Bootstrap

IECD

Revisitemos “La empírica”

X es una variable aleatoria con distribución acumulada $F : \mathbb{R} \to \mathbb{R}$ definida por

$$F(x) = P(X \leq x)$$

La empírica

Dada una muestra aleatoria $X_1, \ldots, X_n$ de la distribución F, su función de distribución empírica $\hat{F}_n$ se define como

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I\left(X_i \leq x\right)}{n}$$

- $\widehat{F}_n(x)$ es una función aleatoria
- Para cada realización $x_1, \ldots, x_n$ de la muestra $\widehat{F}_{n\,obs}(x)$ es una fda que asigna probabilidad $1/n$ a cada x_i.

Para cada x fijo

- $\mathbb{E}(\hat{F}_n(x)) =?$ ¿Es insesgado para $F(x)$?
- $\mathbb{V}(\hat{F}_n(x)) =?$
- ¿Es $\hat{F}_n(x)$ un estimador consistente de $F(x)$?

Para cada x fijo

- $\mathbb{E}(\hat{F}_n(x)) =?$ ¿Es insesgado para $F(x)$?
- $\mathbb{V}(\hat{F}_n(x)) =?$
- ¿Es $\hat{F}_n(x)$ un estimador consistente de $F(x)$?

Teorema (Glivenko - Cantelli) Sea $X_1, \ldots, X_n \sim F$. Entonces

$$\sup_x \left| \widehat{F}_n(x) - F(x) \right| \overset{c.t.p.}{\to} 0$$

Funcionales estadísticos

Un funcional estadístico es una función de F, $T(F)$

Ejemplos

- La esperanza: $F \mapsto \mathbb{E}_F(X) = \int x dF(x)$
- La varianza: $F \mapsto \mathbb{V}_F(X) = \int (x - \int x dF(x))^2 dF(x)$
- La mediana: $F \mapsto F^{-1}(1/2)$

El estimador plug-in

Consideremos $T = T(F)$.

El estimador plug-in de $T = T(F)$ se define como

$$\hat{T} = T(\hat{F}_n)$$

Ejemplo: estimador plug-in de la esperanza

Sea $X_1, \ldots, X_n \sim F$ y $T = T(F) = \mathbb{E}_F(X)$ entonces el estimador plug in de T es

$$\hat{T} = \mathbb{E}_{\hat{F}_n}(X).$$

Ejemplo: estimador plug-in de la esperanza

Sean $x_1, \ldots, x_n$ el conjunto de datos que resultan de las realizaciones de $X_1, \ldots, X_n \sim F$. Para calcular el estimador propuesto, pensemos ¿cuál es la esperanza de una variable aleatoria que toma los valores $x_1, \ldots, x_n$ con probabilidad $1/n$?

$$\sum_{x \in R_X} x p_X(x) = \sum_{i=1}^{n} x_i \frac{1}{n} = \overline{x}$$

Así,

$$\mathbb{E}_{\hat{F}_n}(X) = \overline{X}$$

Estimador plug-in de los cuantiles

Supongamos que F es continua y estrictamente creciente, el p-pésimo cuantil se define como

$T(F) = F^{-1}(p)$. Su estimador plug-in es

$$\hat{T} = T(\hat{F}_n) = \hat{F}_n^{-1}(p)$$

¡¡Pero cuidado!!! $\hat{F}_n$ no es inversible

Estimador plug-in de los cuantiles

Supongamos que F es continua y estrictamente creciente, el p-pésimo cuantil se define como

$T(F) = F^{-1}(p)$. Su estimador plug-in es

$$\hat{T} = T(\hat{F}_n) = \hat{F}_n^{-1}(p)$$

¡¡Pero cuidado!!! $\hat{F}_n$ no es inversible

Definimos la función cuantil

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

$\hat{F}_n^{-1}(p)$ se llama p-ésimo cuantil muestral.

