios: consistencia, normalidad asintótica,…

Los criterios asintóticos para evaluar un estimador se refieren a su comportamiento límite a medida que disponemos de más y más datos.

Se dice que un estimador es **consistente** si su valor converge al del parámetro al aumentar el tamaño muestral. Dado que $\hat{\theta}$ es una sucesión de v.a. debemos considerar algún tipo de convergencia estocástica (véase la sección siguiente). En función de qué tipo de convergencia se utilice la consistencia puede ser fuerte o débil. La consistencia es una propiedad que debería tener cualquier estimador razonable.

Otra buena propiedad que puede tener un estimador es la **normalidad asintótica**, es decir, que la distribución límite del estimador (o de alguna transformación del estimador) sea aproximadamente normal para muestras grandes. Veremos algunos ejemplos también en la sección siguiente. Esta propiedad es útil si queremos calcular intervalos de confianza o llevar a cabo contrastes de hipótesis sobre el parámetro.

Hay otros criterios para valorar la bondad de una estimación. A veces se usan por ejemplo, criterios de robustez. Un estimador es **robusto** si no se ve muy afectado por la presencia de datos atípicos en la muestra. Hay diversas maneras de formalizar matemáticamente esta propiedad.

Convergencias estocásticas

Convergencia casi segura

Sea X_n una sucesión de variables aleatorias. Se dice que X_n **converge casi seguro** a otra variable aleatoria X y se denota $X_n \to_{c.s.} X$ si

$$\mathrm{P}\{\lim_{n\to\infty} X_n = X\} = 1.$$

Con probabilidad uno, la sucesión de los valores X_n converge a X. En general, este tipo de convergencia es muy fuerte. En estadística suele ser suficiente para muchos propósitos considerar modos de convergencia menos exigentes.

Dado un estimador $\hat{\theta}$, se dice que es **fuertemente consistente** para θ si $\hat{\theta} \to_{\text{c.s.}} \theta$.

Convergencia en probabilidad

Sea X_n una sucesión de variables aleatorias. Se dice que X_n **converge en probabilidad** a otra variable aleatoria X y se denota $X_n \to_p X$ si, para todo $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\{|X_n - X| > \epsilon\} = 0.$$

Esto significa que, si n es grande, con probabilidad cercana a uno la diferencia entre X_n y X es inferior a cualquier margen de error predeterminado. Con mucha frecuencia en estadística, la variable límite de la sucesión anterior es degenerada, lo que significa que $\mathrm{P}(X = \theta) = 1$, para cierto $\theta \in \mathrm{R}$. En este caso, escribimos $X_n \to_p \theta$.

Dado un estimador $\hat{\theta}$, se dice que es **débilmente consistente** para θ si $\hat{\theta} \to_p \theta$.

Proposición. Si tanto el sesgo como la varianza de un estimador convergen a cero cuando $n \to \infty$, entonces el estimador es débilmente consistente.

Demostración. Es una consecuencia muy simple de la desigualdad de Markov. Dado $\epsilon > 0$,

$$\mathrm{P}(|\hat{\theta} - \theta| \geq \epsilon) = \mathrm{P}(|\hat{\theta} - \theta|^2 \geq \epsilon^2) \leq \frac{\mathrm{E}[(\hat{\theta} - \theta)^2]}{\epsilon^2} = \frac{\mathrm{Var}(\hat{\theta})}{\epsilon^2} + \frac{\mathrm{Sesgo}(\hat{\theta})^2}{\epsilon^2} \to 0.$$

Convergencia en distribución

El siguiente concepto de convergencia es el más importante en estadística y se refiere a la distribución de las variables, en lugar de a las variables en sí.

Sea X_n una sucesión de variables aleatorias con funciones de distribución F_n. Se dice que X_n **converge en distribución** a otra variable aleatoria X con función de distribución F y se denota $X_n \to_d X$ (o también, más propiamente, $F_n \to_d F$) si, para todo $x \in \mathrm{Cont}(F)$, se verifica $\lim_{n\to\infty} F_n(x) = F(x)$, donde $\mathrm{Cont}(F)$ es el conjunto de puntos en los que F es continua.

En general, se puede demostrar que $X_n \to_p X$ implica $X_n \to_d X$, pero que la implicación recíproca no es cierta en general. No obstante, si la distribución límite es degenerada, ambas convergencias son equivalentes, es decir, $X_n \to_p \theta \Leftrightarrow X_n \to_d \theta$.

Un resultado importante en relación con la convergencia en distribución es el *teorema de la aplicación continua*

Teorema de la aplicación continua. Sea X_n una sucesión de v.a. tal que $X_n \to_d X$ y sea $g\colon \mathrm{R} \to \mathrm{R}$ una función continua. Entonces, $g(X_n) \to_d g(X)$.

