ESTADÍSTICA I
Tema 2: Muestreo aleatorio

- Diferencia entre probabilidad e inferencia estadística
- Conceptos probabilísticos básicos
- Muestra aleatoria
- El problema de inferencia
- Estadísticos. Media y varianza muestrales. Estadísticos de orden
- Ley de los grandes números
- Función de distribución empírica. Teorema de Glivenko-Cantelli
- Histogramas y estimadores kernel. Moda muestral

Diferencia entre probabilidad e inferencia estadística

En Probabilidad I resolvisteis problemas como éste:

> En un ensayo clínico, se administra un medicamento a 200 pacientes. Se sabe que el medicamento es efectivo con probabilidad 0.75. ¿Cuál es la probabilidad de que mejoren más de 155 pacientes?

La información previa para responder a la pregunta es "*el medicamento es efectivo con probabilidad 0.75*". En la práctica no es realista pensar que se pueda disponer de tal información. Ese dato sólo se conocería tras administrar el medicamento a **todos** los individuos enfermos de una población y comprobar que se ha producido una mejoría en el 75% de los casos.

Estamos suponiendo totalmente conocida la distribución de la v.a. $Y =$ "número de pacientes entre los 200 para los que el medicamento es efectivo".

Saber que el medicamento es efectivo en el 75% de los casos equivale a afirmar que la distribución de Y es una binomial $\mathrm{B}(n = 200, p = 0.75)$.

En este caso, podríamos calcular la probabilidad de que mejoren más de 155 pacientes de los 200 a los que se ha administrado el medicamento. La respuesta del problema se puede obtener de la forma siguiente usando el comando `pbinom` de R (que calcula los valores de la función de la distribución binomial):

```
1 - pbinom(155, 200, 0.75)
[1] 0.1852385
```

o, equivalentemente,

```
pbinom(155, 200, 0.75, lower.tail=F)
```

La situación inversa es más realista. Normalmente nos interesa conocer el porcentaje p de casos en los que un medicamento es efectivo a partir de la información obtenida al administrarlo a un subconjunto de la población de enfermos (la muestra).

Este es el tipo de problemas que nos plantearemos en Estadística I:

> En un ensayo clínico, se administra un medicamento a 200 pacientes y se observa que mejoran 150 pacientes. ¿Hay evidencia estadística para afirmar que el medicamento es efectivo en un porcentaje de casos superior al 75%?

Puesto en términos matematicos, si

$$X_i = \begin{cases} 1 & \text{si el medicamento es efectivo en el paciente} \\ 0 & \text{si no lo es,} \end{cases}$$

tenemos un conjunto de 200 v.a. $X_1, \ldots, X_{200}$ independientes e idénticamente distribuidas (si la población es grande) con distribución B(1,p).

Se suelen usar letras mayúsculas para denotar las v.a. y letras minúsculas para las 200 observaciones o realizaciones obtenidas tras realizar el experimento (ceros o unos) $x_1, \ldots, x_{200}$.

Sabiendo que $\sum_{i=1}^{200} x_i = 150$, ¿qué podemos decir sobre p? ¿Podemos afirmar que $p > 0.75$? Si hacemos una afirmación como ésa, ¿cuál es el riesgo de equivocarnos?

En este caso, a partir de la información sobre 200 casos particulares, tenemos que obtener información general sobre p, un parámetro que afecta a toda la población.

Esta asignatura se dedica a estudiar estos y otros problemas similares. Tendréis que utilizar toda la información que habéis adquirido en probabilidad, para recorrer de nuevo el camino en la dirección contraria.

Conceptos probabilísticos básicos

Un espacio de probabilidad es un triplete $(\Omega, \mathcal{A}, \mathbb{P})$ donde

- Ω es un conjunto no vacío
- $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ es una σ-álgebra, es decir,
  - $\Omega \in \mathcal{A}$.
  - Si $A \in \mathcal{A}$, entonces $A^c \in \mathcal{A}$.
  - Si $\{A_i\}_{i=1}^{\infty} \subset \mathcal{A}$, entonces $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.
- $\mathbb{P}$ es una medida de probabilidad sobre $\mathcal{A}$, es decir,

$$\begin{array}{rccc} \mathbb{P}: & \mathcal{A} & \to & [0,1] \\ & A & \mapsto & \mathbb{P}(A) \end{array}$$

satisfaciendo

  - $\mathbb{P}(\Omega) = 1$
  - Si $\{A_i\}_{i=1}^{\infty} \subset \mathcal{A}$ con $A_i \cap A_j = \varnothing$ para $i \neq j$, entonces $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

La σ-álgebra más habitual en $\mathbb{R}$ es la de Borel $\mathcal{B}$, que se puede generar con los intervalos.

Una variable aleatoria es una aplicación medible

$$\begin{array}{rccc} X: & (\Omega, \mathcal{A}, \mathbb{P}) & \to & (\mathbb{R}, \mathcal{B}) \\ & \omega & \mapsto & X(\omega) \end{array}$$

es decir, para cualquier $B \in \mathcal{B}$, se cumple que

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} = \{X \in B\} \in \mathcal{A}.$$

Entonces $\mathbb{P}$ y X inducen una medida de probabilidad P_X en $(\mathbb{R}, \mathcal{B})$ llamada distribución de probabilidad de la v.a. X

$$P_X(B) = \mathbb{P}\{X \in B\}.$$

El espacio muestral de X es el subconjunto de $\mathbb{R}$ que contiene todos los posibles valores de X.

La función de distribución de la v.a. X es la aplicación

$$\begin{array}{rccl} F: & \mathbb{R} & \to & [0,1] \\ & x & \mapsto & F(x) = \mathbb{P}\{X \leq x\} = P_X(-\infty, x]. \end{array}$$

Es una función no decreciente y continua por la derecha.