Ejemplo: Ingreso total familiar en Argentina en 2019

(Fuente: INDEC, [www.indec.gob.ar])

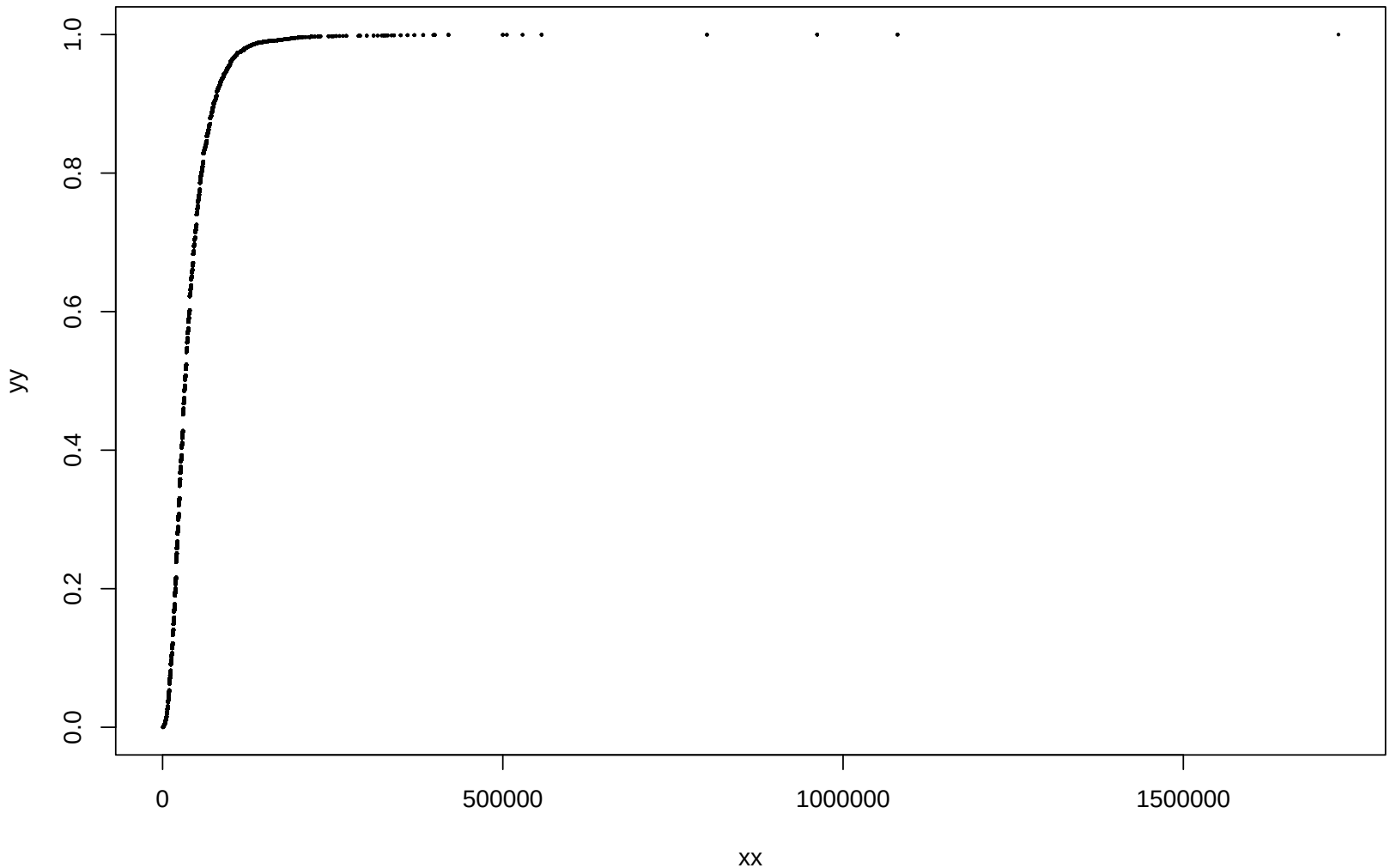
```
ingresos <- read.table("ingresos_argentina_2019.txt")
ingresos <- unlist(ingresos)
ingresos <- ingresos[ingresos>0]
length(ingresos)
```

```
## [1] 50286
```

