

## Tema 2. Estimación de la función de distribución (Parte 2. Aplicaciones de $F_n$ )

Rosa M. Crujeiras

Alberto Rodríguez



Dpto. de Estadística e Investigación Operativa  
Máster en Técnicas Estadísticas  
Curso 2009-2010

Consideremos  $X$  una v.a. con distribución  $F$  y sea  $\mathcal{X} = (X_1, \dots, X_n)$  una m.a.s. de esta variable.

Generalmente, nos interesa estimar las siguientes características:

- ▶ Esperanza:

$$\theta_1(F) = \int x dF(x)$$

- ▶ Momento orden 2:

$$\theta_2(F) = \int x^2 dF(x)$$

- ▶ Cuantil  $(1 - \alpha)$ :

$$\mathbb{P}(X \leq \theta_3(F)) = 1 - \alpha$$

Las características anteriores se pueden estimar a partir de:

- ▶ Media muestral:

$$T_1(\mathcal{X}) = \frac{1}{n} \sum_{i=1}^n X_i = \int x dF_n(x)$$

- ▶ Momento muestral de orden 2:

$$T_2(\mathcal{X}) = \frac{1}{n} \sum_{i=1}^n X_i^2 = \int x^2 dF_n(x)$$

- ▶ Cuantil muestral  $(1 - \alpha)$ :

$$T_3(\mathcal{X}) = (1 - \alpha) - \text{cuantil de } F_n$$

Para determinar la validez de un estimador, se evalúan las siguientes cantidades:

- ▶ Sesgo:

$$\text{Bias}_F(T) = \mathbb{E}_F(T(\mathcal{X})) - \theta(F)$$

- ▶ Varianza:

$$\text{Var}_F(T) = \mathbb{E}_F(T^2(\mathcal{X})) - \mathbb{E}_F^2(T(\mathcal{X}))$$

- ▶ Error Cuadrático Medio:

$$\text{MSE}_F(T) = \mathbb{E}_F((T(\mathcal{X}) - \theta(F))^2) = \text{Bias}_F^2(T) + \text{Var}_F(T)$$

Todas estas cantidades son conocidas si hemos determinado la distribución de  $T(\mathcal{X})$ :

$$G(x) = \text{Distr}_{T,F}(x) = \mathbb{P}_F(T(\mathcal{X}) \leq x)$$

Tanto el sesgo como la varianza de un estimador dependen de  $F$ , que es **desconocida**, pero disponemos de un buen estimador

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \leq x\}.$$

**Plug-in: reemplazar  $F$  por  $F_n$**

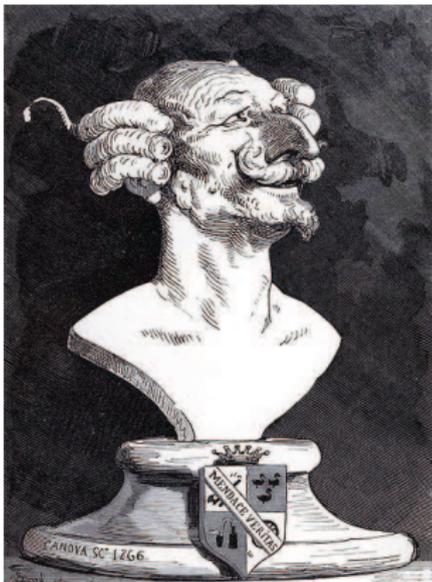
Ya lo hemos hecho para construir estimadores:

- ▶ Esperanza:

$$\theta_1(F) = \int x dF(x)$$

- ▶ Media muestral:

$$T_1(\mathcal{X}) = \hat{\theta}_1(F) = \theta_1(F_n) = \int x dF_n(x) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$



Bootstrap (Efron, 1979) fue propuesto como una alternativa al jackknife, para la estimación de la varianza de un estadístico. Hoy en día el Bootstrap se utiliza en diferentes contextos de la Estadística, para dar solución a las preguntas de la inferencia (estimación, intervalos de confianza y contraste de hipótesis).

