Tema 4. Estimación de errores estándar mediante remuestreo

basado en

B. Efron, R. Tibshirani (1993). An Introduction to the bootstrap.

O. Kirchkamp (2019). Resampling methods.

Curso 2023/2024

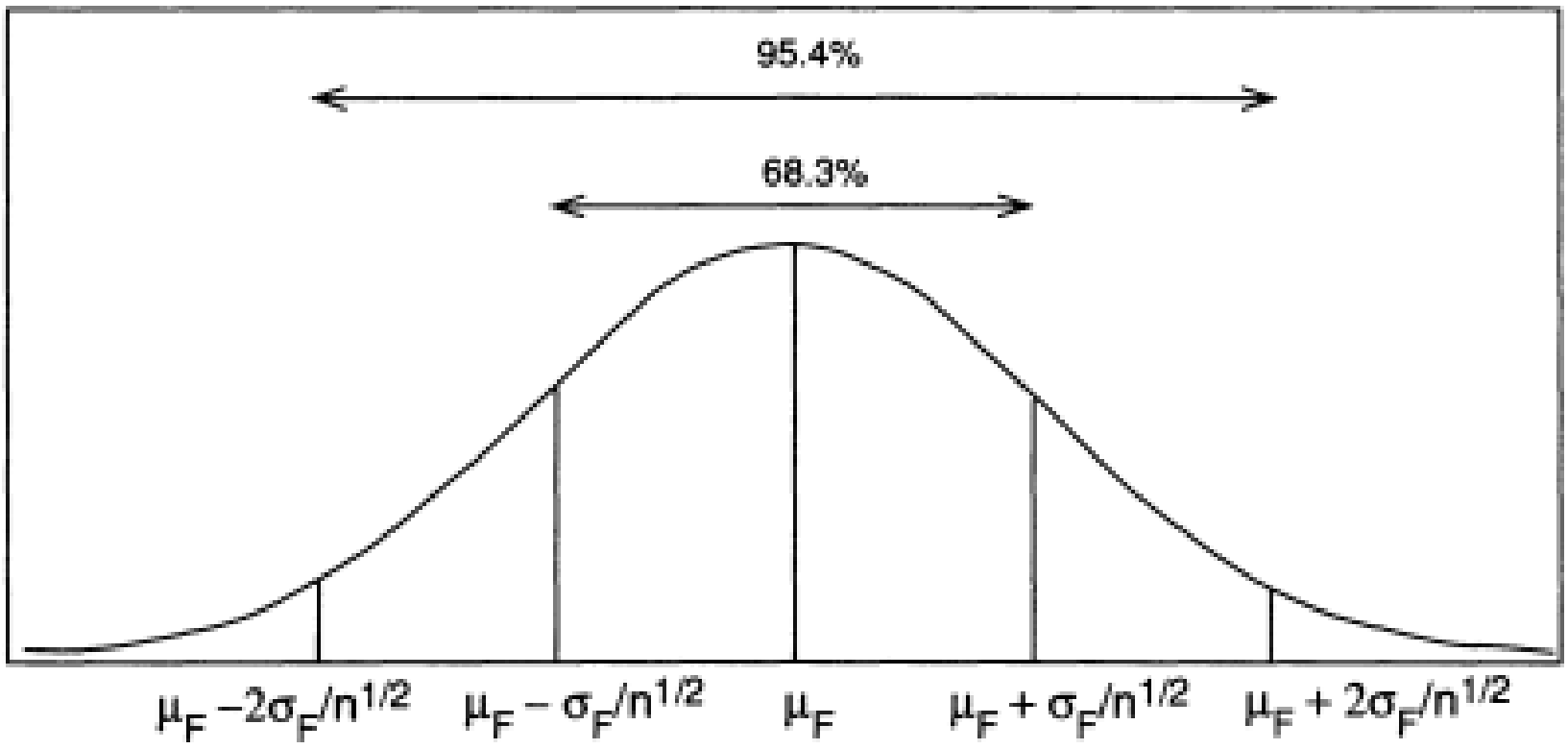
Errores estándar

- Los estadísticos muestrales se usan frecuentemente en Estadística, por ello es necesario conocer su precisión.
- El bootstrap permite encontrar el error estándar de los estadísticos basándose en el principio de *plug-in*.
- Supongamos una v.a. X con media μ_F y varianza σ_F^2
- Sea $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ una m.a.s procedente de la distribución F.
- La media muestral

$$\bar{x} \sim \left(\mu_F, \frac{\sigma_F^2}{n}\right)$$

Cálculo de errores estándar mediante el TCL

- De este modo la esperanza de $\bar{x}$ es la misma que la de la v.a. X original, pero la varianza es igual a $1/n$ veces la varianza de X.
- Así el error estándar de $\bar{x}$ es simplemente $\frac{\sigma_F}{\sqrt{n}}$.
- En una distribución normal se espera que X sea menor que una vez la desviación estándar de μ_F aproximadamente el 68 % de las ocasiones y menor que dos desviaciones estándar alrededor del 95 % de las veces, aplicando el *TCL* (teorema central del límite).


95.4%
68.3%
$\mu_F - 2\sigma_F/n^{1/2}$
$\mu_F - \sigma_F/n^{1/2}$
μ_F
$\mu_F + \sigma_F/n^{1/2}$
$\mu_F + 2\sigma_F/n^{1/2}$

Limitaciones del TCL

- La aproximación del TCL funciona bien cuando el tamaño muestral n es grande pero la aproximación tiene limitaciones.

- Supongamos que X sigue una distribución de Bernoulli:

  $P_F\{X = 1\} = p$
  $P_F\{X = 0\} = 1 - p$

- El parámetro p es la probabilidad de éxito que está entre 0 y 1

Limitaciones del TCL

- Una m.a.s. es una sucesión de unos y ceros de modo que la suma

$$s = \sum_{i=1}^{n} x_i \sim \text{Bin}(n, p)$$

- La media $\overline{x} = \dfrac{s}{n}$ es igual a $\widehat{p}$ que es el estimador *plug-in* de p, de modo que

$$\widehat{p} \sim N\left(p,\ \frac{p(1-p)}{n}\right)$$

Limitaciones del TCL

- Se toma el ejemplo de una distribución binomial con $n = 25$ en los casos de $p = 0{,}25$ y $p = 0{,}9$.