Estimamos la fda a mano

```
f_acu <- function(x, datos){
  mean(datos <= x)
}
```

```
xx <- ingresos
yy <- 0
for(i in 1:length(xx)){
  yy[i] <- f_acu(xx[i], ingresos)
}
plot(xx,yy, cex = 0.2)
```


Función de distribución empírica del ITF en Argentina en 2019
yy
0.0
0.2
0.4
0.6
0.8
1.0
0
500000
1000000
1500000
xx

¿Qué proporción de familias argentinas tenían un ingreso entre 40000 y 50000 pesos en 2019?

```
f_acu(50000, ingresos) - f_acu(39999, ingresos)
```

```
## [1] 0.1383486
```

¿Qué proporción de familias argentinas tenían un ingreso mayor que 100000 en 2019?

```
1 - f_acu(100000, ingresos)
```

```
## [1] 0.03921569
```

Estimar el cuantil 0.9 del ingreso total familiar en Argentina en 2019

Lo hacemos a mano y por comando:

```
x <- seq(60000, 100000)
ims <- sapply(x, f_acu, datos = ingresos)
min(x[ims>=0.9])
```

```
## [1] 75000
```

```
quantile(ingresos, 0.9)
```

```
##   90%
## 75000
```

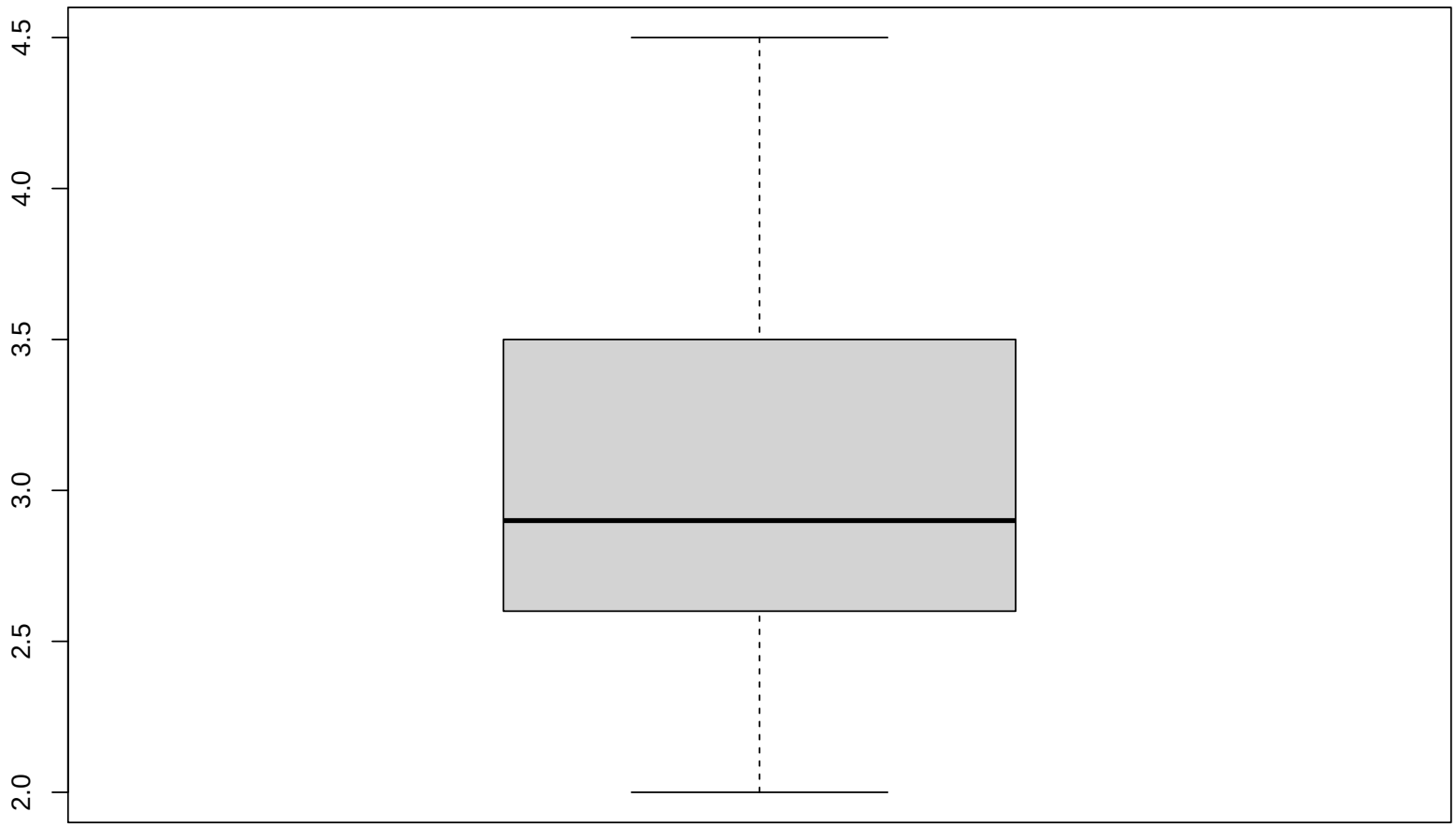
Ejemplo: Tiempo de Vida de Lámparas

Se quiere estudiar el tiempo de vida de ciertas lámparas. Para ello se observa en 30 de ellas el tiempo de vida en días:

```
tiempos
```

```
##
##  3.5 4.5 3.1 4.2 2.4 2.8 3.0 3.8 2.5 2.7
##  2.8 2.6 3.4 2.5 3.6 2.6 3.1 4.0 2.3 3.5
##  4.1 2.9 2.6 2.0 3.0 2.5 2.9 3.1 2.3 2.8
```

```r
boxplot(tiempos)
```



La mediana

¿Cuál es el tiempo de vida que es alzanzado por la mitad de las lámparas?

```
median(tiempos)
```

```
## [1] 2.9
```

¿Cuán preciso es el estimador? ¿Cómo estimamos su error estándar?

¿Cómo calculamos un intervalo de confianza para la mediana?

La mediana

¿Cuál es el tiempo de vida que es alzanzado por la mitad de las lámparas?

```
median(tiempos)
```

```
## [1] 2.9
```

¿Cuán preciso es el estimador? ¿Cómo estimamos su error estándar?