## Mundo real

- $\mathcal{X} = (X_1, \dots, X_n)$  m.a.s. de  $X \sim F$  desconocida
- $T(\mathcal{X})$ , estimator of  $\theta(F)$
- $G(x) = \mathbb{P}_F(T(\mathcal{X}) \leq x)$

## Mundo Bootstrap

- $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$ , extraída de la población  $\mathcal{X}$  con  $F_n$  conocida
- $T(\mathcal{X}^*)$ , estimador of  $\theta(F_n)$
- Las propiedades distribucionales de  $T(\mathcal{X}^*)$  son conocidas (condicionalmente en  $\mathcal{X}$ )

## Principio Bootstrap

Las propiedades distribucionales de  $T(\mathcal{X}^*)$  pueden imitar las de  $T(\mathcal{X})$ .

## Estimaciones Bootstrap:

- ▶ Sesgo Bootstrap:

$$\text{Bias}^*(T) = \text{Bias}_{F_n}(T) = \mathbb{E}_{F_n}(T(\mathcal{X}^*) - \theta(F_n))$$

- ▶ Varianza Bootstrap:

$$\text{Var}^*(T) = \text{Var}_{F_n}(T) = \mathbb{E}_{F_n}(T^2(\mathcal{X}^*)) - \mathbb{E}_{F_n}^2(T(\mathcal{X}^*))$$

- ▶ Error Cuadrático Medio:

$$\text{MSE}^*(T) = \text{MSE}_{F_n}(T) = \mathbb{E}_{F_n}((T(\mathcal{X}^*) - \theta(F_n))^2)$$

- ▶ Distribución Bootstrap:

$$\hat{G}(x) = \text{Distr}_{T, F_n}(x) = \mathbb{P}_{F_n}(T(\mathcal{X}^*) \leq x) = \mathbb{P}(T(\mathcal{X}^*) \leq x | \mathcal{X})$$

La clave para que todo funcione:

Bajo condiciones de regularidad, para  $n$  suficientemente grande:

$$\hat{G}(x) = \text{Dist}_{T, F_n}(x) \approx \text{Dist}_{T, F}(x) = G(x)$$

Algunas puntualizaciones:

- ▶ Normalmente,  $T(\mathcal{X}) = \theta(F_n)$ , aunque de manera más general  $T(\mathcal{X}) = g_n(\theta(F_n))$  ( $g_n$  conocida).
- ▶ Si  $F(\cdot) = F_\tau(\cdot)$  con  $\tau \in \mathbb{R}^k$  conocida, se puede hacer lo mismo reemplazando  $F_n$  por  $F_{\hat{\tau}}$  (bootstrap paramétrico).
- ▶ En el mundo Bootstrap, todo es conocido (condicionalmente en  $\mathcal{X}$ ), aunque puede ser difícil de calcular.

## Aproximación de $G$ por Monte-Carlo

1. Extraer  $M$  muestras  $(\mathcal{X}^{(m)}, m = 1, \dots, M)$  de  $F$  (conocida)
2. Cada muestra  $\mathcal{X}^{(m)} = (X_1^{(m)}, \dots, X_n^{(m)})$ , donde  $X_i^{(m)}$  iid  $F$
3. Calcular  $T(\mathcal{X}^{(m)})$ ,  $m = 1, \dots, M$
4. Aproximación por Monte-Carlo (SLLN)

$$G(x) = \text{Dist}_{T,F}(x) \approx \frac{1}{M} \sum_{i=1}^M \mathbb{I}(T(\mathcal{X}^{(m)}) \leq x)$$

Estimación por Monte-Carlo de sesgo, varianza y MSE:

► Sesgo MC:

$$\text{Bias}_F(T) \approx \frac{1}{M} \sum_{m=1}^M T(\mathcal{X}^{(m)}) - \theta(F)$$

► Varianza MC:

$$\text{Var}_F(T) \approx \frac{1}{M} \sum_{m=1}^M T^2(\mathcal{X}^{(m)}) - \left( \frac{1}{M} \sum_{m=1}^M T(\mathcal{X}^{(m)}) \right)^2$$