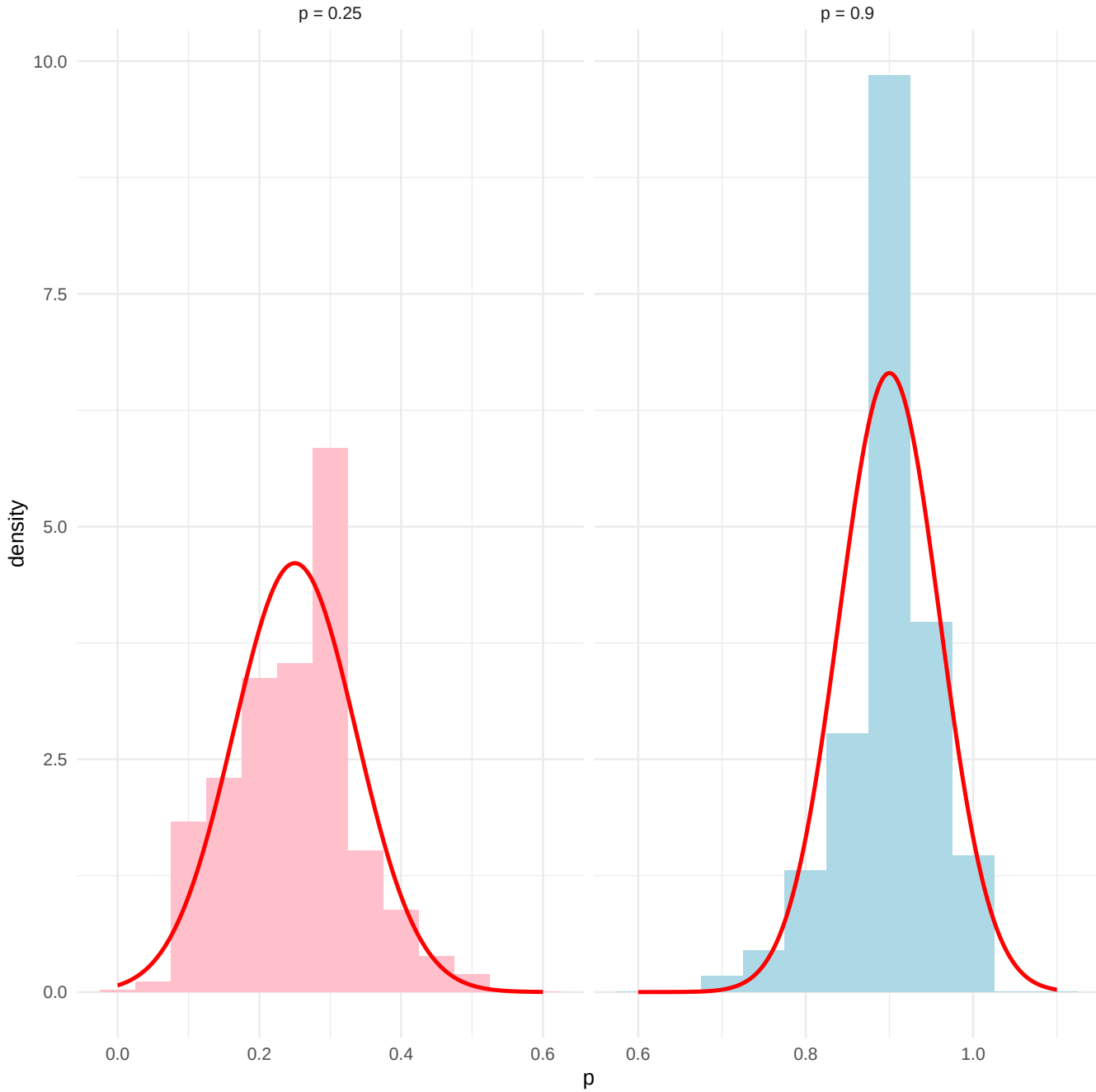
- Para el caso de $p = 0{,}9$ la aproximación a la normal por el TCL **no** es muy buena.

```
n = 25

p09 = rbinom(20000, n, 0.9)/n
p025 = rbinom(20000, n, 0.25)/n

head(p09)
```

```
[1] 0.88 0.88 0.92 0.88 0.96 0.88
```


p = 0.25
p = 0.9
10.0
7.5
5.0
2.5
0.0
density
0.0
0.2
0.4
0.6
0.6
0.8
1.0
p

Bootstrap y errores estándar

- El bootstrap permite calcular errores estándar sin que tenga importancia lo complicado que sea el estimador que se considere.
- Los métodos bootstrap dependen del concepto de *muestra bootstrap*.
- Partimos de la función de distribución empírica $\widehat{F}$ que asigna probabilidad $1/n$ a cada uno de los elementos de la muestra observada.
- Una muestra bootstrap se define como una muestra aleatoria de tamaño n extraída de $\widehat{F}$

$$\widehat{F} \rightarrow \mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$$

- La notación *estrella* * indica que $\mathbf{x}^*$ no es el conjunto de datos original sino una versión *remuestreada* de la muestra original $\mathbf{x}$.

Bootstrap y errores estándar

- Alternativamente, se puede decir que una muestra bootstrap $x_1^*, x_2^*, \ldots x_n^*$ es una muestra aleatoria de tamaño n tomada **con reemplazamiento** de la muestra original (que hace el papel de *población*).
- El algoritmo se denomina **bootstrap no paramétrico** porque depende solo de la función de distribución empírica.
- Por ejemplo podríamos tener una muestra bootstrap como

$$\begin{aligned} x_1^* &= x_7 \\ x_2^* &= x_3 \\ x_3^* &= x_3 \\ \ldots &\quad \ldots \quad \ldots \\ x_n^* &= x_2 \end{aligned}$$

Algoritmo Bootstrap y errores estándar

(I) Seleccionar B muestras bootstrap $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ cada una obtenida a partir de la muestra original $\mathbf{x}$ con reemplazamiento.

(II) Evaluar la réplica bootstrap en el estimador correspondiente

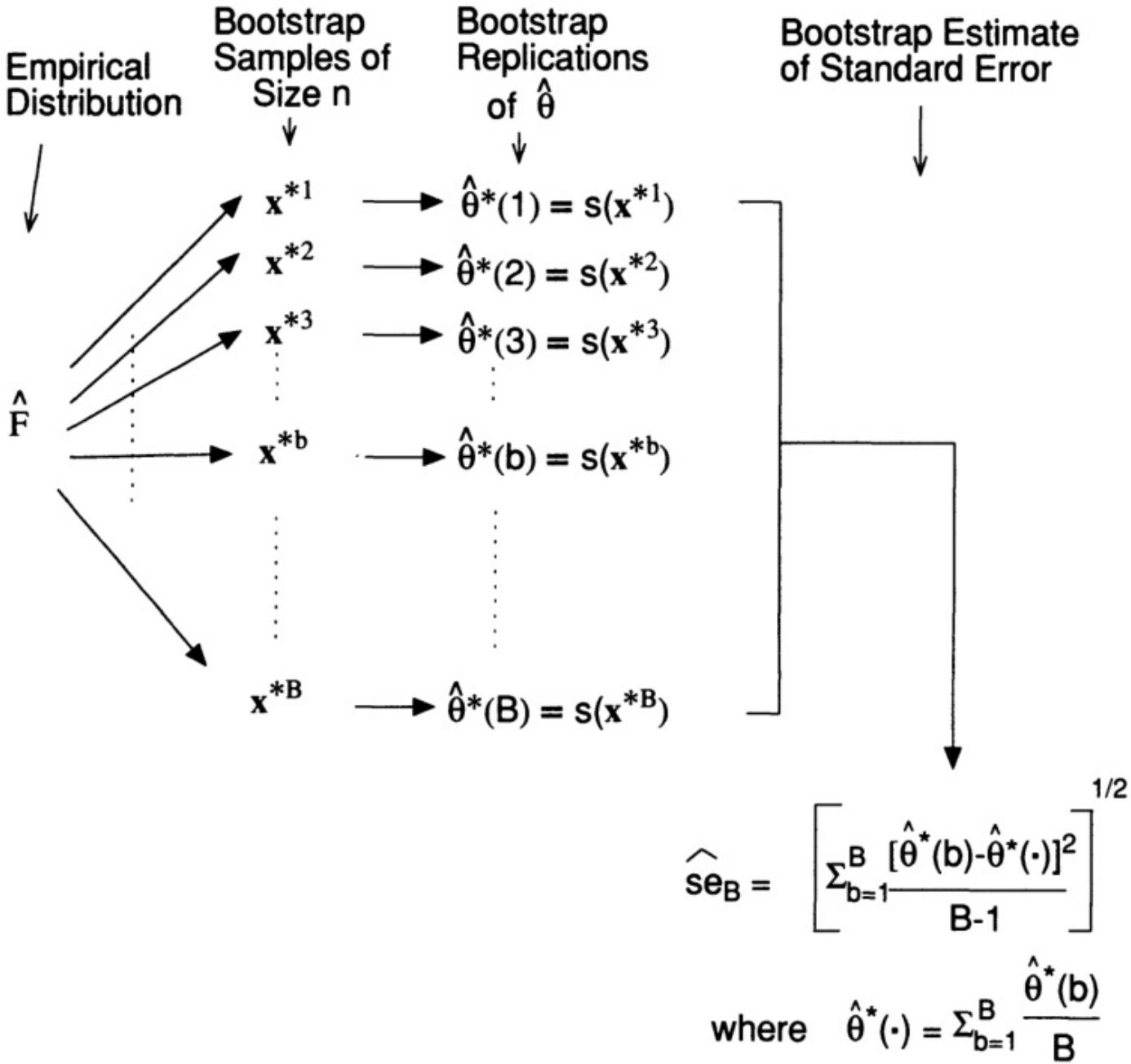
$$\widehat{\theta}^*(b) = s(\mathbf{x}^{*b})$$

para $b = 1, 2, \ldots, B$.

