Introducción a la metodología bootstrap

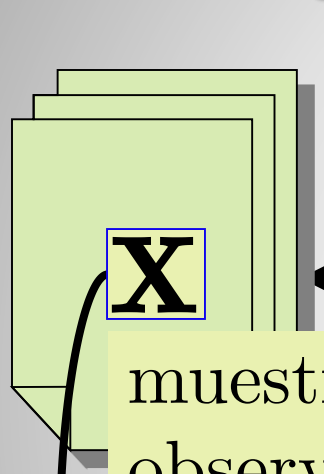
Mètodes no paramètrics i de remostratge

Grau Interuniversitari en Estadística UB - UPC
Departament d'Estadística
Universitat de Barcelona
Jordi Ocaña Rebull

Puntos a tratar

- Elementos de un problema de inferencia estadística
- Determinación de la distribución muestral (o de alguna de sus características)
- Principio “plug-in” y bootstrap
- Principio de Montecarlo y bootstrap
- Necesaria correspondencia entre “mundo real” y “mundo bootstrap”
- Ejemplos

Elementos de un problema de inferencia estadística



$\mathbf{X}$

$\mathbf{x}$

muestra observada

$F \in \mathcal{F}$

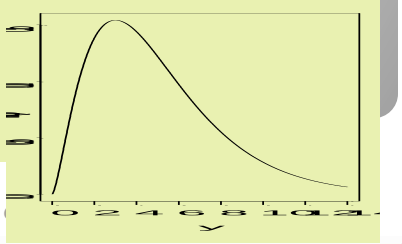
Modelo probabilístico, mecanismo "generador" de los datos

$t \quad \mathbf{x}$

estadísticos

$R \quad t, F, \mathbf{x}$

Medidas de precisión



Elementos de un problema de I.E. Ejemplo introductorio

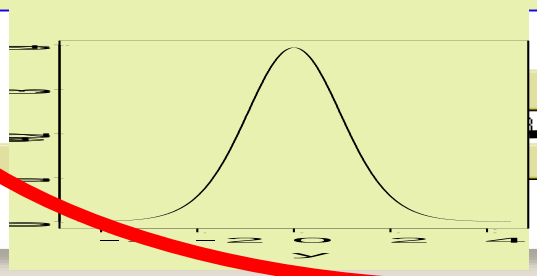
$$\mathbf{x} = x_1, \ldots, x_n$$

muestra aleatoria simple de tamaño n

$$f\ \mathbf{x}; \mu, \sigma^2 = \prod_{i=1}^{n} \sigma\sqrt{2\pi}^{-1} e^{-\frac{x_i - \mu^2}{2\sigma^2}}$$

Normal de media y varianza desconocidas

$$t\ \mathbf{x}$$



Distribución muestral de un estadístico

- Cálculo de p-valores, valores críticos...: indispensable conocer la distribución muestral del estadístico de interés

$$G\left(t, n, F\left(x;\theta\right), \ldots\right)$$

- Cómo determinarla:
  - Paramétrico exacto
  - Asintótico
  - Rangos...

Principio "plug-in" y bootstrap (en sentido amplio)

- Fijémonos en el paso $G = G(F(;\theta),\ldots)$
- Si $\hat{F}$ es buena estimación de F a partir de los datos, razonable aproximar G mediante:

$$G\left(\hat{F},\ldots\right)$$

→ Principio "plug-in"

- **Metodologia bootstrap ≡ inferencia basada en el Principio "plug-in"**

Ejemplo: aplicación automàtica del Principio "plug-in"

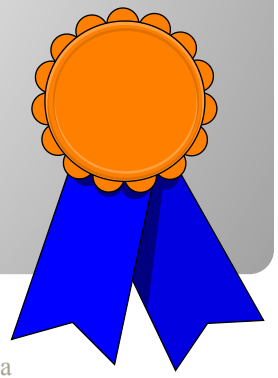
- A menudo $\hat{F}$ es la distribución empírica, F_n, discreta, que assigna probabilidad $1/n$ a cada valor muestral y 0 a cualquier otro