► Error Cuadrático Medio MC:

$$\text{MSE}_F(T) \approx \frac{1}{M} \sum_{m=1}^M \left( T(\mathcal{X}^{(m)}) - \theta(F) \right)^2$$

Las aproximaciones por Monte-Carlo siempre se pueden obtener en el Mundo Bootstrap:

1. Extraer  $B$  muestras  $(\mathcal{X}^{*(b)}, b = 1, \dots, B)$  de  $F_n$  (conocida)
2. Cada muestra  $\mathcal{X}^{*(b)} = (X_1^{*(b)}, \dots, X_n^{*(b)})$ , donde  $X_i^{*(b)} \text{ iid } F_n$
3. Calcular  $T(\mathcal{X}^{*(b)})$ ,  $b = 1, \dots, B$
4. Aproximación por Monte-Carlo (SLLN)

## Estimaciones por Monte-Carlo en el Mundo Bootstrap:

- Sesgo:

$$\text{Bias}^*(T) \approx \frac{1}{B} \sum_{b=1}^B T(\mathcal{X}^{*(b)}) - \theta(F_n)$$

- Varianza:

$$\text{Var}^*(T) \approx \frac{1}{B} \sum_{b=1}^B T^2(\mathcal{X}^{*(b)}) - \left( \frac{1}{B} \sum_{b=1}^B T(\mathcal{X}^{*(b)}) \right)^2$$

- Error Cuadrático Medio:

$$\text{MSE}^*(T) \approx \frac{1}{B} \sum_{b=1}^B \left( T(\mathcal{X}^{*(b)}) - \theta(F_n) \right)^2$$

- Distribución:

$$\hat{G}(x) = \mathbb{P}_F(\bar{X} \leq x) = \text{Dist}_{T,F}(x) \approx \frac{1}{B} \sum_{b=1}^B I\{T(\mathcal{X}^{*(b)}) \leq x\}$$

Queremos construir un IC para  $\theta(F)$  y para ello disponemos de una raiz  $(T(\mathcal{X}) - \theta(F))$ , con distribución (conocida)  $H$ :

$$H(x) = \mathbb{P}_F(T(\mathcal{X}) - \theta(F) \leq x)$$

El IC de nivel  $(1 - \alpha)\%$  para  $\theta(F)$  se obtiene de:

$$\mathbb{P}_F \left( u \left( \frac{\alpha}{2} \right) \leq T(\mathcal{X}) - \theta(F) \leq u \left( 1 - \frac{\alpha}{2} \right) \right) = 1 - \alpha$$

donde  $u(a)$  denotan los cuantiles de  $H$ . Por tanto, el intervalo de nivel  $(1 - \alpha)\%$  para  $\theta(F)$  es:

$$\left[ T(\mathcal{X}) - u \left( 1 - \frac{\alpha}{2} \right), T(\mathcal{X}) - u \left( \frac{\alpha}{2} \right) \right]$$

## Raíces pivotaes asintóticas (MLE)

Sea  $T$  el MLE de  $\theta(F)$  y  $l(\mathcal{X}; \theta)$  la función de log-verosimilitud ( $\dot{l}(\mathcal{X}; \theta)$  la *score function*). Entonces:

$$T(\mathcal{X}) - \theta(F) \sim AN(0, \mathcal{F}_n^{-1}(\theta)),$$

donde  $\mathcal{F}_n$  es la información de Fisher:

$$\mathcal{F}_n(\theta) = \text{Var}(\dot{l}(\mathcal{X}; \theta)) = -E(\ddot{l}(\mathcal{X}; \theta))$$

$\mathcal{F}_n(\theta)$  se puede estimar como:

- $\mathcal{F}_n(\hat{\theta})$
- información de Fisher observada  $-\ddot{l}(\mathcal{X}; \theta)$