La función característica de la v.a. X es

$$\varphi(t) = \mathbb{E}(e^{itX}) = \int_{\mathbb{R}} e^{itx} dP_X(x).$$

Una v.a. X es discreta cuando existe un conjunto finito o numerable $S = \{a_i\} \subset \mathbb{R}$ tal que

$$1 = P(S) = \sum_i \mathbb{P}\{X = a_i\} = \sum_i (F(a_i) - F(a_i-)).$$

La distribución de X es (absolutamente) continua cuando existe una función de densidad f tal que

$$\mathbb{P}\{X \in B\} = \int_B f(t)dt, \ \forall B \in \mathcal{B},$$

o, de manera equivalente,

$$F(x) = \int_{-\infty}^{x} f(t)dt, \ \forall x \in \mathbb{R}.$$

Entonces, $F'(t) = f(t)$.

La densidad f debe satisfacer

- $f(t) \geq 0$ para todo t;
- $\int_{\mathbb{R}} f(t)dt = 1$.

Definimos la media poblacional o esperanza de X como

$$\mu = \mathbb{E}(X) = \int_\Omega X d\mathbb{P} = \int_{\mathbb{R}} x \, dP_X(x) = \int_{\mathbb{R}} x dF(x),$$

supuesto que esta integral es finita.

Teorema de cambio de espacio de integración: Si g es una función real medible tal que $\mathbb{E}(g(X))$ es finita, entonces

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) \, dP_X(x).$$

En particular,

$$\sigma^2 = \mathbb{V}(X) = \int_{\mathbb{R}} (x - \mu)^2 dP_X(x).$$

Si $\mathbb{E}|g(X)| = \infty$, entonces decimos que $\mathbb{E}g(X)$ no existe.

El momento de orden k de la v.a. X respecto al origen es $\mathbb{E}(X^k)$.

El momento de orden k de X respecto a la media es $\mathbb{E}\left((X-\mu)^k\right)$.

Principales distribuciones discretas y continuas:
Ver enlace en la web de la asignatura.

Desigualdades básicas:

Desigualdad de Markov: Sea X v.a. no negativa. Entonces, para todo $\epsilon > 0$, $\mathbb{P}\{X > \epsilon\} \leq \dfrac{\mu}{\epsilon}$.

Desigualdad de Chebyshev: $\mathbb{P}\{|X-\mu| > \epsilon\} \leq \dfrac{\sigma^2}{\epsilon^2}$.

Convergencias estocásticas

Sean $X, X_n : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ v.a., $n \in \mathbb{N}$.

¿Qué significa $X_n \xrightarrow[n\to\infty]{} X$?

- **Convergencia en probabilidad**

Decimos que $\{X_n\}_{n\in\mathbb{N}}$ converge a X en probabilidad y lo denotamos $X_n \xrightarrow[n\to\infty]{P} X$ si, para todo $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| \geq \epsilon\} = 0$$

o equivalentemente

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| < \epsilon\} = 1.$$

En Análisis este tipo de convergencia se llama *convergencia en medida*.

• Convergencia casi segura

Decimos que $\{X_n\}_{n\in\mathbb{N}}$ converge a X casi seguro (o con probabilidad uno o en casi todo punto) y lo denotamos $X_n \xrightarrow[n\to\infty]{\text{c.s.}} X$ si

$$\mathbb{P}\{\omega \in \Omega : X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)\} = 0$$

o equivalentemente si, para todo $\epsilon > 0$,

$$\mathbb{P}\{\lim_{n\to\infty} |X_n - X| < \epsilon\} = 1.$$

Ejemplo de convergencia en probabilidad pero no c.s.

Consideramos una sucesión de v.a. construidas del siguiente modo. Primero definimos una v.a. U uniforme en el intervalo [0,1] y luego le aplicamos ciertas funciones indicatrices $\mathbb{1}_{A_n^i}$, siendo $A_n^i = \left[\frac{i-1}{n}, \frac{i}{n}\right]$, $i = 1, \ldots, n$, $n \geq 1$:

$$(\Omega, \mathcal{A}, \mathbb{P}) \xrightarrow{U} (\mathbb{R}, \mathcal{B}) \xrightarrow{\mathbb{1}_{A_n^i}} (\mathbb{R}, \mathcal{B}).$$

Para un $\omega \in \Omega$ fijo, $U(\omega)$ es una observación concreta extraída de la distribución uniforme en [0,1].

La sucesión de v.a.

$$\begin{aligned} &X_1^1 = \mathbb{1}_{A_1^1}(U), \quad X_2^1 = \mathbb{1}_{A_2^1}(U), \quad X_2^2 = \mathbb{1}_{A_2^2}(U), \\ &X_3^1 = \mathbb{1}_{A_3^1}(U), \quad X_3^2 = \mathbb{1}_{A_3^2}(U), \quad X_3^3 = \mathbb{1}_{A_3^3}(U), \ldots \end{aligned}$$

converge a 0 en probabilidad pero no c.s.

• Convergencia débil o en distribución

Sean F y F_n las funciones de distribución de X y X_n respectivamente. Decimos que $\{X_n\}_{n\in\mathbb{N}}$ converge a X débilmente o en distribución y lo denotamos $X_n \xrightarrow[n\to\infty]{\text{d}} X$ si

$$\lim_{n\to\infty} F_n(x) = F(x)$$

para todo $x \in \mathbb{R}$ en el que F sea continua.

Sean ϕ y ϕ_n las funciones características de X y X_n respectivamente. Se cumple que

$$X_n \xrightarrow[n\to\infty]{\text{d}} X \Leftrightarrow \phi_n(t) \xrightarrow[n\to\infty]{} \phi(t), \quad \forall t \in \mathbb{R}.$$

También se cumple que

$$X_n \xrightarrow[n\to\infty]{\text{d}} X \Leftrightarrow \mathbb{E}(g(X_n)) \xrightarrow[n\to\infty]{} \mathbb{E}(g(X))$$

para toda $g : \mathbb{R} \to \mathbb{R}$ continua y acotada.