- Si interessa característica concreta como

$$\text{var}_F\left(\bar{X}\right) = \frac{\text{var}_F\left(X\right)}{n}$$

. . . Según Principio "plug-in":

$$\text{var}_{F_n}\left(\bar{X}\right) = \frac{\text{var}_{F_n}\left(X\right)}{n} = \frac{s^2}{n}$$



Detalles del cálculo anterior

$$\mathrm{var}_{F_n}\left(\bar{X}^*\right) = \frac{E_{F_n}\left(\left(X^* - E_{F_n}\left(\bar{X}^*\right)\right)^2\right)^2}{n}$$

$$E_{F_n}\left(\bar{X}^*\right) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}\left(= E_{F_n}\left(\bar{X}^*\right)\right)$$

$$E_{F_n}\left(\left(\bar{X}^* - \bar{x}\right)^2\right) = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \frac{1}{n} = s^2$$

- Conveniencia de notación X^* en lugar de X: no es la misma v.a

Dificultades en la aplicación del Principio "plug-in"

- No tan (o a veces nada) clara su aplicación en situaciones más complejas:
  - otras características de la distribución muestral, incluso para estadísticos sencillos como la media muestral (p.e. un cuantil, ...)
  - otros estadísticos que no sean medias ni funciones sencillas de medias
  - determinación de la distribución muestral completa

$$G\left(;\hat{F}\right)$$

El método de Montecarlo

F

↓

Modelo probabilístico, **completamente especificado**

p.e. n réplicas $N \quad \mu, \sigma^2 \quad iid$

$$\mathbf{x}_1 = (x_{11}, \ldots, x_{1n}) \mapsto U(\mathbf{x}_1) = u_1$$
$$\mathbf{x}_2 = (x_{21}, \ldots, x_{2n}) \mapsto U(\mathbf{x}_2) = u_2$$
$$\vdots$$
$$\mathbf{x}_m = (x_{m1}, \ldots, x_{mn}) \mapsto U(\mathbf{x}_m) = u_m$$

(gran) muestra de m valores del estadístico

"Leyes de los grandes números"

$$\frac{1}{m-1}\sum_{j=1}^{m}(u_j - \overline{u})^2 \cong \mathrm{var}_F \; U$$

Generación de m muestras independientes (o no) según F

$$\hat{G} \cong G_{;F}, \quad \text{etc.}$$

Bootstrap y Montecarlo

$\hat{F}$ estimación del Modelo probabilístico,

p.e. $P^*\left[X^* = x^*\right] = \begin{cases} \frac{1}{n} & \text{si } x^* \in x_1,\ldots,x_n \\ 0 & \text{en caso contrario} \end{cases}$

$$\mathbf{x}^*_1 = \left(x^*_{11},\ldots,x^*_{1n}\right) \mapsto U\left(\mathbf{x}^*_1\right) = u^*_1$$
$$\mathbf{x}^*_2 = \left(x^*_{21},\ldots,x^*_{2n}\right) \mapsto U\left(\mathbf{x}^*_2\right) = u^*_2$$
$$\vdots$$
$$\mathbf{x}^*_B = \left(x^*_{B1},\ldots,x^*_{Bn}\right) \mapsto U\left(\mathbf{x}^*_B\right) = u^*_B$$

muestra de B valores del estadístico

Generación de B "remuestras" de tamaño n (muestras aleatorias con reemplazo de los elementos de $\mathbf{x}$)

"Leyes de los grandes números"

$$\frac{1}{B-1}\sum_{b=1}^{B}(u^*_b - \bar{u}^*)^2 \cong \mathrm{var}_{\hat{F}}\ U^*$$

$$\hat{G}^* \cong G_{\hat{F}}, \quad \text{etc.}$$

Qué estimamos a partir del Montecarlo bootstrap?