(III) Estimar el error estándar $se_F(\widehat{\theta})$ mediante

$$\widehat{se}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}^*(b) - \widehat{\theta}^*(\cdot)\right)^2}$$

donde $\widehat{\theta}^*(\cdot) = \dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\widehat{\theta}^*(b)$


Empirical Distribution
Bootstrap Samples of Size n
Bootstrap Replications of $\hat{\theta}$
Bootstrap Estimate of Standard Error
$\hat{F}$
x^{*1}
x^{*2}
x^{*3}
x^{*b}
x^{*B}
$\hat{\theta}^*(1) = s(x^{*1})$
$\hat{\theta}^*(2) = s(x^{*2})$
$\hat{\theta}^*(3) = s(x^{*3})$
$\hat{\theta}^*(b) = s(x^{*b})$
$\hat{\theta}^*(B) = s(x^{*B})$
$\widehat{se}_B = \left[\Sigma_{b=1}^{B} \frac{[\hat{\theta}^*(b)-\hat{\theta}^*(\cdot)]^2}{B-1}\right]^{1/2}$
where $\hat{\theta}^*(\cdot) = \Sigma_{b=1}^{B} \frac{\hat{\theta}^*(b)}{B}$

Ejemplo de los institutos de máster en leyes

- La correlación entre GPA y LSAT es

```
library(bootstrap)
(lawCor = with(law,cor(GPA,LSAT)))
```

```
[1] 0.7763745
```

- ¿Cómo es de preciso el estimador del coeficiente de correlación lineal?
- Si la distribución conjunta de ambas variables F es normal bivariante, entonces $\hat{\rho}$ (siguiendo a Efron&Tibshirani) tiene un error estándar igual a

$$\widehat{\sigma}_{\widehat{\rho}} = \frac{1-\widehat{\rho}^2}{\sqrt{n-3}} \approx 0{,}115$$

Ejemplo de los institutos de máster en leyes

- Usando bootstrap se puede **evitar** asumir que F se distribuye como una normal bivariante.

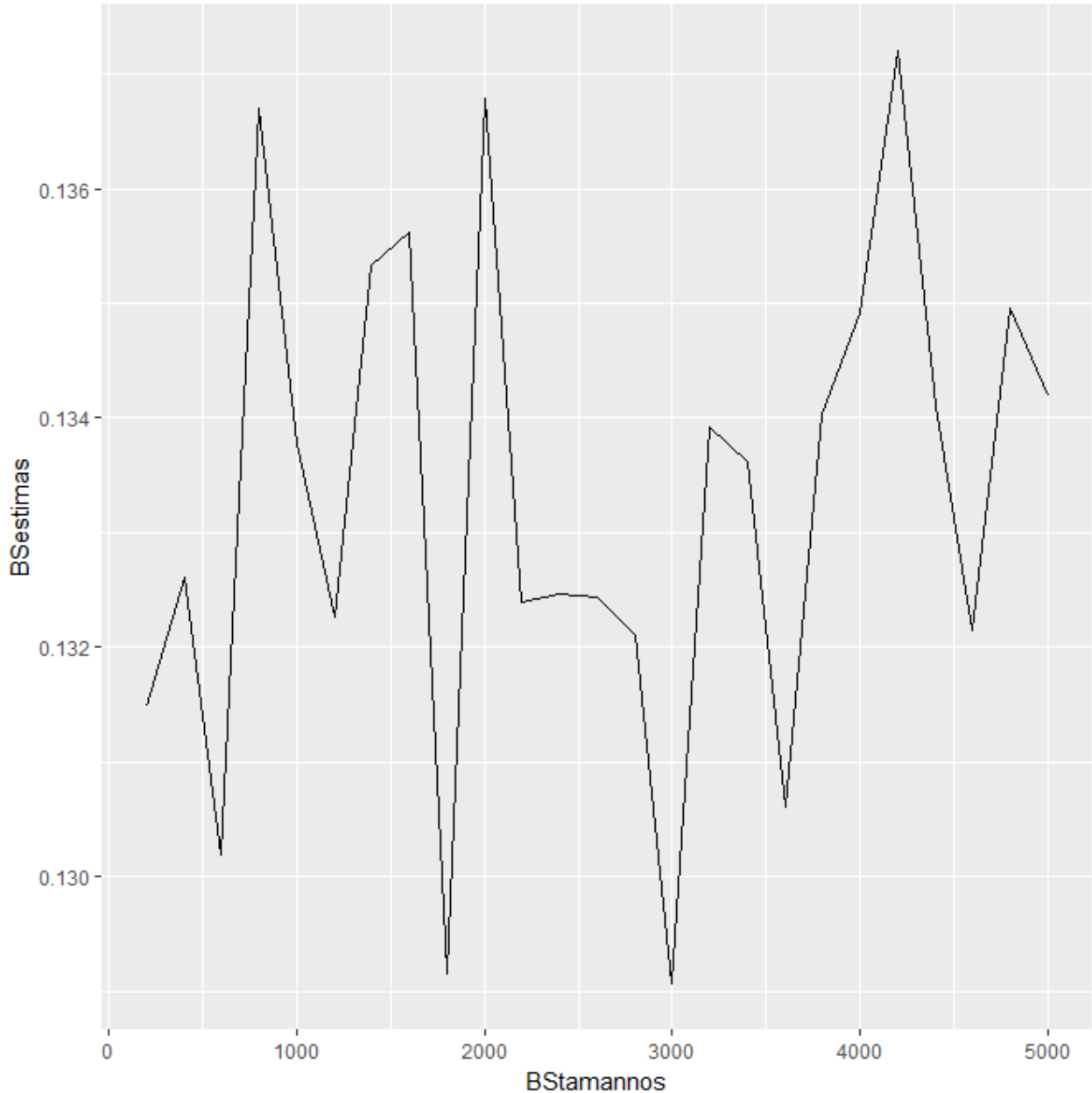
```
ssamplesize = dim(law)[1]
ind = 1:samplesize

law.boot =
replicate(1000, {indB = sample(ind,replace=TRUE);
with(law[indB,], cor(GPA,LSAT))})

sd(law.boot)
```

```
[1] 0.1336493
```

- ¿Cómo converge de rápido el estimador bootstrap?