Se satisfacen las siguientes implicaciones:

$$X_n \xrightarrow[n\to\infty]{\text{c.s.}} X \quad \Rightarrow \quad X_n \xrightarrow[n\to\infty]{P} X$$

$$X_n \xrightarrow[n\to\infty]{P} X \quad \Rightarrow \quad X_n \xrightarrow[n\to\infty]{\text{d}} X$$

$$X_n \xrightarrow[n\to\infty]{P} c, \text{ con } c \text{ constante} \quad \Leftrightarrow \quad X_n \xrightarrow[n\to\infty]{\text{d}} c$$

Teorema de Slutsky: Sean $\{X_n\}_n$ e $\{Y_n\}_n$ sucesiones de v.a. y X una v.a. Si $X_n \xrightarrow[n\to\infty]{\text{d}} X$ e $Y_n \xrightarrow[n\to\infty]{P} c$, siendo $c \in \mathbb{R}$ una constante, entonces

(i) $X_n + Y_n \xrightarrow[n\to\infty]{\text{d}} X + c$

(ii) $X_n\, Y_n \xrightarrow[n\to\infty]{\text{d}} c\, X$

Teorema de la aplicación continua: Sea $\{X_n\}_n$ una sucesión de v.a. tal que $X_n \xrightarrow[n\to\infty]{\text{d}} X$ y sea $g : \mathbb{R} \to \mathbb{R}$ una función continua. Entonces, $g(X_n) \xrightarrow[n\to\infty]{\text{d}} g(X)$.

Muestra

Se supone que los datos $x_1, \ldots, x_n$ se obtienen mediante observaciones reiteradas e independientes de una v.a. X. Se dice entonces que los datos constituyen una muestra (observada) de X.

La muestra observada es una "realización" de una muestra aleatoria $X_1, \ldots, X_n$ de X.

Desde el punto de vista probabilístico, la muestra (aleatoria) está constituida por n variables aleatorias $X_1, \ldots, X_n$ independientes e idénticamente distribuidas (i.i.d.).

Se dice a veces, en terminología estadística informal (pero muy habitual) que la muestra se extrae de una población, descrita por la v.a. X, y se adjetivan como poblacionales a las características de interés de la distribución de X (por ejemplo, $\mu = \mathbb{E}(X)$, $\sigma^2 = \mathbb{E}(X-\mu)^2$ ó $\mathbb{E}(X^3)$).

Estadísticos

Cuando extraemos una muestra $X_1, \ldots, X_n$ de X se suelen calcular algunas medidas resumen. Cualquiera de ellas se puede expresar matemáticamente como una función $T = T(x_1, \ldots, x_n)$ de la muestra $X_1, \ldots, X_n$.

Dada una función medible T, la v.a. $T = T(X_1, \ldots, X_n)$ se denomina estadístico. La definición de estadístico es muy amplia.

Como la distribución de probabilidad de T se calcula a partir de la distribución de las variables X_i que constituyen la muestra, la denominaremos distribución de T en el muestreo (*sampling distribution*). Obviamente la distribución de $T(X_1, \ldots, X_n)$ depende de la distribución de X y de la expresión matemática de la función $T = T(x_1, \ldots, x_n)$.

A veces, si la situación es suficientemente simple, se puede calcular analíticamente la distribución en el muestreo de un estadístico. Una herramienta útil para ello es la función característica.

Si no, podemos determinar la distribución en el muestreo de un estadístico, tal vez podamos aproximarla asintóticamente (cuando $n \to \infty$) utilizando algún resultado de convergencia de sucesiones de v.a.'s.

Si esto tampoco es posible, entonces podemos aproximar la distribución en el muestreo de T mediante el método de Montecarlo (simulaciones intensivas de los valores de T).

Ejemplo: Supongamos, por ejemplo, que la distribución del tiempo de espera en la caja de un supermercado es exponencial de parámetro λ. ¿Cuál es la la distribución del promedio de los tiempos de espera de n clientes? Se supone que los clientes se han seleccionado en días diferentes de manera que sus tiempo de espera se pueden suponer independientes.

Si $X \sim \exp(\lambda)$, entonces su función característica es

$$\varphi_X(t) = \left(1 - \frac{it}{\lambda}\right)^{-1}.$$

La función característica del promedio es

$$\varphi_{\bar{X}}(t) = \mathbb{E}\left(e^{it\bar{X}}\right) = \mathbb{E}\left(e^{it(X_1+\ldots+X_n)/n}\right) = \left[\varphi_X\left(\frac{t}{n}\right)\right]^n = \left(1 - \frac{it}{n\lambda}\right)^{-n},$$

la de una distribución gamma de parámetros $\alpha = n$ y $\beta = n\lambda$.