Montecarlo bootstrap		Plug-in		"Verdadero" valor del funcional
$\hat{G}^* = \hat{G}\left(u_1^*, \ldots, u_B^*\right)$	$\cong$	$G\left(;\hat{F}\right)$	$\cong$	$G\left(;F\right)$
$\bar{u}_* = \frac{1}{B}\sum_{b=1}^{B} u_b^*$	$\cong$	$E_{\hat{F}}\left(U^*\right)$	$\cong$	$E_F\left(U\right)$
$\mathrm{var}_*\left(U^*\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(u_b^* - \bar{u}_*\right)^2$	$\cong$	$\mathrm{var}_{\hat{F}}\left(U^*\right)$	$\cong$	$\mathrm{var}_F\left(U\right)$
$\hat{P}_*\left[U^* \geq U(\mathbf{x})\right] = \frac{\#\left\{u_b^* \geq U(\mathbf{x})\right\}}{B}$	$\cong$	$P_{\hat{F}}\left[U^* \geq U(\mathbf{x})\right]$	$\cong$	$P_F\left[U \geq U(\mathbf{x})\right]$

Error de aproximación de Montecarlo

Problema "clásico" de precisión estadística

Validez de la aproximación bootstrap

- Resultado general (pero no muy útil):
  - Según Leyes de los grandes números, $F_n(x)$ tiende (en diversos sentidos) hacia $F(x)$. Extensible a funciones suficientemente "suaves"
- Validez: resultado sobre funcionales, funciones globales de F_n (u otras estimaciones) y de F: teoremas límite sobre distancias entre distribuciones
- Más interés práctico: comparación entre aproximación bootstrap y otras, para n finito

Características generales de los ejemplos

- Modelo probabilístico subyacente conocido
  - Normal $\mu = 15$, $\sigma = 3$, o bien
  - Exponencial $\alpha = 1/\mu = 1/15$

  ($\Rightarrow$ distribución muestral conocida)
- Análisis de única muestra (pequeña, *n* = 10), generada según uno u otro modelo.
  - caso normal: 15.54, 21.06, 16.52, 13.62, 16.14, 10.98, 13.53, 16.02, 16.79, 15.90
  - caso exponencial: 8.51, 8.71, 69.19, 10.05, 23.64, 8.67, 1.51, 20.36, 1.23, 5.27

Características generales de los ejemplos

- Estadístico bajo estudio: t
- aproximaciones: normal, bootstrap no paramétrico y bootstrap paramétrico
- aproximaciones bootstrap: estima "kernel" a partir de $B = 10000$ valores del estadístico (media o t, según el caso)
- Cada uno de estos valores calculado sobre una remuestra de tamaño $n = 10$

Estadístico t, caso normal: $n = 10$, $\mu = 15$, $\sigma = 3$

Verdadera distribución: $t \sim \mathbf{t}\ \ n-1=9$

Aproximación normal: $t \approx N\ \ 0{,}1$

Bootstrap: 1000 valores $t^* = t\ \ \mathbf{x}^*$

para remuestras $\mathbf{x}^* = \ \ x_1^*,\ldots,x_n^*$

no paramétrico: cada x_i^* escogido con probabilidad $1/n$ entre los de la muestra original

paramétrico: cada x_i^* generado según $N\ \ 15.62, 2.63$

Detalle y justificación del proceso de remuestreo

"Mundo real"

$$\mu = E(X, F)$$

$$F$$

$$\downarrow$$

$$\mathbf{x} = (x_1, \ldots, x_n)$$

$$\downarrow$$

$$\bar{x} = \bar{X}(\mathbf{x})$$

$$\hat{s} = \hat{S}(\mathbf{x}) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\downarrow$$

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{\hat{s}}$$

"Mundo bootstrap"