¿Cómo calculamos un intervalo de confianza para la mediana?

Para responder estas preguntas tenemos 2 problemas:

1 - No conocemos F

2 - Aunque conocieramos F, $\hat{\theta}$ podría ser una función complicada de $X_1, \ldots, X_n$ y podría ser difícil hallar su distribución o su desvío estándar.

Revisemos qué sabemos hacer

La media

a) Estimar el tiempo medio de vida mediante un intervalo de confianza asumiendo un modelo paramétrico.

b) Estimar el tiempo medio de vida mediante un intervalo de confianza asumiendo sólamente que la media y la varianza de los tiempos son finitas (sin asumir un modelo paramétrico).

Revisemos qué sabemos hacer

La media

a) Estimar el tiempo medio de vida mediante un intervalo de confianza asumiendo un modelo paramétrico.

b) Estimar el tiempo medio de vida mediante un intervalo de confianza asumiendo sólamente que la media y la varianza de los tiempos son finitas (sin asumir un modelo paramétrico).

La mediana

a) Estimar la mediana asumiendo un modelo paramétrico.

b) Estimar la mediana sin asumir ningún modelo.

- ¿Cómo estimar el error standard de un estimador sin asumir ningún modelo?

Revisemos qué sabemos hacer

La media

a) Estimar el tiempo medio de vida mediante un intervalo de confianza asumiendo un modelo paramétrico.

b) Estimar el tiempo medio de vida mediante un intervalo de confianza asumiendo sólamente que la media y la varianza de los tiempos son finitas (sin asumir un modelo paramétrico).

La mediana

a) Estimar la mediana asumiendo un modelo paramétrico.

b) Estimar la mediana sin asumir ningún modelo.

- ¿Cómo estimar el error standard de un estimador sin asumir ningún modelo? BOOTSTRAP!!

Supongamos que conocemos F

- Nos interesa $T = T(F)$, estimamos con $\hat{T}$.
- ¿Cómo hallamos la distribución de $\hat{T}$?