0.136
0.134
0.132
0.130
BSestimas
0
1000
2000
3000
4000
5000
BStamannos

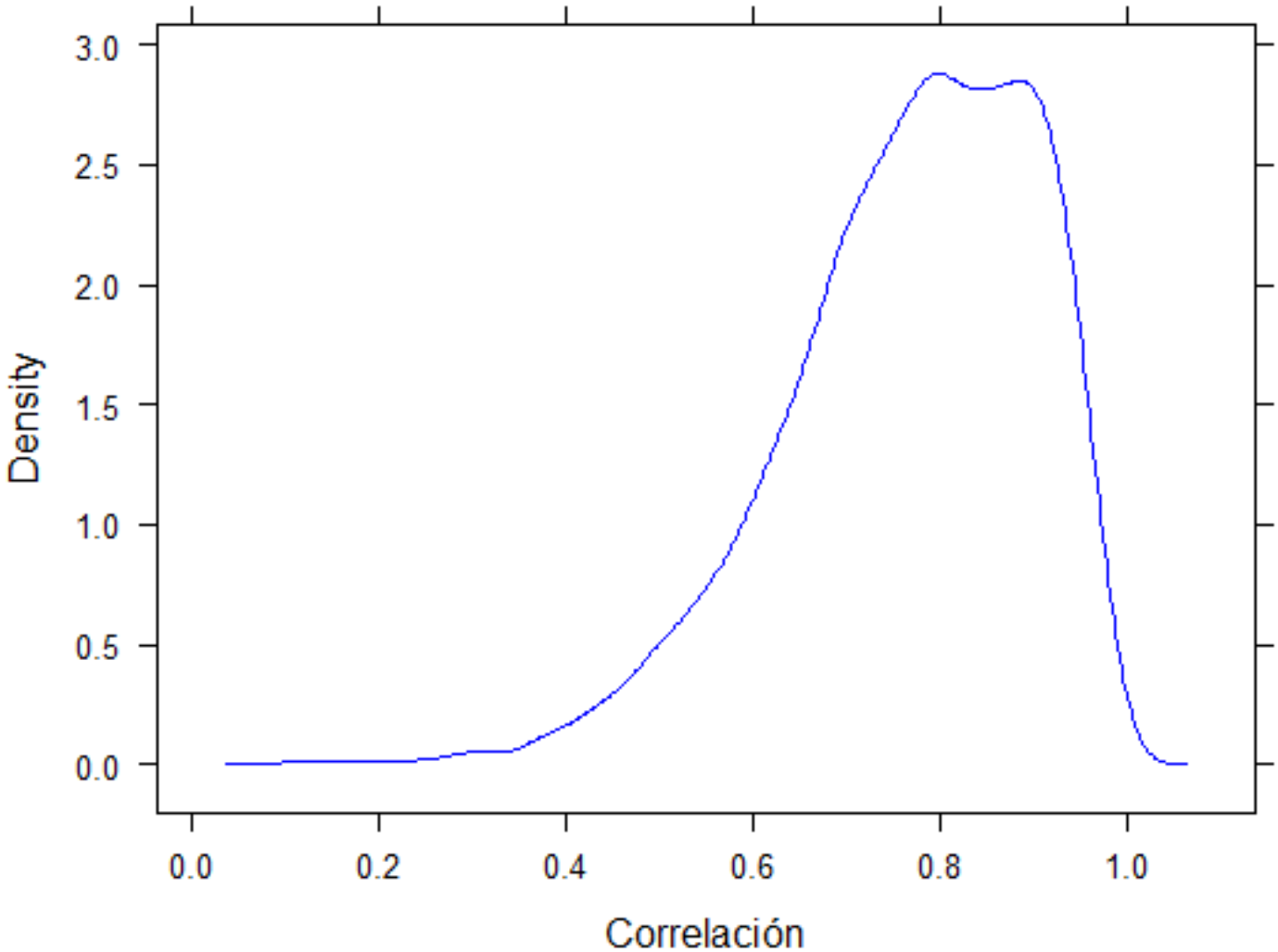


Figura: Densidad de la distribución de $\widehat{\theta}^*$

Bootstrap Paramétrico

- En muchas ocasiones se tienen fórmulas analíticas para calcular los errores. En este caso se puede aplicar el bootstrap aprovechando que éstas se conocen.

- Se denomina a este tipo de remuestreo como ***bootstrap paramétrico*** y se define el estimador bootstrap del error estándar como

$$se_{\widehat{F}_{par}}(\widehat{\theta}^*)$$

- donde $\widehat{F}_{par}$ es un estimador de F que se obtiene a partir de un modelo paramétrico aplicado a los datos.

Bootstrap Paramétrico

- En el ejemplo `law82`, en lugar de estimar la función de distribución F mediante la función de distribución empírica, se puede asumir que la población se distribuye como una normal bivariante.

- Para la media y la matriz de covarianzas de esta distribución, los estimadores razonables serían respectivamente

$$(\bar{x}, \bar{y})$$

$$\frac{1}{14}\begin{pmatrix} \sum_i (x_i - \bar{x})^2 & \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_i (x_i - \bar{x})(y_i - \bar{y}) & \sum_i (y_i - \bar{y})^2 \end{pmatrix}$$

- Se denota a la población normal bivariante que se obtiene con esta media y matriz de covarianzas como $\widehat{F}_{par}$.

Bootstrap Paramétrico

- Se denomina al estimador bootstrap paramétrico del error estándar del parámetro como $se_{\hat{F}_{par}}(\widehat{\theta}^*)$.

- En lugar de muestrear con reemplazamiento a partir de los datos originales, se sacan B muestras de tamaño n del estimador paramétrico de la población $\widehat{F}_{par}$

$$\widehat{F}_{par} \rightarrow (x_1^*, x_2^*, \ldots, x_n^*)$$

- Posteriormente se siguen los mismos pasos 2 y 3 del algoritmo general del bootstrap *no paramétrico*: se calcula el correspondiente estadístico en cada muestra bootstrap y luego se calcula la desviación estándar de las B réplicas.

Ejemplo de los centros de estudios de máster

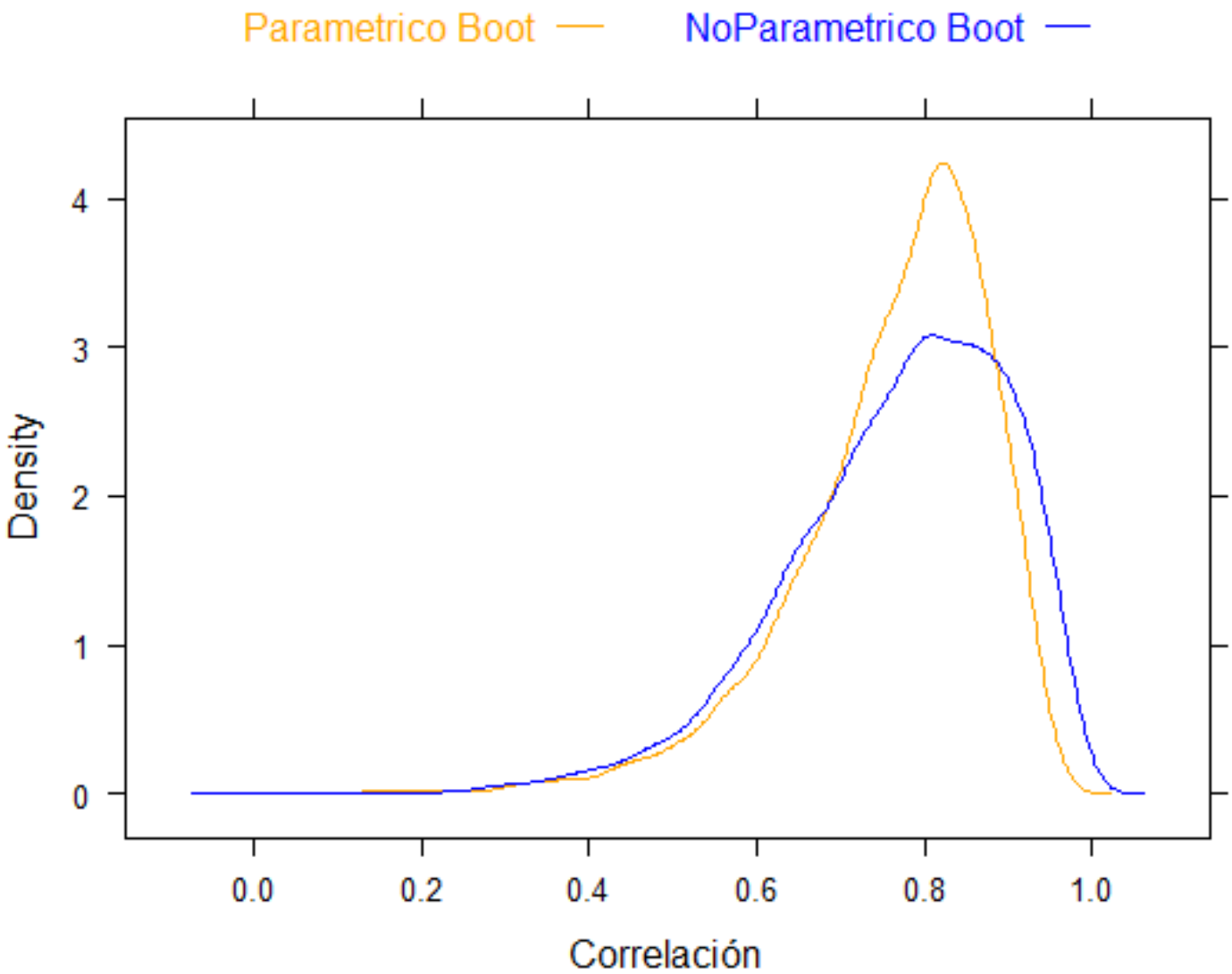
- En el ejemplo de los datos de los centros de estudios de máster en leyes se tiene que si (x, y) se distribuyen como una normal bivariante entonces se pueden generar observaciones de este vector, definiendo