$$\bar{x} = \hat{\mu} = E(X^*, F_n)$$

$$F_n$$

$$\downarrow$$

$$\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$$

$$\downarrow$$

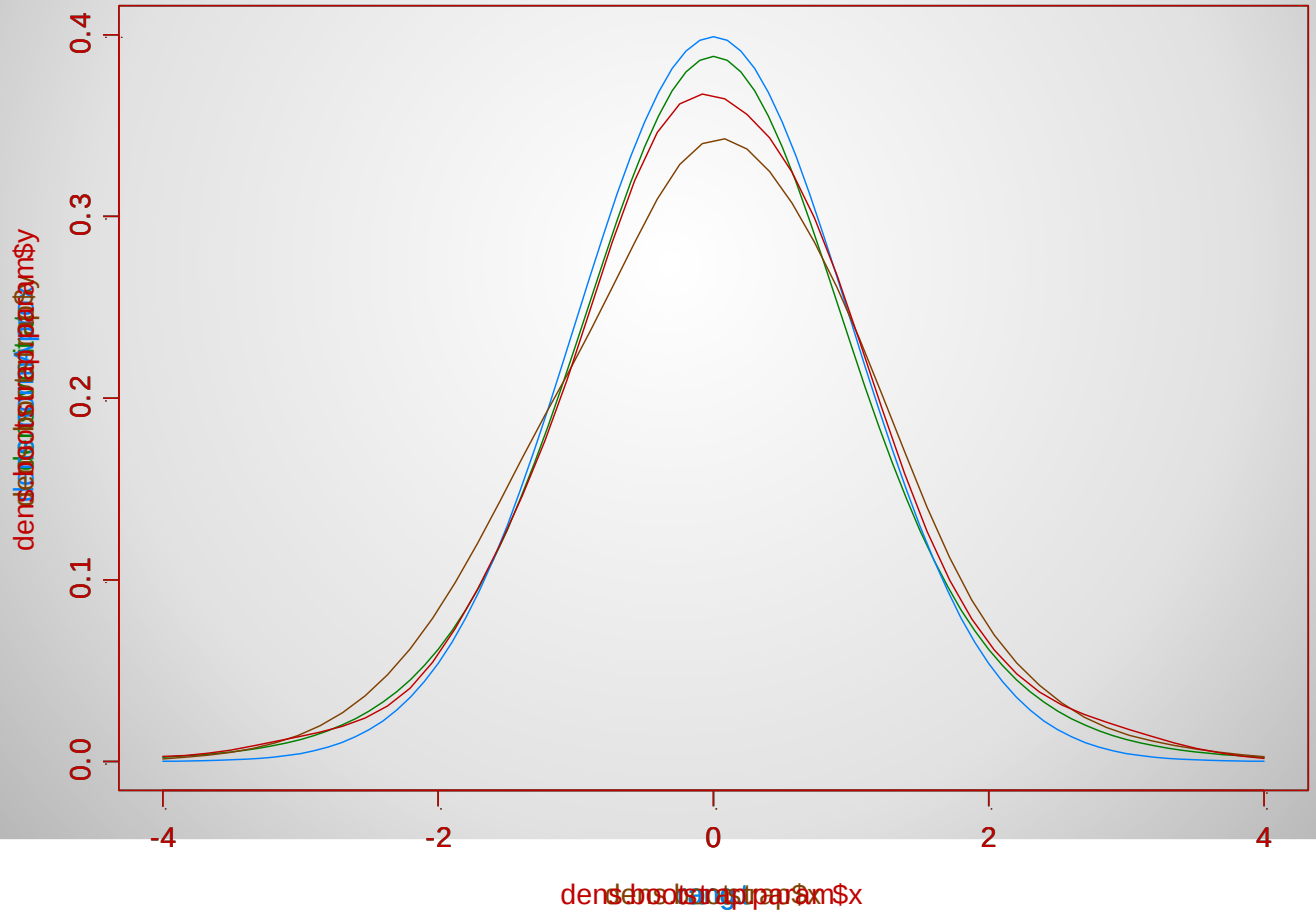
$$\bar{x}^* = \bar{X}(\mathbf{x}^*)$$

$$\hat{s}^* = \hat{S}(\mathbf{x}^*) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i^* - \bar{x}^*)^2}$$

$$\downarrow$$

$$t^* = \frac{\sqrt{n}(\bar{x}^* - \bar{x})}{\hat{s}^*}$$

Estadístico *t*, normal: verdadera densidad, aprox normal, bootstrap no paramétrico y paramétrico