Supongamos que conocemos F

- Nos interesa $T = T(F)$, estimamos con $\hat{T}$.
- ¿Cómo hallamos la distribución de $\hat{T}$?
- Simulando

Supongamos que conocemos F

- Nos interesa $T = T(F)$, estimamos con $\hat{T}$.
- ¿Cómo hallamos la distribución de $\hat{T}$?
- Simulando

1 - Generamos muchos (digamos $B = 1000$) conjuntos de datos de tamaño $n = 30$. Para cada conjunto de datos calculamos el valor de $\hat{T}^*$.

2 - La distribución empírica de los valores resultantes:

$$\hat{T}_1^*, \hat{T}_2^*, \ldots, \hat{T}_B^*$$

es una aproximación a la distribución de $\hat{T}$.

¿Cuando F es desconocida?

Alternativa 1: Bootstrap paramétrico

- Nos interesa estudiar $\hat{T}$.
- Si estamos en condiciones de asumir un modelo paramétrico, por ejemplo que $X_i \sim F(\theta)$, podemos utilizar el **método bootstrap paramétrico**
- Tenemos una muestra: $X_1, \ldots, X_n$

¿Cuando F es desconocida?

Alternativa 1: Bootstrap paramétrico

- Nos interesa estudiar $\hat{T}$.
- Si estamos en condiciones de asumir un modelo paramétrico, por ejemplo que $X_i \sim F(\theta)$, podemos utilizar el **método bootstrap paramétrico**
- Tenemos una muestra: $X_1, \ldots, X_n$

Bootstrap paramétrico

1 - Computar $\hat{\theta}$ a partir de la muestra y luego estimamos con $\hat{T} = T(F(\hat{\theta}))$.

¿Cuando F es desconocida?

Alternativa 1: Bootstrap paramétrico

- Nos interesa estudiar $\hat{T}$.
- Si estamos en condiciones de asumir un modelo paramétrico, por ejemplo que $X_i \sim F(\theta)$, podemos utilizar el **método bootstrap paramétrico**
- Tenemos una muestra: $X_1, \ldots, X_n$

Bootstrap paramétrico

1 - Computar $\hat{\theta}$ a partir de la muestra y luego estimamos con $\hat{T} = T(F(\hat{\theta}))$.
2 - Generar B remuestras de tamaño n de $F(\hat{T})$. Para cada una de ellas calcular el estimador de interés: $\hat{T}^*$.

¿Cuando F es desconocida?

Alternativa 1: Bootstrap paramétrico

- Nos interesa estudiar $\hat{T}$.
- Si estamos en condiciones de asumir un modelo paramétrico, por ejemplo que $X_i \sim F(\theta)$, podemos utilizar el **método bootstrap paramétrico**
- Tenemos una muestra: $X_1, \ldots, X_n$

Bootstrap paramétrico

1 - Computar $\hat{\theta}$ a partir de la muestra y luego estimamos con $\hat{T} = T(F(\hat{\theta}))$.
2 - Generar B remuestras de tamaño n de $F(\hat{T})$. Para cada una de ellas calcular el estimador de interés: $\hat{T}^*$.

3 - La distribución empírica de los valores resultantes: $\hat{T}_1^*, \hat{T}_2^*, \ldots, \hat{T}_B^*$ es una aproximación a la distribución de $\hat{T}$.

En nuestro ejemplo: lámparas

- Nos interesa la mediana, usamos la mediana muestral $\hat{T}$.
- Asumimos un modelo paramétrico, por ejemplo, $X_i \sim \Gamma(\alpha, \lambda)$.
- Tenemos una muestra: $X_1, \ldots, X_n$

Bootstrap paramétrico

1 - Estimar los parámetros usando la muestra: $\rightarrow \hat{\alpha}$ y $\hat{\lambda}$.

2 - Generar B remuestras de tamaño n de $\Gamma(\hat{\alpha}, \hat{\lambda})$:
$\mathbf{X}^* = X_1^*, \ldots X_n^*$. Para cada una de ellas estimar la mediana por la mediana muestral: $\hat{T}^*$

3 - La distribución empírica de los valores resultantes:
$\hat{T}_1^*, \hat{T}_2^*, \ldots, \hat{T}_B^*$ es una aproximación a la distribución de $\hat{T}$.