$$\begin{aligned} x &= \mu_x + \sigma_x z_1 \\ y &= \mu_y + \sigma_y \frac{z_1 + c \cdot z_2}{\sqrt{1 + c^2}} \end{aligned}$$

donde $z_1, z_2 \sim N(0, 1)$

$$c = \sqrt{\frac{\sigma_x^2 \sigma_y^2}{\sigma_{xy}^2} - 1}$$

Ejemplo de los institutos de máster en leyes



Bootstrap Paramétrico

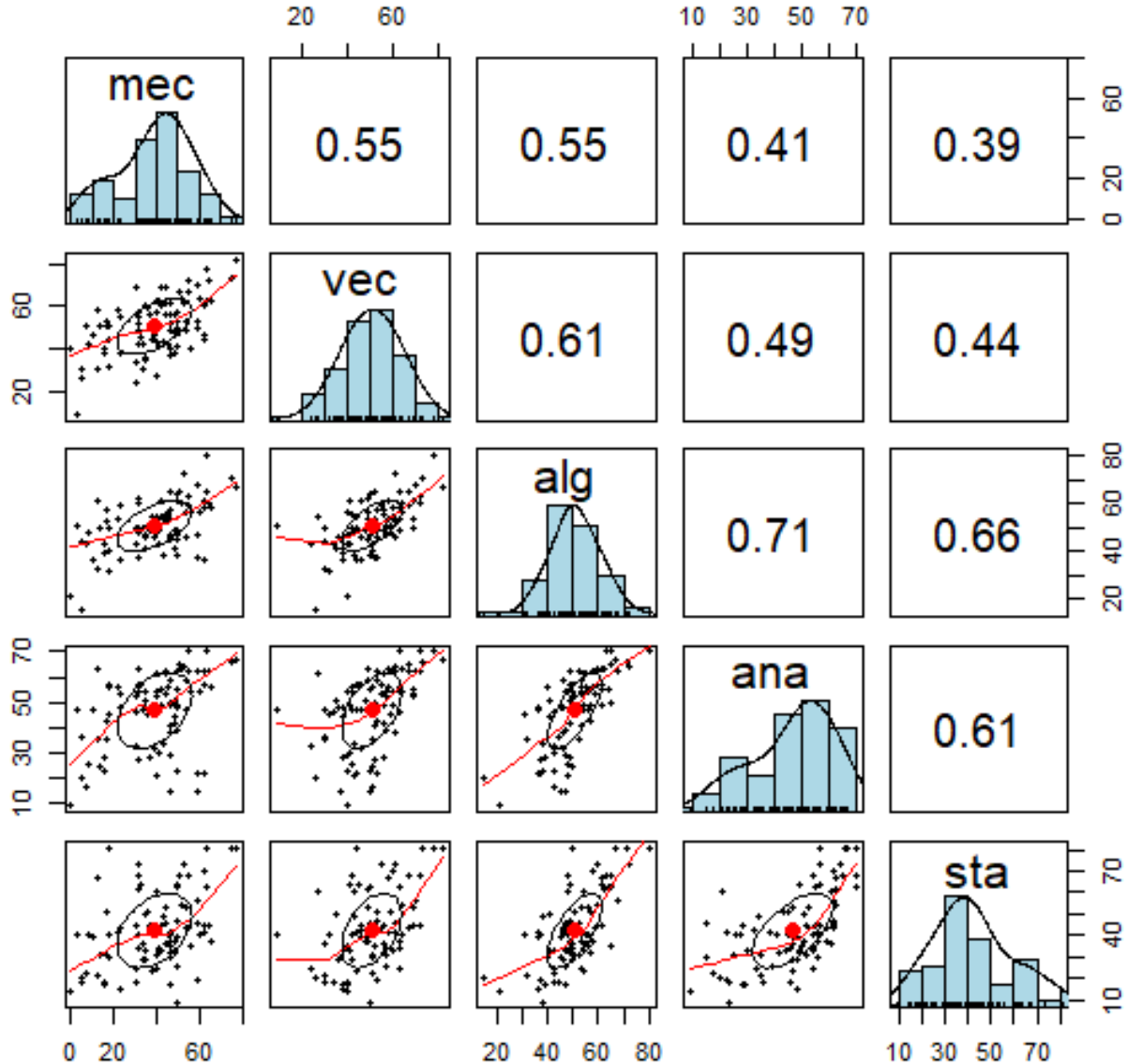
- La mayor parte de los errores estándar son aproximaciones basadas en la distribución normal.
- Estas aproximaciones se parecen a los resultados que se obtienen con el bootstrap paramétrico cuando se hace remuestreo de la distribución normal.
- Cuando se usa bootstrap paramétrico se obtienen resultados más precisos que en las aproximaciones asintóticas cuando éstas existen.

Aplicación a datos multivariantes

- **Ejemplo**: Se tienen unos datos sobre calificaciones en 5 asignaturas de 88 alumnos (ver el libro sobre Análisis Multivariante, de Mardia, Kent and Bibby, (1979)):
  - **mec**: mechanics
  - **vec**: vectors
  - **alg**: algebra
  - **ana**: analysis
  - **sta**: statistics

```
library(bootstrap)
data(scor)
plot(scor)
```

Matriz de variables



Aplicación a datos multivariantes

- El vector de medias y la correspondiente matriz de covarianzas son:

```
colMeans(scor)
```

```
     mec       vec       alg       ana       sta
38.95455  50.59091  50.60227  46.68182  42.30682
```

```
cov(scor)
```

```
          mec       vec       alg       ana       sta
mec  305.7680 127.22257 101.57941 106.27273 117.40491
vec  127.2226 172.84222  85.15726  94.67294  99.01202
alg  101.5794  85.15726 112.88597 112.11338 121.87056
ana  106.2727  94.67294 112.11338 220.38036 155.53553
sta  117.4049  99.01202 121.87056 155.53553 297.75536
```

Aplicación a datos multivariantes

- Se calculan los autovalores y autovectores de la matriz de covarianzas.
- La matriz 5×5 de covarianzas tiene 5 autovalores positivos en orden decreciente: $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \widehat{\lambda}_3 \geq \widehat{\lambda}_4 \geq \widehat{\lambda}_5$ y a cada uno de ellos le corresponde un autovector diferente.

```
round(eigen(cov(scor))$values,3)     # Autovalores
```

```
[1] 686.990 202.111 103.747  84.630  32.153
```

```
round(eigen(cov(scor))$vectors,3)     # Autovectores
```

```
       [,1]   [,2]   [,3]   [,4]   [,5]
[1,] -0.505  0.749 -0.300  0.296 -0.079
[2,] -0.368  0.207  0.416 -0.783 -0.189
[3,] -0.346 -0.076  0.145 -0.003  0.924
[4,] -0.451 -0.301  0.597  0.518 -0.286
[5,] -0.535 -0.548 -0.600 -0.176 -0.151
```

Aplicación a datos multivariantes

- Los autovalores y autovectores de la matriz de covarianzas son importantes para explicar la estructura multivariante de los datos.
- Se observa que las calificaciones en los exámenes están altamente correlacionados entre sí: un estudiante con calificaciones altas en mecánica suele tenerlas altas también en cálculo vectorial.
- Un modelo posible para las medidas correlacionadas sería

$$x_i = Q_i \mathbf{v}$$

para $i = 1, \ldots, 88$

Aplicación a datos multivariantes

- Donde Q_i es un número que representa la capacidad del estudiante i mientras que $\mathbf{v}$ es un vector de valores fijos para todos los estudiantes
- Q_i se puede interpretar como el *coeficiente intelectual* (IQ) del estudiante i-ésimo.
- Si el modelo anterior fuese cierto, entonces solo el primer autovalor $\widehat{\lambda}_1$ sería positivo y el resto de autovalores serían igual a 0.
- También, $\mathbf{v}$ sería igual al primer autovector $\widehat{\mathbf{v}}_1$.