La densidad de la gamma(α, β) es

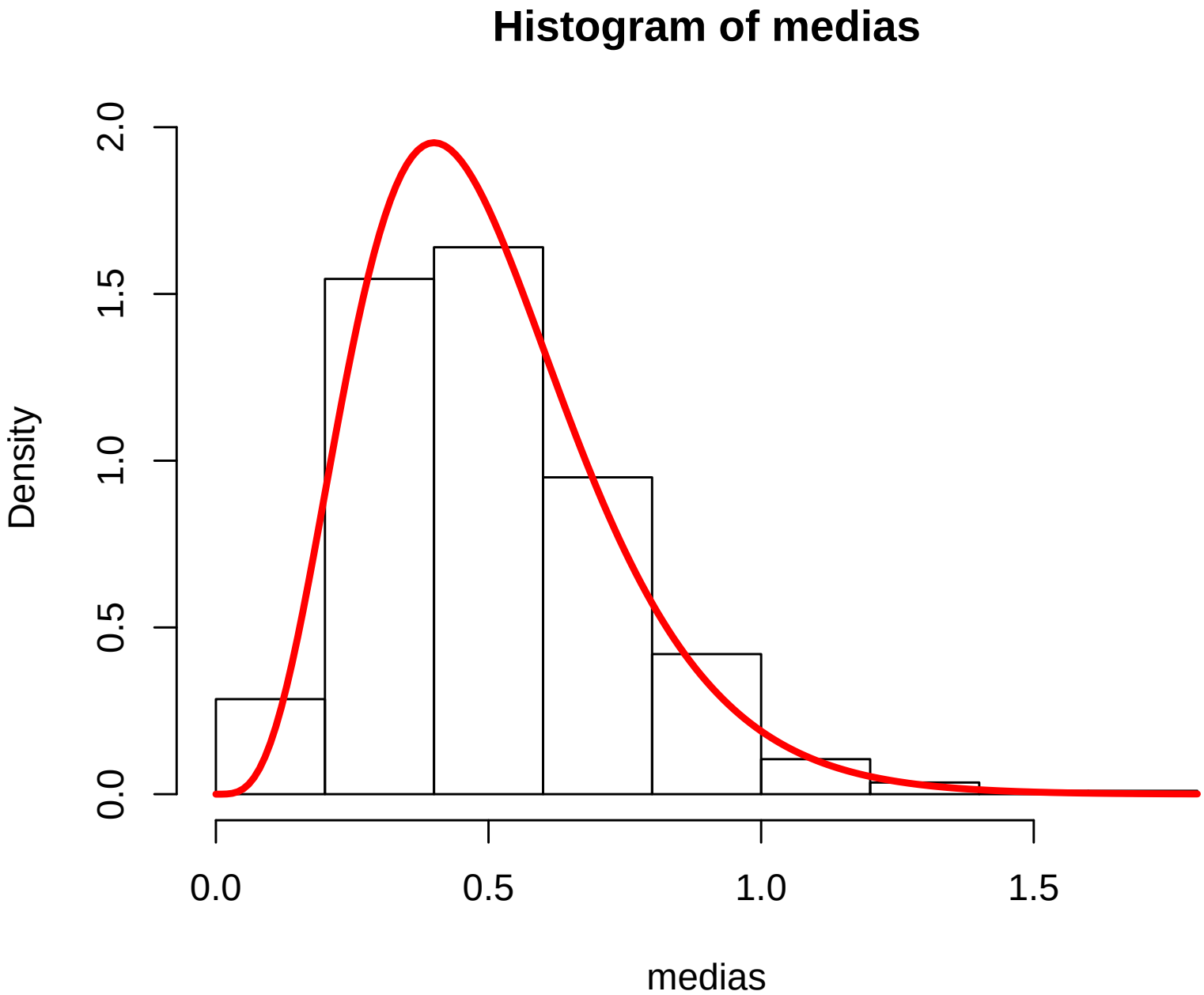
$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

donde

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx.$$

Comprobemos este resultado empíricamente mediante simulación:

```
# Comprobacion de la distribucion en el muestreo
# de la media muestral de n exponenciales:
lambda = 2 # parametro de la exponencial
n = 5 # tamanio muestral
B = 1000 # numero de muestras Montecarlo
Bmuestras = matrix(rexp(B*n,rate=lambda),nrow=n)
medias = colMeans(Bmuestras)
H = hist(medias,plot=F)
x = seq(0,max(H$breaks),0.01)
f = dgamma(x,shape=n,scale=1/(n*lambda))

M = max(max(f),max(H$density))

hist(medias,freq=F,ylim=c(0,M))
lines(x,f,col="red")
```


Histogram of medias
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
medias

El error estándar o error típico de un estadístico T es la desviación típica de su distribución en el muestreo:

$$\sqrt{\mathbb{V}(T(X_1, \ldots, X_n))}. \tag{1}$$

Como a menudo (**??**) depende de alguna cantidad desconocida, también se denomina error típico a una estimación de (**??**).

Ejemplo: Si $X_1 \ldots, X_n$ es una muestra de $X \sim N(\mu, \sigma)$, entonces

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \text{Error típico de } \bar{X} =$$

Planteamiento general del problema de inferencia

Las características de la v.a. X que genera los datos (por ejemplo, los momentos, los cuantiles, la distribución, etc.) se denominan momentos, cuantiles, etc. poblacionales.

En general, uno de los objetivos principales de la inferencia estadística es estimar o "aproximar" las características poblacionales a partir de la información proporcionada por la muestra.

Otras técnicas estadísticas no van orientadas directamente a aproximar el valor de una característica de interés (como por ejemplo la media), sino más bien a decidir entre dos posibles opciones acerca de ella (por ejemplo, si es mayor o menor que 1). La correspondiente metodología se denomina contraste de hipótesis.

Veamos ejemplos de algunos problemas de estimación.

La función de distribución empírica

La función de distribución empírica asociada a la muestra $X_1, \ldots, X_n$ se define mediante

$$\mathbb{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, t]}(X_i)$$

Ésta es la función de distribución que corresponde a una medida de probabilidad discreta que asigna masa $1/n$ a cada uno de los valores $X_1, \ldots, X_n$.