Ejemplo: bootstrap paramétrico

Supongamos que $X_1, \ldots, X_{30} \sim \Gamma(\alpha, \lambda)$. Queremos conocer el error estandar del estimador de la mediana.

```
library(MASS)
est <- fitdistr(tiempos, densfun = "gamma")$estimate
#estimador mediana
qgamma(0.5, shape = est[1],rate = est[2])
```

```
## [1] 2.997314
```

```
B <- 1000; estmed <-  0
for(i in 1:B){
  xboot <- rgamma(n, shape = est[1], rate = est[2])
  estboot <- fitdistr(xboot, densfun = "gamma")
  estshape <- estboot$estimate[1];  estrate <- estboot$estimate[
  estmed[i] <- qgamma(0.5, shape = estshape,rate = estrate)
}
```

Podríamos calcular el desvío estándar:

```
sd(estmed)
```

```
## [1] 0.1083144
```

¿Cuando F es desconocida?

Alternativa 2: Bootstrap uniforme

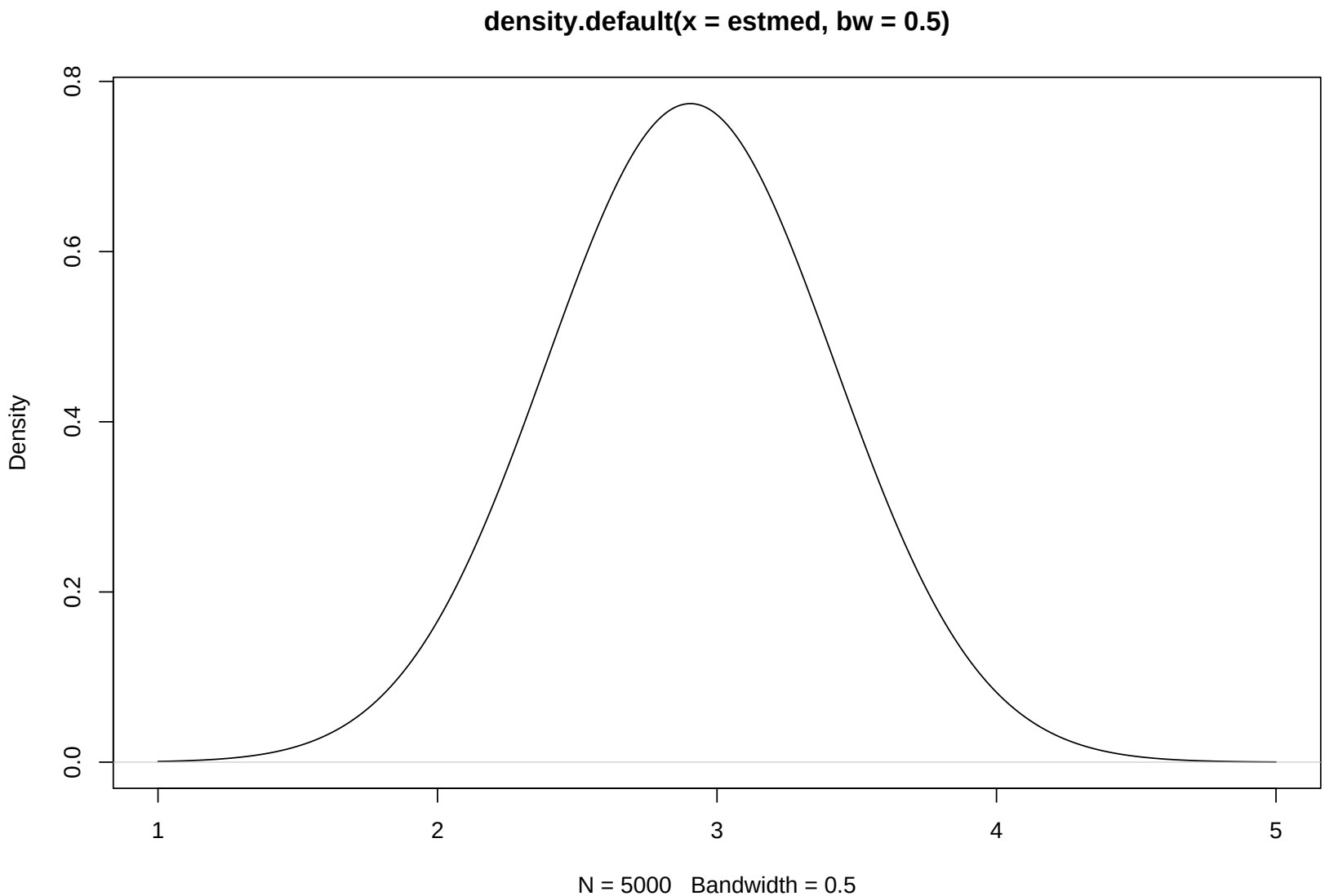
- Supongamos que tenemos un conjunto de datos $x_1, \ldots, x_n$. Asumimos que los datos son una realización de una muestra $\mathbf{X} = X_1, \ldots, X_n$.
- Sabiendo que la empírica $\hat{F}_n$ es un estimador consistente de la distribución F, la solución bootstrap es generar muestras con distribución $\hat{F}_n(x)$.
- Una muestra aleatoria de tamaño n de la distribución $\hat{F}_n$ es entonces una muestra de tamaño n tomada **con reposición** de $X_1, \ldots, X_n$.