Aplicación a datos multivariantes

- Se define el ratio del mayor autovalor con respecto al total $\widehat{\theta}$,

$$\widehat{\theta} = \frac{\widehat{\lambda}_1}{\sum_{i=1}^{5} \widehat{\lambda}_i}$$

- Así el modelo anterior es equivalente a $\widehat{\theta} = 1$.

- Aunque, en la práctica, no se espera que sea *exactamente* igual a 1

- En el caso de las calificaciones, la estimación de $\widehat{\theta}$ es

$$\widehat{\theta} = \frac{686{,}990}{686{,}990 + 202{,}111 + 103{,}747 + 84{,}630 + 32{,}153} = 0{,}619$$

Aplicación a datos multivariantes

- En muchas circunstancias es interesante tener un valor alto de $\widehat{\theta}$ porque eso indica un alto poder explicativo del modelo.
- El valor de $\widehat{\theta}$ mide el porcentaje de varianza explicada por el primer componente.
- Cuanto más cerca estén los puntos respecto al eje del componente principal, mayor será el valor de $\widehat{\theta}$.
- ¿Qué precisión tiene $\widehat{\theta}$? ¿Cuál es el error estándar de $\widehat{\theta}$?
- Esta sería una aplicación directa del bootstrap en este caso.

Aplicación a datos multivariantes

- La complejidad del cálculo de $\widehat{\theta}$ no resulta relevante, en tanto que se pueda calcular $\widehat{\theta}^*$ para cualquier muestra bootstrap.

- En este caso, una muestra bootstrap es una matriz remuestreada $\mathbf{X}^*$ de tamaño 88×5.

- Las filas $\mathbf{x}_i^*$ de $\mathbf{X}^*$ proceden de una m.a.s de tamaño 88 de las filas de la matriz de datos original.

$$x_1^* = x_{i_1}, \; x_2^* = x_{i_2}, \; \ldots, \; x_{88}^* = x_{i_{88}}$$

- De este modo, algunas filas de $\mathbf{X}$ aparecerán varias veces y otras ninguna en la matriz remuestreada $\mathbf{X}^*$.

Aplicación a datos multivariantes

- Una vez generada $\mathbf{X}^*$ se calcula la matriz de covarianzas $\mathbf{G}^*$ de la manera habitual y luego se calculan los autovalores correspondientes.
- Se calcula la réplica bootstrap de $\widehat{\theta}$

$$\widehat{\theta}^* = \frac{\widehat{\lambda}_1^*}{\sum_{j=1}^{5} \widehat{\lambda}_j^*}$$

- Y se aplica el algoritmo general bootstrap para calcular el error estándar.

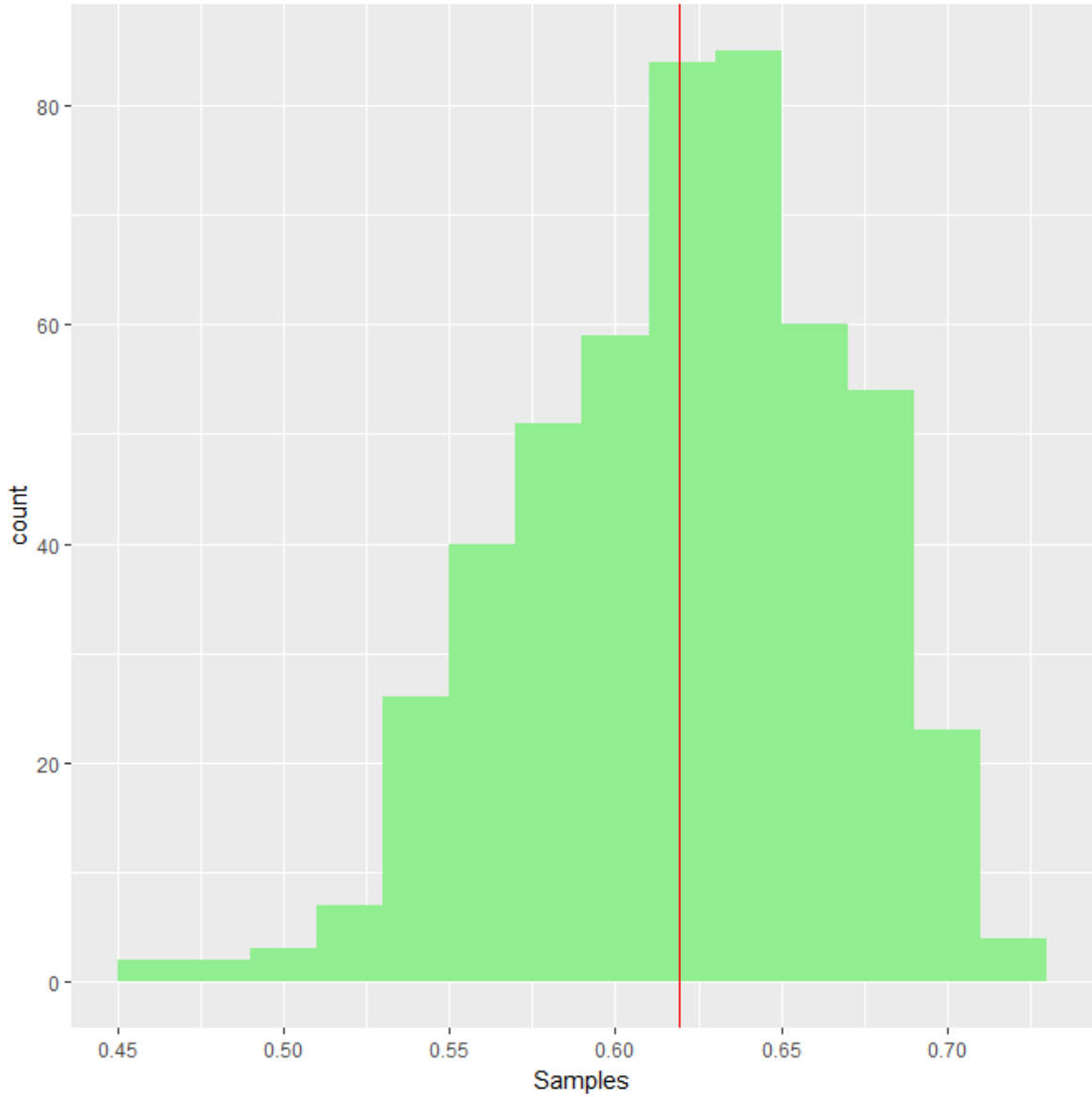


Figura: Distribución bootstrap

Aplicación a datos multivariantes

- El autovector $\widehat{\mathbf{v}}_1$ que corresponde al mayor autovalor se le denomina primer componente principal de $\mathbf{G}$
- Supongamos que se trata de de resumir el rendimiento de los estudiantes mediante un solo número, en lugar de con 5 notas.
- Se puede demostrar que la mejor combinación lineal de las 5 notas es

$$y_i = \sum_{k=1}^{5} \widehat{v}_{1k} x_{ik}$$

es decir, una combinación lineal donde los componentes $\widehat{\mathbf{v}}_1$ equivalen a los pesos de las notas originales.