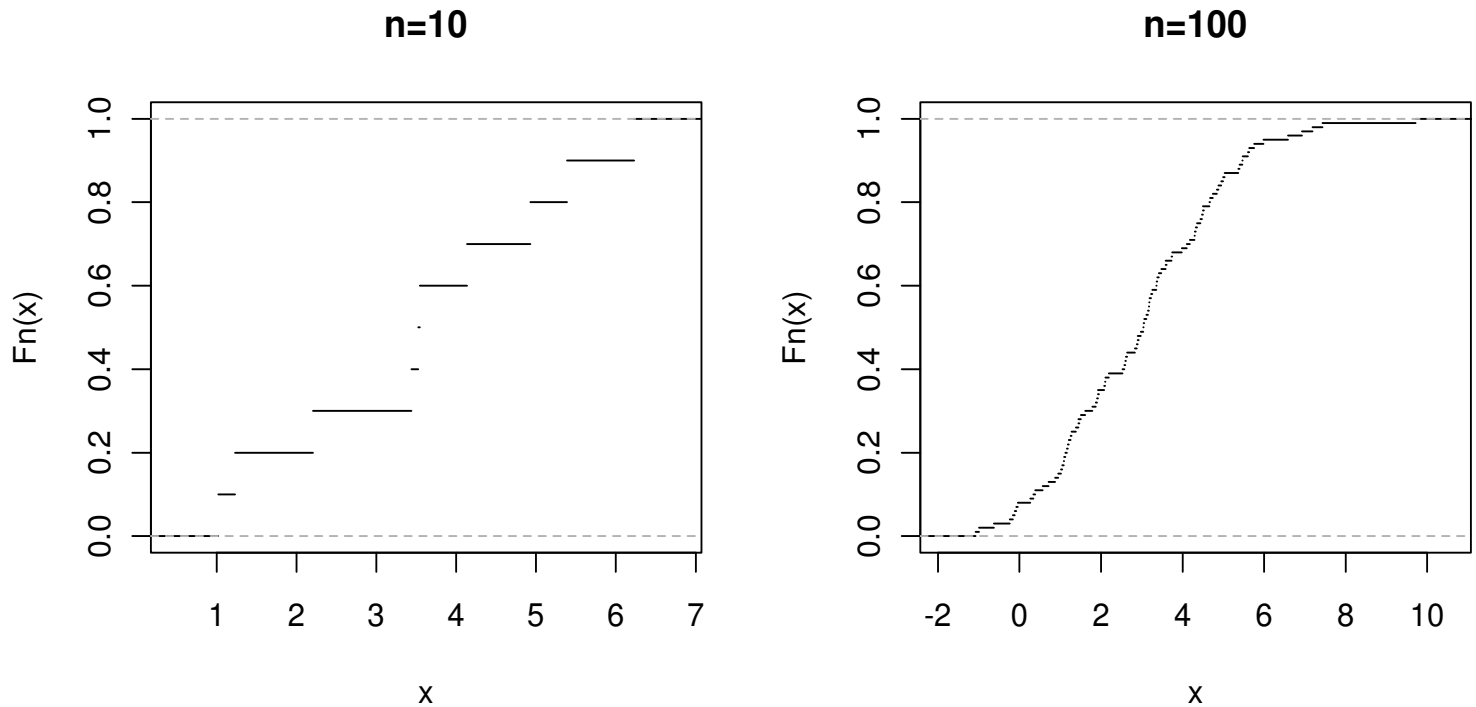
Obsérvese que, para valores prefijados de la muestra, $\mathbb{F}_n$ es una función de distribución discreta y que para cada t fijo $\mathbb{F}_n(t)$ es una v.a. (porque depende de los valores muestrales $X_1, \ldots, X_n$).

```
# Extraccion de una muestra (n=10) de una N(3,1)
x = rnorm(10,mean=3,sd=2)
# Representacion de la  distribucion empirica:
plot(ecdf(x),main="n=10",do.points=F)
```

o también

```
plot.ecdf(x,main="n=10",do.points=F)
```

En el gráfico se muestran dos funciones de distribución empírica obtenidas de este modo, para $n = 10$ y $n = 100$:



El estadístico de Kolmogorov-Smirnov

$$\|\mathbb{F}_n - F\|_\infty := \sup_t |\mathbb{F}_n(t) - F(t)|$$

es una manera de medir la "distancia" entre la función de distribución empírica F_n y la función de distribución real F.

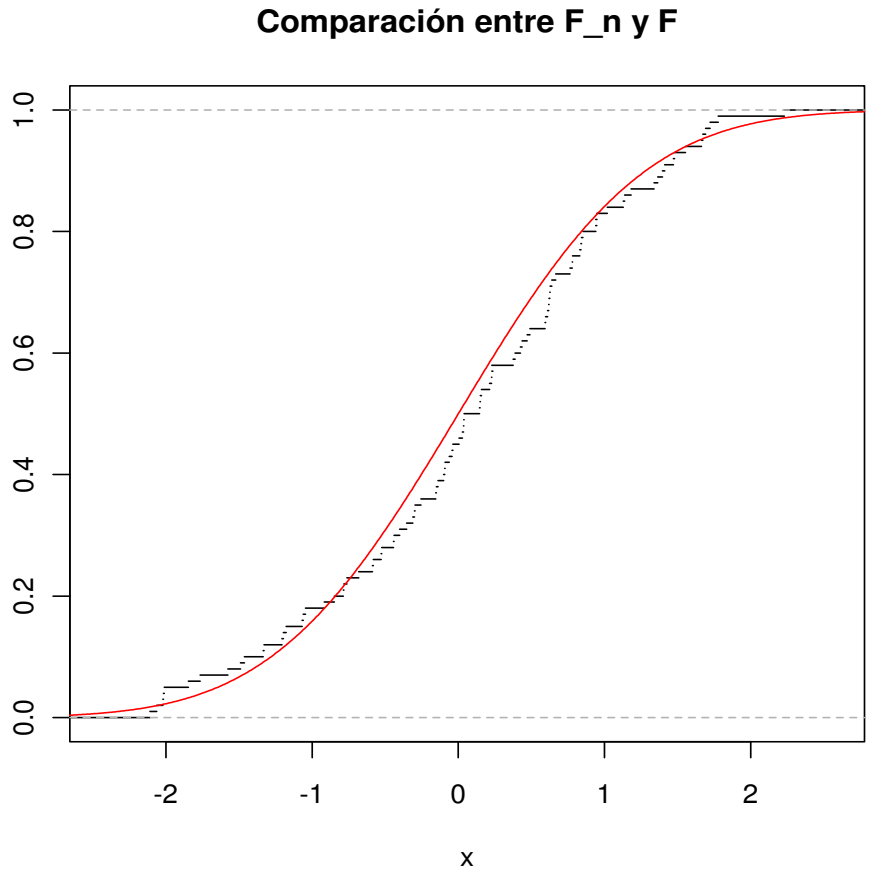
TEOREMA DE GLIVENKO-CANTELLI.- *Sean $X_1, \ldots, X_n$ v.a.i.i.d con función de distribución F. Se cumple que $\|\mathbb{F}_n - F\|_\infty \to 0$ c.s., cuando $n \to \infty$.*

La demostración de este resultado se hará en clase.

Se puede demostrar además que, cuando la muestra $X_1, \ldots, X_n$ procede de una función de distribución F continua, entonces la distribución de $\|\mathbb{F}_n - F\|_\infty$ es conocida y no depende de F. Esto se utiliza para comprobar si es plausible que un cierto modelo paramétrico F haya generado la muestra observada $X_1, \ldots, X_n$ (test de bondad de ajuste).