Ejemplo: lámparas

```
mediana<- median(tiempos)
B <- 5000
n <- length(tiempos)
estmed <- 0
for(i in 1:B){
  xboot <- sample(tiempos, n, replace = TRUE )
  estmed[i] <- median(xboot)
}
sd(estmed)
```

```
## [1] 0.1265239
```

```
plot(density(estmed, bw=0.5))
```



Resumiendo:

- El Bootstrap nos dio un camino para estimar el desvío estándar de un estimador.
- Dado el estimador $\hat{T}_n = \delta(X_1, \ldots, X_n)$, la idea del bootstrap es:

Paso 1: Estimar $\mathbb{V}ar_F(\hat{T}_n)$ con $\mathbb{V}ar_{\hat{F}_n}(\hat{T}_n)$.

Paso 2: Aproximar $\mathbb{V}ar_{\hat{F}_n}(\hat{T}_n)$ por simulación.

- Veremos que también permite calcular intervalos de confianza.

Muestra bootstrap no paramétrico

Sea $\hat{F}_n$ la función de distribución empírica, que asigna probabilidad $1/n$ a cada observación X_i.

Una **muestra bootstrap** es una muestra aleatoria de la distribución empírica $\hat{F}_n$.

¿Cómo obtenemos una muestra bootstrap

$$\mathbf{X}^* = X_1^*, \ldots X_n^*$$

con esta alternativa?

Muestra bootstrap no paramétrico

Sea $\hat{F}_n$ la función de distribución empírica, que asigna probabilidad $1/n$ a cada observación X_i.

Una **muestra bootstrap** es una muestra aleatoria de la distribución empírica $\hat{F}_n$.

¿Cómo obtenemos una muestra bootstrap

$$\mathbf{X}^* = X_1^*, \ldots X_n^*$$

con esta alternativa?

Extrayendo una muestra con reposición de la muestra original $X_1, \ldots, X_n$: $X_1^*, \ldots X_n^*$

Muestra bootstrap no paramétrico

Sea $\hat{F}_n$ la función de distribución empírica, que asigna probabilidad $1/n$ a cada observación X_i.

Una **muestra bootstrap** es una muestra aleatoria de la distribución empírica $\hat{F}_n$.

¿Cómo obtenemos una muestra bootstrap

$$\mathbf{X}^* = X_1^*, \ldots X_n^*$$

con esta alternativa?