- Esta combinación lineal es óptima en el sentido de que captura la mayor parte de la variabilidad de las 5 puntuaciones originales de entre todos los posibles $\mathbf{v}$.

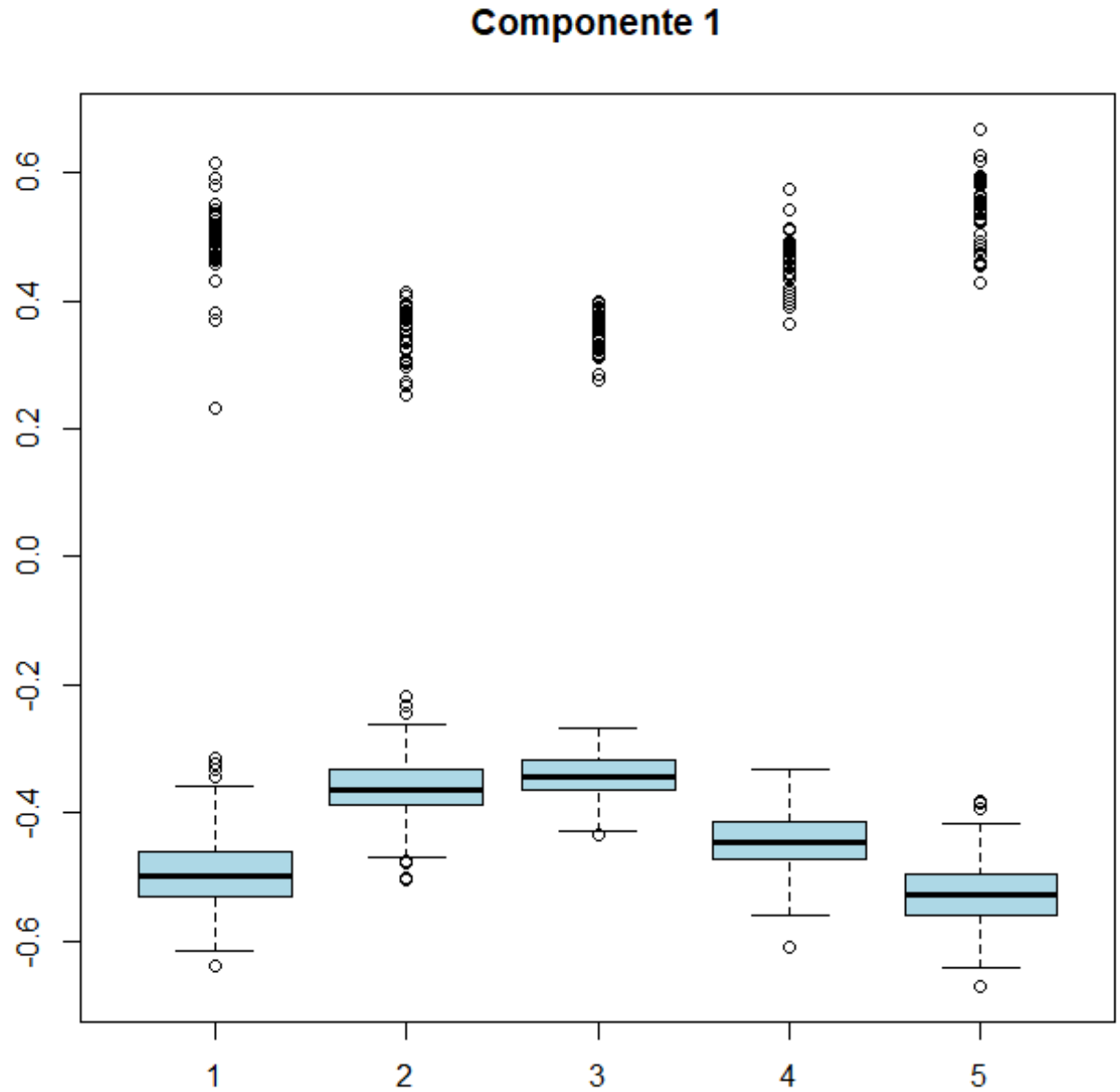
Aplicación a datos multivariantes

- La segunda combinación lineal

$$z_i = \sum_{k=1}^{5} \widehat{v}_{2k} x_{ik}$$

es el segundo componente principal $\widehat{\mathbf{v}}_2$ es decir, el segundo autovector de $\mathbf{G}$.

- El primer componente se puede asociar a la *media de puntuaciones* de un estudiante, mientras que el segundo parece asociarse más bien a la relación que hay entre exámenes con libro *abierto* frente a *cerrado*.