Comprobación empírica del teorema de Glivenko-Cantelli:

```
plot(ecdf(rnorm(100)),do.points=F,
     main="Comparacion entre Fn y F")
x = seq(-3.2,3.2,0.01)
lines(x,pnorm(x),col="red")
```



La media muestral y la media poblacional

Observemos que la media muestral

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

puede expresarse en la forma

$$\bar{X} = \int_{\mathbb{R}} x \, d\mathbb{F}_n(x).$$

Esto pone de relieve la analogía entre la media muestral y la media poblacional

$$\mu = \int_{\mathbb{R}} x \, dF(x)$$

Otras relaciones, muy importantes, entre $\bar{X}$ y μ son

1. $\bar{X}$ es estimador insesgado o centrado de μ:

$$\mathbb{E}(\bar{X}) = \mu.$$

2.

$$\mathbb{V}(\bar{X}) = \frac{\sigma^2}{n}.$$

   Error típico de la media muestral $= \sigma/\sqrt{n}$.

3. Ley fuerte de los grandes números (Kolmogorov): *Sea $\{X_k\}$ una sucesión de v.a.i.i.d. con media finita μ. Se satisface entonces*

$$\bar{X} := \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{c.s.} \mu, \text{cuando } n \to \infty. \qquad (2)$$

   En términos estadísticos, la LGN establece que "la media muestral es un estimador consistente de la media poblacional".

4. Teorema Central del Límite:

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma),$$

donde el símbolo $\xrightarrow{d}$ denota convergencia en distribución (o débil) cuando $n \to \infty$

Es decir,

$$\lim_{n\to\infty} \mathbb{P}\{\sqrt{n}(\bar{X} - \mu) \leq \sigma t\} = \Phi(t),$$

donde Φ denota la función de distribución de la $N(0, 1)$.

Por tanto, para n "grande" se tiene $\mathbb{P}\{\sqrt{n}(\bar{X} - \mu) \leq x\} \approx \Phi\left(\frac{x}{\sigma}\right)$, aunque las X_i no tengan distribución normal.

La varianza muestral y la varianza poblacional

La medida de dispersión habitual para una v.a. X es la varianza

$$\mathbb{V}(X) = \sigma^2 = \int_{\mathbb{R}} (x-\mu)^2 dF(x).$$

El análogo muestral de σ^2 es la varianza muestral (sesgada)

$$S_n^2 = \int_{\mathbb{R}} (x-\bar{X})^2 d\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Puede comprobarse que

$$\mathbb{E}(S_n^2) = \frac{n-1}{n}\sigma^2 \qquad \text{y} \qquad S_n^2 \xrightarrow{c.s.} \sigma^2.$$

Frecuentemente, en lugar de S_n^2 se utiliza la varianza muestral (insesgada)

$$S_{n-1}^2 = \frac{n}{n-1}\hat{\sigma}_n^2.$$

Se tiene que $\mathbb{E}(S_{n-1}^2) = \sigma^2$ y $\hat{S}_{n-1}^2 \xrightarrow[n\to\infty]{c.s.} \sigma^2$.

El método bootstrap

Para aproximar la distribución en el muestreo de un estimador cuando no se puede obtener de manera explícita y cerrada, Efron (1979) presentó el método *bootstrap*, que combina técnicas de simulación con el principio de sustitución o *plug-in*. Este principio, muy simple y potente, consiste en estimar cualquier cantidad que dependa de F, por ejemplo $\theta = g(F)$, por su análogo muestral, que resulta de sustituir F por $\mathbb{F}_n$, es decir, $\hat{\theta} = g(\mathbb{F}_n)$.

La palabra *bootstrap* alude a una de las aventuras del Barón de Münchausen, escritas en el siglo XVIII por R.E. Raspe, según la cual el barón cayó a las aguas de un profundo lago y consiguió salir tirando de los cordones de sus botas (de donde procede la expresión en inglés *to pull oneself up by one's bootstraps*).

Supongamos que queremos estimar la distribución del estadístico $T(X_1, \ldots, X_n; F)$ (por ejemplo, $T = \sqrt{n}(\hat{\theta} - \theta)$). Su función de distribución es

$$H_n(x) = \mathbb{P}_F\{T(X_1, \ldots, X_n; F) \leq x\},$$

donde la notación $\mathbb{P}_F$ indica que la probabilidad se calcula suponiendo que las X_i tienen función de distribución F.

El estimador *plug-in* de H_n es

$$\hat{H}_n(x) = \mathbb{P}_{\mathbb{F}_n}\{T(X_1^*, \ldots, X_n^*; \mathbb{F}_n) \leq x\}$$

y recibe el nombre de estimador bootstrap ideal. La muestra $X_1^*, \ldots, X_n^*$ es de la distribución empírica $\mathbb{F}_n$.

En la práctica, suele ser imposible obtener una expresión cerrada para $\hat{H}_n(x)$. Sin embargo, al haber sustituido F por $\mathbb{F}_n$ hemos pasado de la población original F, en la que sólo disponemos de una muestra de tamaño n, a la distribución empírica $\mathbb{F}_n$ de la que podemos muestrear tanto como nuestra capacidad de cálculo lo permita.

En concreto, se simulan B muestras bootstrap $X_1^{*b}, \ldots, X_n^{*b}$, $b = 1, \ldots, B$, de $\mathbb{F}_n$. Para ello, basta con muestrear con reemplazamiento entre los datos originales $X_1, \ldots, X_n$. Podemos calcular el valor del estadístico T correspondiente a cada una de estas muestras artificiales

$$T^{*(b)} = T(X_1^{*b}, \ldots, X_n^{*b}; \mathbb{F}_n^{*b}).$$

El valor de $\hat{H}_n(x)$ se aproxima mediante

$$\hat{H}_n^*(x) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\{T^{*(b)} \leq x\}}.$$

Algoritmo *bootstrap* para aproximar la función de distribución $H_n(x)$ de un estadístico $T = T(X_1, \ldots, X_n; F)$

B.1. Se estima F mediante $\mathbb{F}_n$.

B.2. Se obtienen B muestras *bootstrap* $X_1^{*b}, \ldots, X_n^{*b}$, $b = 1, \ldots, B$, de la distribución empírica $\mathbb{F}_n$, sorteando con reemplazamiento entre los datos originales $X_1, \ldots, X_n$.

B.3. Se determina el estadístico T en cada muestra *bootstrap*, obteniendo $T^{*(b)} = T(X_1^{*b}, \ldots, X_n^{*b}; \mathbb{F}_n^{*b})$, $b = 1, \ldots, B$.