Estadístico t, exponencial: $n = 10$, $\alpha = 1/\mu = 1/15$

Verdadera distribución:
estimada por simulación

Aproximación normal: $t \approx N\ \ 0,1$

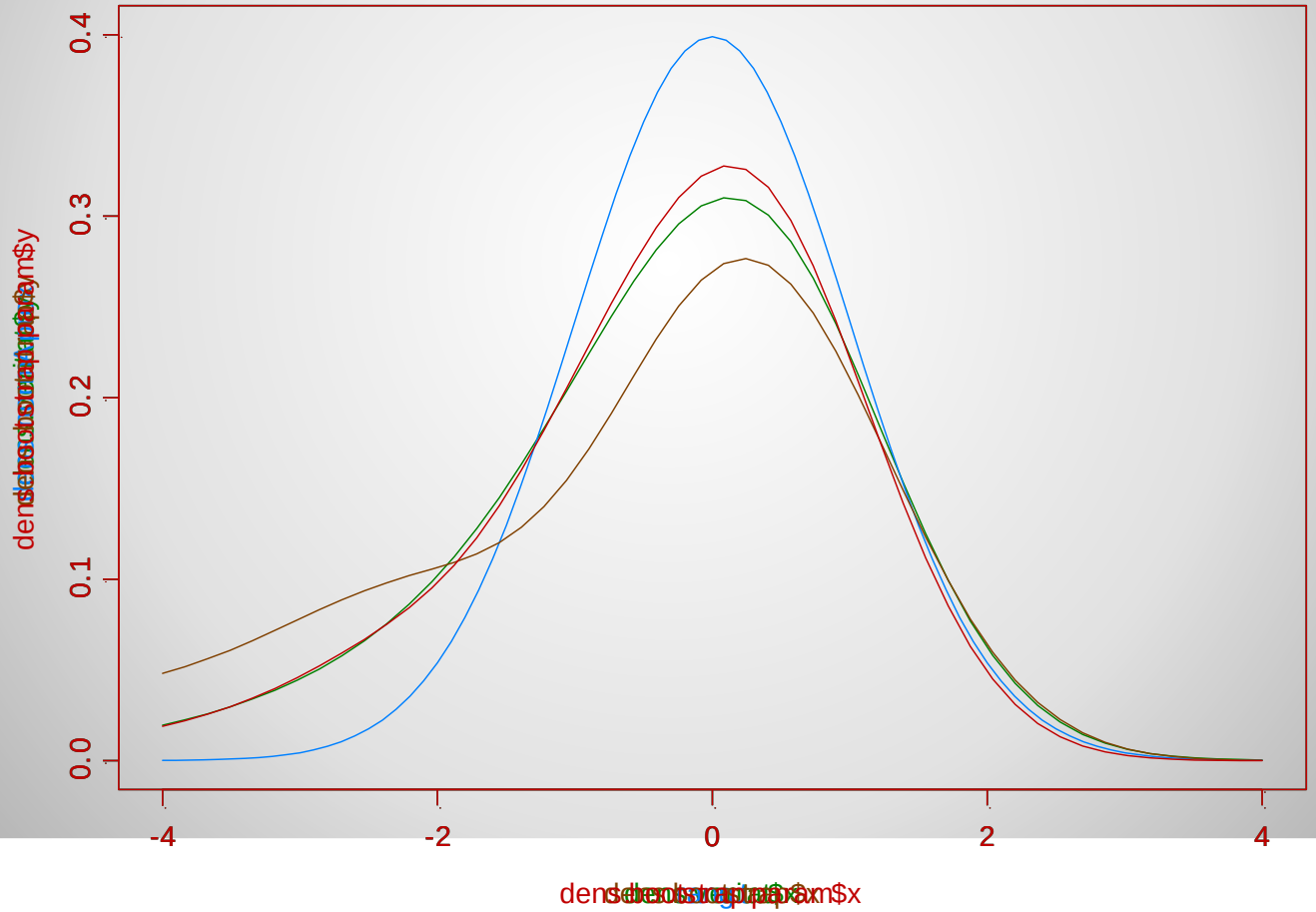
Bootstrap: 1000 valores $t^* = t\ \ \mathbf{x}^*$

para remuestras $\mathbf{x}^* = \ \ x_1^*,\ldots,x_n^*$

no paramétrico: cada x_i^* escogido con probabilidad $1/n$ entre los de la muestra original

paramétrico: cada x_i^* generado según $Exp\ \ 1/15.62$

Estadístico *t*, exponencial:
verdadera dens, aprox normal, boot no paramétrico y paramétrico



Caso exponencial, *t*, *n* = 40