Para construir el **IC Bootstrap básico** para  $\theta(F)$  se utiliza la estimación  $\widehat{G}(x)$  de la distribución de  $T(\mathcal{X})$  para obtener los cuantiles  $u(a)$ . El IC resultante es:

$$\left[ T(\mathcal{X}) - u^* \left( 1 - \frac{\alpha}{2} \right), T(\mathcal{X}) - u^* \left( \frac{\alpha}{2} \right) \right]$$

donde  $u^*(a)$  es el  $a$ -cuantil de  $\widehat{H}(x) = \widehat{G}(x + \theta(F_n))$ , estimación Bootstrap de  $H(x)$ :

$$\begin{aligned} \widehat{H}(x) &= \mathbb{P}_{F_n} (T(\mathcal{X}^*) - \theta(F_n) \leq x) \\ &= \mathbb{P} (T(\mathcal{X}^*) - \theta(F_n) \leq x \mid \mathcal{X}) \\ &\approx \frac{1}{B} \sum_{b=1}^B \mathbb{I} \left( T(\mathcal{X}^{*(b)}) - \theta(F_n) \leq x \right). \end{aligned}$$

Los cuantiles de  $\widehat{H}(x)$  se obtienen fácilmente de los cuantiles  $v^*(a)$  de la distribución  $\widehat{G}(x) = Dist_T^*(x)$ :

$$\widehat{G}(x) = Dist_T^*(x) \approx \frac{1}{B} \sum_{b=1}^B \mathbb{I}(T(\mathcal{X}^{*(b)}) \leq x).$$

Entonces  $v^*(a) = \widehat{G}^{-1}(a)$ , y por tanto,

$$\mathbb{P}_{F_n} \left( T(\mathcal{X}^*) \leq v^*(a) \right) = \widehat{G}(v^*(a)) = a$$

Dado que  $\widehat{H}(x) = \widehat{G}(x + \theta(F_n))$  simplemente se trasladan los cuantiles para obtener  $u^*(a)$ :

$$u^*(a) = v^*(a) - \theta(F_n)$$

Los cuantiles  $v^*(a)$  de  $\widehat{G}(x) = Dist_T^*(x)$  se obtienen del siguiente algoritmo :

- Consideremos los estadísticos ordenados

$$T(\mathcal{X}^{*(b)}), b = 1, \dots, B:$$

$$T_{(1)}^* \leq \dots \leq T_{(B)}^*.$$

- Sean  $k1 = [B \frac{\alpha}{2}] + 1$  y  $k2 = [B (1 - \frac{\alpha}{2})]$ , donde  $[\cdot]$  denota la parte entera.
- Finalmente:

$$v^* \left( \frac{\alpha}{2} \right) = T_{(k1)}^*, \quad v^* \left( 1 - \frac{\alpha}{2} \right) = T_{(k2)}^*$$

Si  $T(\mathcal{X}) = \theta(F_n)$  (situación usual), el IC bootstrap para  $\theta(F)$  viene dado por:

$$\begin{aligned} & \left[ T(\mathcal{X}) + \theta(F_n) - v^* \left( 1 - \frac{\alpha}{2} \right), T(\mathcal{X}) + \theta(F_n) - v^* \left( \frac{\alpha}{2} \right) \right] \\ & = \left[ 2 T(\mathcal{X}) - v^* \left( 1 - \frac{\alpha}{2} \right), 2 T(\mathcal{X}) - v^* \left( \frac{\alpha}{2} \right) \right] \end{aligned}$$

Supongamos que  $X \sim N(\mu, \sigma^2)$  y queremos obtener un IC para  $\mu$ , con  $\sigma^2$  desconocida. Tenemos que:

$$T(\mathcal{X}) = \sqrt{n} \frac{\bar{X} - \mu}{S} \sim t_{n-1}$$

donde  $S^2$  denota la cuasivarianza muestral y  $t_{n-1}$  es una distribución T-Student con  $(n - 1)$  grados de libertad. El IC para  $\mu$  se puede obtener como:

$$\left[ \bar{X} - t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \right]$$

Si la población no es Normal, entonces habrá que determinar  $u(\alpha/2)$ ,  $u(1 - \alpha/2)$  tales que:

$$\mathbb{P}(u(\alpha/2) \leq T(\mathcal{X}) \leq u(1 - \alpha/2))$$

para poder obtener:

$$\left[ \bar{X} - u(1 - \alpha/2) \frac{S}{\sqrt{n}}, \bar{X} - u(\alpha/2) \frac{S}{\sqrt{n}} \right]$$

Intervalos Bootstrap estudentizados (ejemplo para  $\mu$ ):

1. Generar  $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$  muestra Bootstrap (de  $F_n$ ).
2. Evaluar el estadístico:

$$T^* = \bar{n} \frac{\overline{X^*} - \bar{X}}{S^*}$$

donde  $\overline{X^*}$  y  $S^*$  son la media y la cuasidesviación típica de la muestra Bootstrap.