B.4. Se calcula la distribución empírica de la muestra *bootstrap* de T: $\hat{H}_n^*(x) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\{T^{*(b)} \leq x\}}$.

El *bootstrap* implica, pues, una aproximación en dos pasos:

$$H_n(x) \simeq \hat{H}_n(x) \simeq H_n^*(x).$$

La LFGN (aplicada a observaciones de la distribución empírica $\mathbb{F}_n$) garantiza que

$$H_n^*(x) \xrightarrow[B\to\infty]{\text{c.s.}} H_n(x).$$

Por esta razón se suele escoger un valor grande de B.

Sin embargo, probar que los términos $H_n(x)$ y $\hat{H}_n(x)$ convergen al mismo límite cuando $n \to \infty$ requiere trabajo teórico adicional. Se denomina establecer la consistencia del bootstrap. Intuitivamente, hay que demostrar que al aumentar el tamaño muestral n el mundo bootstrap se parece lo suficiente al mundo real.

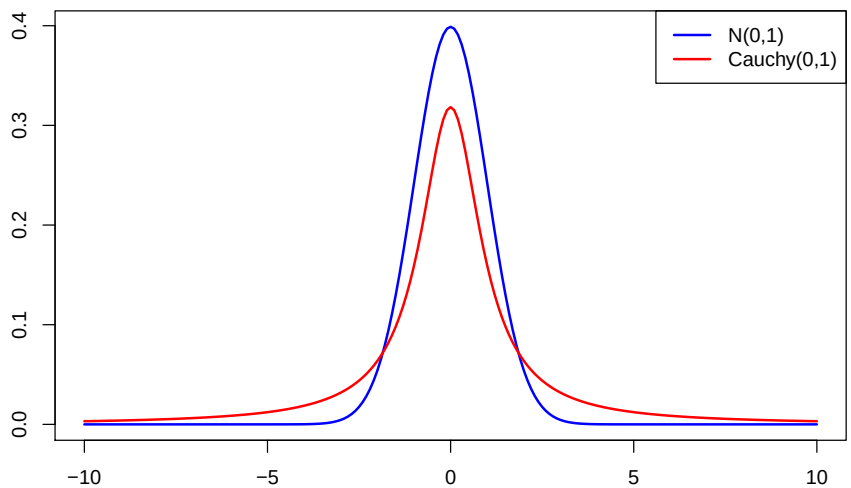
El mismo algoritmo sirve esencialmente para estimar cualquier aspecto de la distribución, en lugar de la función de distribución completa. Por ejemplo, quizá interesa estimar la varianza del estimador $\hat{\theta}$ de un parámetro θ. En este caso el estimador *bootstrap* ideal de $\mathbb{V}_F(\hat{\theta})$ es $\mathbb{V}_{\mathbb{F}_n}(\hat{\theta}^*)$, que aproximaremos mediante la varianza muestral de las B muestras *bootstrap*

$$\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b}-\bar{\theta}^*)^2.$$

Ejemplo: Sean $X_1, \ldots, X_n$ v.a.i.i.d. de una distribución de Cauchy centrada en θ y con parámetro de escala 1. La densidad es

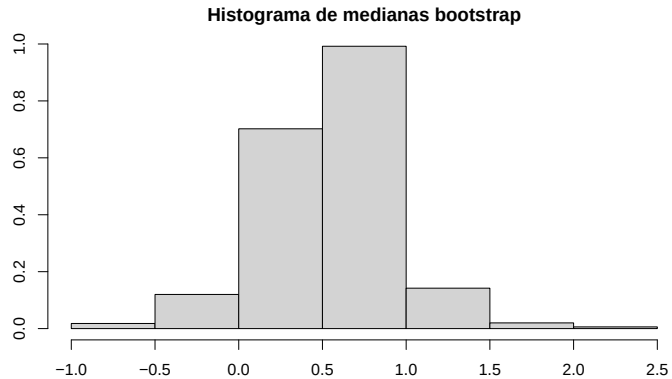
$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}, \; x \in \mathbb{R}.$$

La esperanza de esta distribución no existe, por lo que para estimar θ se usa la mediana muestral ($\hat{\theta}$). ¿Cuál es la distribución de esta mediana? ¿Y la desviación típica?



En el código siguiente generamos una muestra original de una distribución de Cauchy con $n = 30$ datos, y aplicamos el método *bootstrap* con $B = 1000$.

```r
set.seed(150)

B <- 1000
n <- 30
theta <- 1

muestra_original <- rcauchy(n)
mediana_original <- median(muestra_original)

# Generamos las remuestras (matriz n x B, cada columna una remuestra)
muestras_bootstrap <- sample(muestra_original, n*B, rep = TRUE)
muestras_bootstrap <- matrix(muestras_bootstrap, nrow = n)

# Medianas de las remuestras
medianas_bootstrap <- apply(muestras_bootstrap, 2, median)
hist(medianas_bootstrap,freq=F,main="Histograma de medianas bootstrap")

## Estimador bootstrap de la desv. tip. de la mediana
sd(medianas_bootstrap)
[1] 0.4243034
```



Distribución empírica y estimadores kernel

Obsérvese que

$$\hat{f}_n(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(t - X_i\right)$$

$$= \int_{\mathbb{R}} K_h(t-x)d\mathbb{F}_n(x),$$

es decir, que el estimador kernel $\hat{f}_n(t)$ puede considerarse como la convolución del núcleo re-escalado $K_h(z) = \frac{1}{h}K\left(\frac{z}{h}\right)$ con la medida de probabilidad empírica $\mathbb{F}_n$.

Intuitivamente esto significa que la distribución correspondiente a la función de densidad $\hat{f}_n$ puede considerarse como una "versión suavizada" de la distribución empírica.

TEOREMA.- *Sean $X_1, X_2, \ldots,$ v.a.i.i.d. con distribución común absolutamente continua de densidad f.*
Supongamos que

(a) el núcleo K es una función de densidad acotada con $|x|K(x) \to 0$ cuando $|x| \to \infty$;

(b) $h = h_n \to 0$ y que $nh_n \to \infty$;

(c) la densidad f es acotada y continua en un punto t.