Ley de los grandes números y teorema central del límite

Ley de los grandes números

Bajo condiciones de regularidad, las leyes de los grandes números permiten establecer la convergencia de promedios de variables aleatorias. El ejemplo más sencillo es el siguiente:

Ley débil de los grandes números (LDGN). Sea X_n una sucesión de v.a.i.i.d. con media μ. Entonces,

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \to_p \mu.$$

La demostración cuando se supone $\mathrm{Var}(X_i) = \sigma^2 < \infty$ se reduce a una aplicación elemental de la desigualdad de Chebychev:

$$\mathrm{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.$$

En el caso de varianza infinita el resultado sigue siendo cierto pero la demostración ya no es elemental.

Teorema central del límite

El teorema central del límite nos da información acerca de la distribución aproximada de la diferencia entre la media muestral y la media poblacional $|\bar{X}_n - \mu|$, si el tamaño muestral n es suficientemente grande.

Teorema central del límite (TCL). Sea X_n una sucesión de v.a.i.i.d. con media μ y varianza σ^2. Entonces,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to_d \mathrm{N}(0, 1).$$

Es fácil ver que la conclusión del TCL es equivalente a escribir

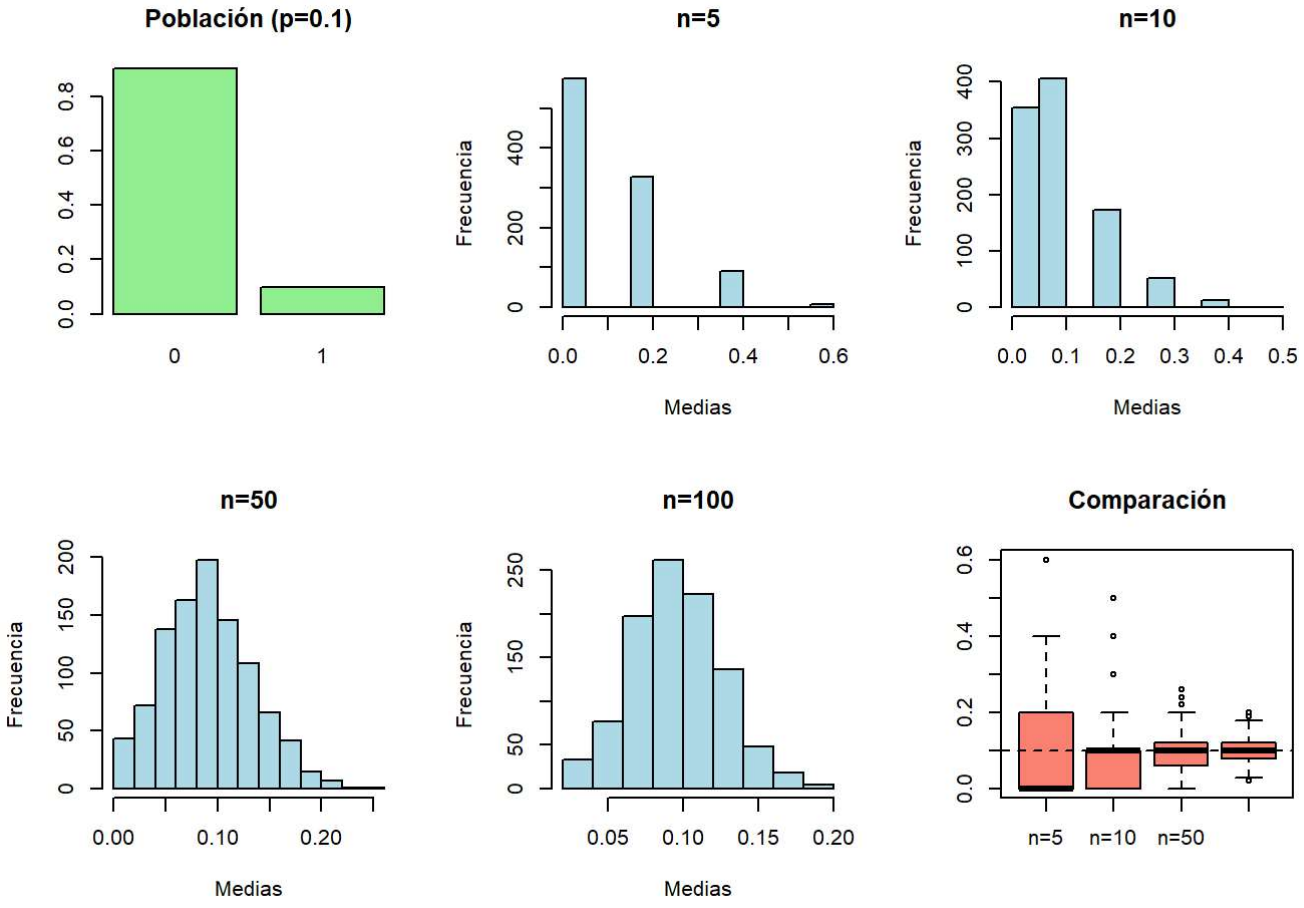
$$\sqrt{n}(\bar{X}_n - \mu) \to_d \mathrm{N}(0, \sigma^2).$$