3. Repetir  $B$  veces los pasos 1 y 2 y obtener  $t_1^*, \dots, t_B^*$  evaluaciones de  $T^*$ .
4. Ordenar de menor a mayor los  $t_b^*$  y extraer los cuantiles (posiciones  $\alpha/2 * B$  y  $(1 - \alpha/2) * B$ ).
5. Calcula el IC Bootstrap como:

$$\left[ \bar{X} - u^*(1 - \alpha/2) \frac{S}{\sqrt{n}}, \bar{X} - u(\alpha/2) \frac{S}{\sqrt{n}} \right]$$

Ejercicio: comprobaremos el funcionamiento del método anterior, considerando una m.a.s. de tamaño  $n = 100$  de una  $Exp(1)$ .

1. Genera una m.a.s. de tamaño  $n$  de  $Exp(1)$ .
2. Calcula el IC Bootstrap para  $\mu = \mathbb{E}(X) = 1$ , con  $B = 1000$  y  $\alpha = 0'05$ .
3. Comprueba si  $\mu$  está en el IC.
4. Repite el proceso anterior  $M = 100$  veces y comprueba cuántas veces  $\mu$  está contenida en los intervalos.
5. Obtén IC por la aproximación Normal y comprueba cuántas veces  $\mu$  está dentro de esos intervalos.

MUNDO REAL	MUNDO BOOTSTRAP
$F$ desconocida $\theta(F)$ desconocida	Para $\mathcal{X} = (X_1, \dots, X_n)$ , $F_n$ es conocida $\theta(F_n)$ conocida
<div style="border: 1px solid black; padding: 5px; display: inline-block;">DGP <math>F : X \sim F</math></div>	<div style="border: 1px solid black; padding: 5px; display: inline-block;">DGP* <math>F_n : X^* \sim F_n</math></div>
$\Downarrow$ $\mathcal{X} = (X_1, \dots, X_n)$	$\Downarrow$ $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$

MUNDO REAL	MUNDO BOOTSTRAP
$T(\mathcal{X})$ estimador de $\theta(F)$	$T(\mathcal{X}^*)$ estimador de $\theta(F_n)$
$G(x) = \mathbb{P}_F(T(\mathcal{X}) \leq x)$	$\hat{G}(x) = P_{F_n}(T(\mathcal{X}^*) \leq x)$
Características	Características
$\mathbb{E}_F(T(\mathcal{X}))$	$\mathbb{E}_{F_n}(T(\mathcal{X}^*))$
$\text{Var}_F(T(\mathcal{X}))$	$\text{Var}_{F_n}(T(\mathcal{X}^*))$

## MUNDO REAL

$$W = T(\mathcal{X}) - \theta(F)$$

$$H(x) = \mathbb{P}_F(W \leq x) \Rightarrow u(a) = H^{-1}(a)$$

$$\Downarrow$$
IC para  $\theta$ 

$$[T(\mathcal{X}) - u(1 - \alpha/2), T(\mathcal{X}) - u(\alpha/2)]$$

## MUNDO BOOTSTRAP

$$W^* = T(\mathcal{X}^*) - \theta(F_n)$$

$$\hat{H}(x) = \mathbb{P}_{F_n}(W^* \leq x) \Rightarrow \hat{u}(a) = \hat{H}^{-1}(a)$$

$$\Downarrow$$
IC Bootstrap para  $\theta$ 

$$[T(\mathcal{X}) - \hat{u}(1 - \alpha/2), T(\mathcal{X}) - \hat{u}(\alpha/2)]$$