Entonces

$$\hat{f}_n(t) \xrightarrow{P} f(t).$$

La demostración se hará en clase.

Este resultado indica que los estimadores kernel pueden utilizarse para estimar la función de densidad de las v.a. X_i. Hay versiones mucho más generales de este resultado. Aquí se ha elegido ésta por la sencillez de su demostración.

Aplicación de los estimadores kernel para definir la moda muestral

Sea X una v.a. con densidad f. Supongamos que f es continua y que tiene un único máximo. Se define entonces **la moda** de f como el valor θ que verifica

$$f(\theta) = \max_x f(x).$$

Sea $\hat{f}_n$ una sucesión de estimadores kernel basados en una función núcleo K que es una densidad tal que $\lim_{z\to\pm\infty} K(z) = 0$. Se define una **moda muestral** como un valor θ_n que verifica

$$\hat{f}(\theta_n) = \max_x \hat{f}_n(x).$$

TEOREMA(Consistencia de la moda muestral).- *Supongamos que*

(a) *la densidad f es uniformemente continua en $\mathbb{R}$ y alcanza un único máximo (moda) en θ.*

(b) *$\hat{f}_n$ una sucesión de estimadores kernel cuya función núcleo K es una densidad tal que $\lim_{z\to\pm\infty} K(z) = 0$.*

(c) $\sup_t |\hat{f}_n(t) - f(t)| \xrightarrow{c.s.} 0$, *cuando $n \to \infty$.*

Entonces

$$\theta_n \xrightarrow{c.s.} \theta, \tag{3}$$

*siendo $\{\theta_n\}$ cualquier sucesión de modas muestrales. Si en la hipótesis (c) se reemplaza la convergencia c.s. por convergencia en probabilidad, la consistencia (**??**) se obtiene también en probabilidad.*

Puede probarse que $h \to 0$ y $nh/\log n \to \infty$ son condiciones suficientes para que (c) se cumpla (bajo ciertas condiciones sobre K que se verifican para el núcleo gaussiano y otros núcleos usuales).

Estadísticos de orden

Dada una muestra $X_1, \ldots, X_n$ se denotan por

$$X_{(1)} \leq \ldots \leq X_{(n)}$$

las variables de la muestra ordenadas de menor a mayor, es decir, $X_{(1)}$ es el mínimo de la muestra, $X_{(2)}$ la siguiente observación más pequeña y $X_{(n)}$ es el máximo de la muestra.

En algunos libros, cuando se quiere señalar el papel del tamaño muestral en las propiedades del estadístico de orden, se denota $X_{(k)} = X_{k:n}$.

Los estadísticos de orden $X_{(k)}$ pueden utilizarse para definir la mediana o los cuartiles. Sin embargo, la función cuantílica proporciona una manera más directa de definir estos conceptos.

Sean $X_1, \ldots, X_n$ v.a.i.i.d con densidad (resp. función de masa) f y $(Y_1, \ldots, Y_n) = (X_{(1)}, \ldots, X_{(n)})$ la muestra ordenada. La función de densidad (resp. masa) conjunta de la muestra ordenada es

$$g(y_1, \ldots, y_n) = n! f(y_1) \cdots f(y_n), \quad \text{si } y_1 < \cdots < y_n$$

y 0 en caso contrario.

Determinemos la distribución marginal de $X_{(k)}$. Para ello, dado $x \in \mathbb{R}$, definimos la v.a. auxiliar $\xi =$ "número de observaciones muestrales menores o iguales que x" $\sim B(n, F(x))$. Entonces

$$F_{X_{(k)}}(x) = \mathbb{P}\{X_{(k)} \leq x\} = \mathbb{P}\{\xi \geq k\} = \sum_{j=k}^{n} \binom{n}{j} (F(x))^j (1-F(x))^{n-j}$$

En el caso de que F tenga densidad f, derivando la expresión anterior y arreglando los términos resultantes se obtiene:

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} f(x) F(x)^{k-1} (1 - F(x))^{n-k}.$$

La función cuantílica

Sea F la función de distribución de una v.a. X. Se define la función cuantílica correspondiente a F, como la función F^{-1}, definida en el intervalo $(0, 1)$ mediante

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}.$$

Se llama cuantil poblacional de orden p al valor $F^{-1}(p)$ de la función cuantílica en p.

El estimador natural del cuantil poblacional de orden p es el análogo cuantil muestral de orden p definido a partir de la distribución empírica, es decir, $\mathbb{F}_n^{-1}(p)$.

El valor $p = 1/2$ proporciona la mediana poblacional, $F^{-1}(1/2)$, y muestral, $\mathbb{F}_n^{-1}(1/2)$.

En general,

$$\mathbb{F}_n^{-1}(p) = Q_p = c_{n,p} X_{([np])} + (1 - c_{n,p}) X_{([np]+1)},$$

donde $c_{n,p} = 1$ si $np \in \mathbb{N}$ y $c_{n,p} = 0$ si $np \notin \mathbb{N}$.

Comportamiento asintótico

Si, para todo $\epsilon > 0$, se satisface que $p < F(Q_p + \epsilon)$, entonces se cumple que

$$\mathbb{F}_n^{-1}(p) \xrightarrow{c.s.} F^{-1}(p).$$

Si F es derivable con derivada f continua y estrictamente positiva en la mediana $\theta = F^{-1}(1/2)$, entonces se cumple que

$$\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}(\mathbb{F}_n^{-1}(1/2) - F^{-1}(1/2)) \xrightarrow{\mathrm{d}} N\left(0, \sigma = \frac{1}{2f(\theta)}\right),$$

para el estimador *plug-in* de la mediana, $\hat{\theta}_n = \mathbb{F}_n^{-1}(1/2)$.

Referencias

Casella, G., Berger, R.L. (2002). *Statistical Inference*. Duxbury/Thomson Learning. Capítulos 1 al 5.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.

Gentle, J.E. (2013). *Theory of Statistics*. George Mason University. `https://mason.gmu.edu/~jgentle/books/MathStat.pdf`

Knight, K. (2000). *Mathematical Statistics*. Chapman & Hall CRC. Capítulos 1 al 3.