Extrayendo una muestra con reposición de la muestra original $X_1, \ldots, X_n$: $X_1^*, \ldots X_n^*$

Para cada muestra bootstrap hay una replicación bootstrap de $\hat{T}$: $\hat{T}^*$ estadístico bootstrap

Diagrama (Wasserman, 2003)

$$
\begin{array}{lccccc}
\text{Mundo real} & F & \Longrightarrow & X_1, \ldots, X_n & \Longrightarrow & \hat{T} \\
\text{Mundo bootstrap} & \widehat{F}_n & \Longrightarrow & X_1^*, \ldots, X_n^* & \Longrightarrow & \hat{T}^*
\end{array}
$$

El algoritmo bootstrap

Este algoritmo permite estimar el error estándar de un estimador $\hat{T}_n = \delta\left(X_1, \ldots, X_n\right)$

1 - Generar $X_1^*, \ldots, X_n^*$ a partir de $\hat{F}_n$.

2 - Calcular $\hat{T}_n^* = \delta\left(X_1^*, \ldots, X_n^*\right)$.

3 - Repetir los pasos 1 y 2, B veces, para obtener $\hat{T}_{n,1}^*, \ldots, \hat{T}_{n,B}^*$.

4 - Calculamos

$$\widehat{\mathsf{se}}_{\mathsf{boot}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{T}_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}\hat{T}_{n,r}^*\right)^2}$$

Método 1. Intervalo bootstrap normal

- $\hat{T}_n$ asintóticamente normal si

$$\frac{\hat{T}_n - T(F)}{\text{se}} \approx \mathcal{N}(0,1)$$

con $\text{se} = \text{se}(\hat{T}_n)$

- Sea $\widehat{\text{se}}_{\text{boot}}$ el estimador bootstrap de $\text{se}(\hat{T}_n)$

intervalo bootstrap normal nivel aproximado $1-\alpha$:

$$\hat{T}_n \pm z_{\alpha/2}\, \widehat{\text{se}}_{\text{boot}}$$

Ejemplo

```
ingresos<- read.csv("ingresos_argentina_2019.txt", sep="")
ingresos <- unlist(ingresos)
```

Fuente: Indec

Objetivo 1: Estimar el ingreso familiar medio en Argentina en 2019 mediante un intervalo de confianza.

Método 1: IC asintótico basado en el TCL

$$\left[\overline{X} - 1.96\frac{s}{\sqrt{n}}, \overline{X} + 1.96\frac{s}{\sqrt{n}}\right]$$

```
x <- ingresos
n <- length(x)
mean(x)
```

```
## [1] 34750.76
```

```
sd(x)
```

```
## [1] 35907.18
```

```
intervalo <- c(mean(x) - 1.96 * sd(x)/sqrt(n),
               mean(x) + 1.96 * sd(x)/sqrt(n))
intervalo
```

```
## [1] 34461.65 35039.87
```

Objetivo 2: Estimar el ingreso familiar mediano en Argentina en 2019 mediante un intervalo de confianza.

```
median(ingresos)
```

```
## [1] 29000
```

Estimación bootstrap del error estándar de la mediana

```r
estimar_se_mediana <-
  function(x, B = 1000){
    titahatboot <- rep(0, B)
    for(i in 1:B){
      Xboot <- sample(x, length(x), replace = TRUE)
      titahatboot[i] <- median(Xboot)
    }
    sqrt(mean((titahatboot-mean(titahatboot))^2))
  }
```

IC para la mediana por bootstrap uniforme

Intervalo bootsrap normal

```
x <- ingresos
se_boot <- estimar_se_mediana(x)
se_boot
```

```
## [1] 186.2835
```

```
intervalo_boot <- c(median(x) - 1.96*se_boot,
                    median(x) + 1.96*se_boot)
intervalo_boot
```

```
## [1] 28634.88 29365.12
```

Método 2: Intervalo bootstrap percentil

- B: número de iteraciones bootstrap
- Sean $\hat{T}_1^*, \ldots, \hat{T}_B^*$ estadísticos bootstrap del estimador.

$$\text{intervalo boot percentil } 1-\alpha: \quad \left(\hat{T}^*_{\alpha/2}\ ,\ \hat{T}^*_{1-\alpha/2}\right)$$

donde $\hat{T}^*_\beta$ es el $\beta-$cuantil muestral de los estadísticos bootstrap.