La siguiente aproximación también se deriva del TCL:

$$\bar{X}_n - \mu \cong \mathrm{N}\left(0, \frac{\sigma^2}{n}\right).$$

Por la LDGN, $\bar{X}_n - \mu \to_p 0$. Al multiplicar por $\sqrt{n}$, resulta que el límite ya no es cero ni tampoco infinito, sino una distribución no degenerada (normal). En este sentido podemos decir que la velocidad con la que $\bar{X}_n$ converge a μ es la misma con la que $1/\sqrt{n}$ va a cero. También a veces se dice que la convergencia es “a tasa $\sqrt{n}$”. Esta es la velocidad de convergencia usual en estimación paramétrica.

En la siguiente figura se calculan los promedios de 1000 muestras de tamaño n de una distribución de Bernoulli para distintos valores de n:



Dos resultados útiles

El lema de Slutsky y el método delta se utilizan para combinar la LDGN y el TCL con el fin de estudiar propiedades asintóticas de estadísticos un poco más complicados.

Lema de Slutsky. Sean X_n e Y_n dos sucesiones de v.a. tales que $X_n \to_d X$ e $Y_n \to_d \theta$, donde $\theta \in \mathrm{R}$. Sea $g\colon \mathrm{R}^2 \to \mathrm{R}$ una función continua. Entonces, $g(X_n, Y_n) \to_d g(X, \theta)$.

La aplicación habitual de este lema permite conocer el límite de sucesiones de sumas $X_n + Y_n$, productos $X_n Y_n$ o cocientes X_n / Y_n.

Para determinar el comportamiento asintótico de funciones suaves de sucesiones cuyo límite es conocido se usa el llamado *método delta*:

Proposición (método delta). Sean X_n una sucesión de v.a. tal que $n^b(X_n - \theta) \to_d X$ para $b > 0$ y $\theta \in \mathrm{R}$. Sea $g\colon \mathrm{R} \to \mathrm{R}$ una función derivable con derivada continua. Entonces,

$$n^b[g(X_n) - g(\theta)] \to_d g'(\theta)X$$

Algunas aplicaciones

1. Supongamos que $X_1, X_2, \ldots$ son v.a.i.i.d. con distribución uniforme en el intervalo $(0, \theta)$. Determina el límite en distribución de $2\bar{X}_n$ y $\sqrt{n}(2\bar{X}_n - \theta)$.
2. Supongamos que $X_1, X_2, \ldots$ son v.a.i.i.d. con distribución $\mathrm{B}(1, p)$. Determina el comportamiento asintótico de:

- La proporción muestral: $\hat{p} = (X_1 + \cdots + X_n)/n$.
- La proporción muestral estandarizada: $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$.

- La proporción muestral estandarizada, pero usando en el denominador $\hat{p}$ en lugar de p: $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$.

3. Supongamos que $X_1, X_2, \ldots$ son v.a.i.i.d. con media μ, varianza σ^2 y $\mathrm{E}(X_i^4) < \infty$. Entonces, $S_n^2 \to_p \sigma^2$ y $\sqrt{n}(S_n^2 - \sigma^2) \to_d \mathrm{N}(0, \sigma^4(\kappa - 1))$, donde $\kappa = \mathrm{E}[(X_i - \mu)^4]/\sigma^4$ es el llamado **coeficiente de curtosis**.

Sabemos que la varianza muestral es invariante por traslaciones, por lo tanto si definimos $H_i = X_i - \mu$, tenemos que

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(H_i - \bar{H})^2 = \frac{n}{n-1}\left(\frac{\sum_{i=1}^n H_i^2}{n} - \bar{H}^2\right) = \frac{n}{n-1}(\bar{U} - \bar{H}^2),$$

donde $U_i = H_i^2 = (X_i - \mu)^2$. Si reordenamos adecuadamente los términos,

$$\sqrt{n}(S^2 - \sigma^2) = \frac{n}{n-1}\sqrt{n}(\bar{U} - \sigma^2) + \frac{\sqrt{n}}{n-1}\sigma^2 - \frac{n}{n-1}\sqrt{n}\bar{H}^2.$$

Por el TCL, el primer término converge en distribución a $\mathrm{N}(0, \sigma^4(\kappa - 1))$, mientras que los dos términos restantes convergen en distribución a cero. Por el lema de Slutsky, se tiene el resultado.

4. Supongamos que $X_1, X_2, \ldots$ son v.a.i.i.d. con distribución de Poisson de parámetro λ. ¿Cuál es la distribución asintótica de $\sqrt{n}(\bar{X}_n - \lambda)$? ¿Cuál es la distribución asintótica de $\sqrt{n}(2\sqrt{\bar{X}_n} - 2\sqrt{\lambda})$? (Esto es un ejemplo de lo que se conoce como **transformación estabilizadora de la varianza**)