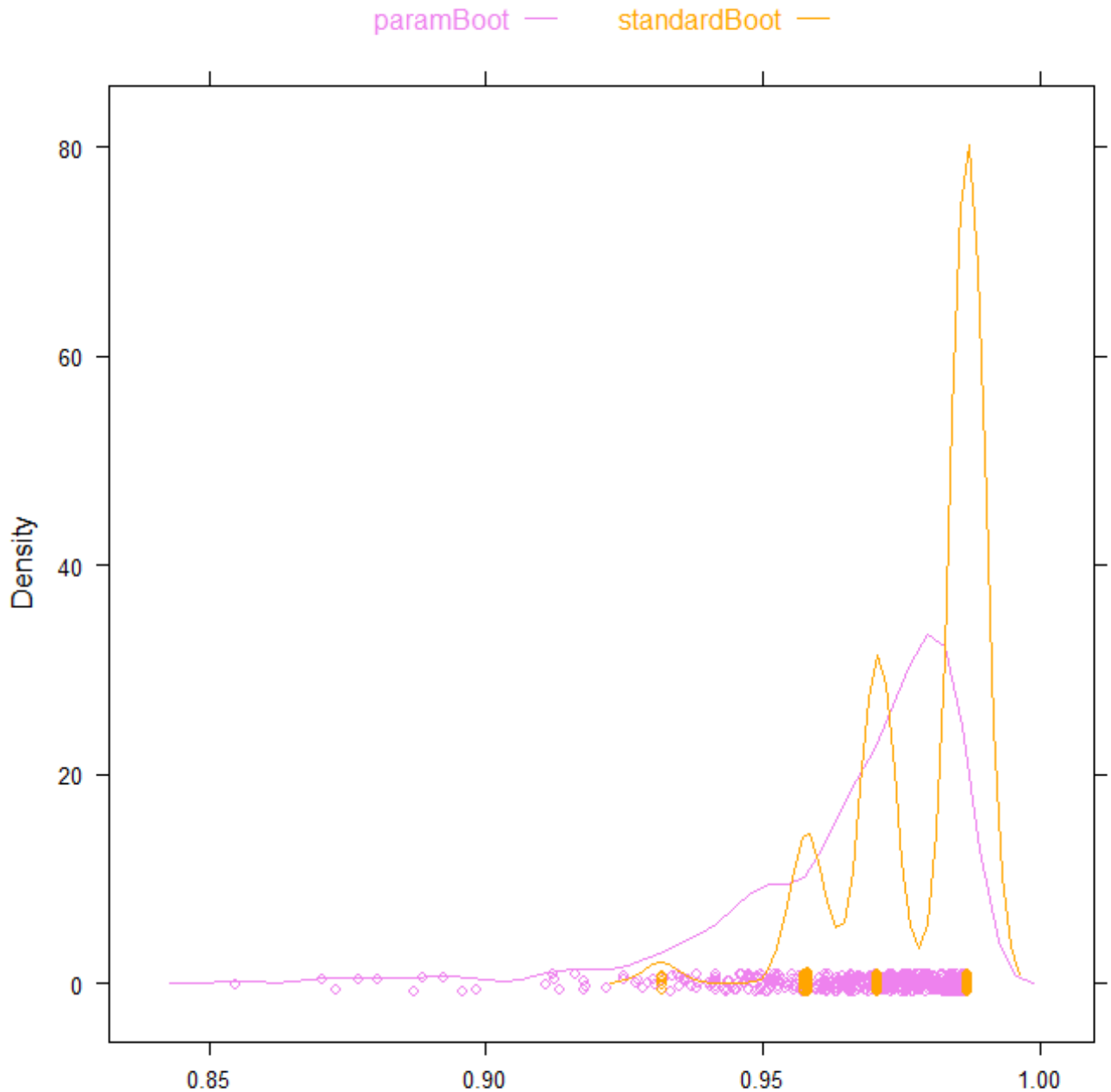

Componente 1
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
1
2
3
4
5

Cuando puede fallar el bootstrap

- En general el bootstrap no funciona cuando se aplica en parámetros que se encuentran en el frontera del espacio paramétrico.
- Consideramos el siguiente ejemplo:
  X se distribuye como una distribución uniforme en $(0, \theta)$.
  El estimador de máxima verosimilitud para θ es el máx(X_i)
- Tenemos una muestra de 50 observaciones.
  Comparamos el estimador bootstrap no paramétrico de θ con respecto al estimador paramétrico del mismo.

```
N = 50
X = runif(N)
(thetaHat = max(X))
```

```
[1] 0.990335
```


paramBoot
standardBoot
Density
80
60
40
20
0
0.85
0.90
0.95
1.00

Estructuras de datos generales

- Hasta ahora se ha considerado una estructura simple de los datos: el modelo unimuestral donde una distribución de probabilidad desconocida F genera los datos X mediante muestreo aleatorio.
- Pero algunos datos x_i pueden ser bastante complejos, como vectores, mapas o imágenes.
- Estructuras complejas de datos aparecen en modelos como series temporales, análisis de varianza, modelos de regresión, datos censurados o muestreo estratificado.
- Pero el método bootstrap se puede adaptar a estructuras de datos generales.

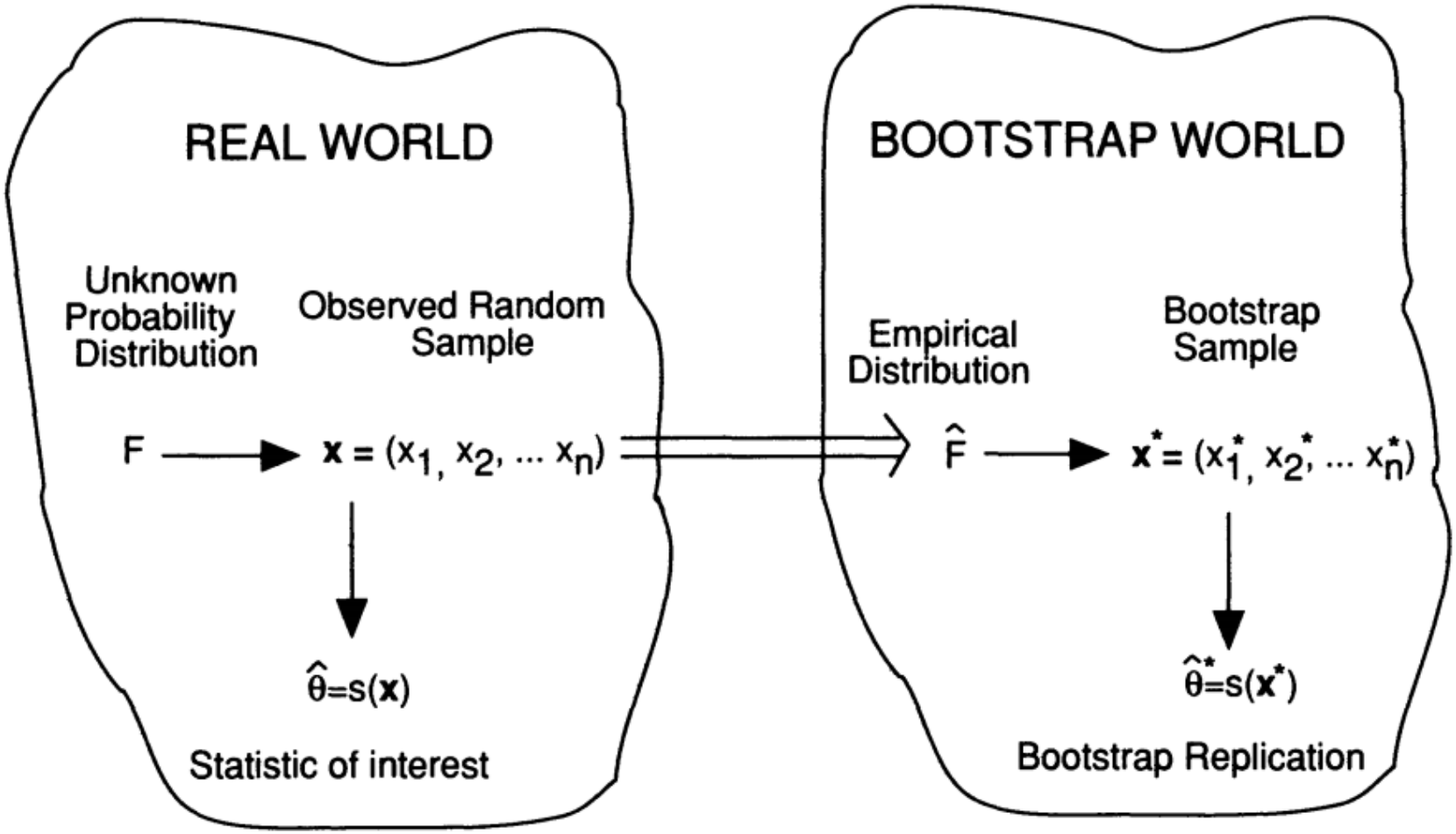
Problemas unimuestrales

- El esquema del método bootstrap para problemas unimuestrales se basa en la existencia de dos *mundos paralelos*.

- Por un lado está el mundo real con una distribución desconocida F de la que se toma una muestra aleatoria y se calcula un estadístico a partir de $\mathbf{x}$ digamos $\widehat{\theta} = s(\mathbf{x})$. Despúes se trata de estudiar su comportamiento: errores, intervalos de confianza, etc.

- Por otro lado está el *mundo bootstrap* de modo que la población se reduce a la muestra original y a partir de la distribución empírica $\widehat{F}$ se obtienen la muestras bootstrap $\mathbf{x}^*$.

- A partir de ella se calcula el estadístico de interés $\widehat{\theta}^* = s(\mathbf{x}^*)$ y se estudia su comportamiento.


REAL WORLD
Unknown Probability Distribution
Observed Random Sample
F
x = (x1, x2, ... xn)
θ̂=s(x)
Statistic of interest
BOOTSTRAP WORLD
Empirical Distribution
Bootstrap Sample
F̂
x* = (x1*, x2*, ... xn*)
θ̂*=s(x*)
Bootstrap Replication

Problemas unimuestrales

- La doble flecha del esquema indica el cálculo de $\widehat{F}$ a partir de F.
- Conceptualmente este es el paso fundamental del bootstrap y el resto de pasos se definen por analogía.
- El procedimiento bootstrap para estructuras más complejas es inmediato una vez que se sabe como realizar el proceso de la *doble flecha*, es decir cómo estimar el mecanismo probabilístico a partir de los datos.
- Se usa la notación $P \rightarrow \mathbf{x}$ para indicar que un modelo de probabilidad desconocido P ha generado el conjunto de datos $\mathbf{x}$.

Problemas de dos muestras

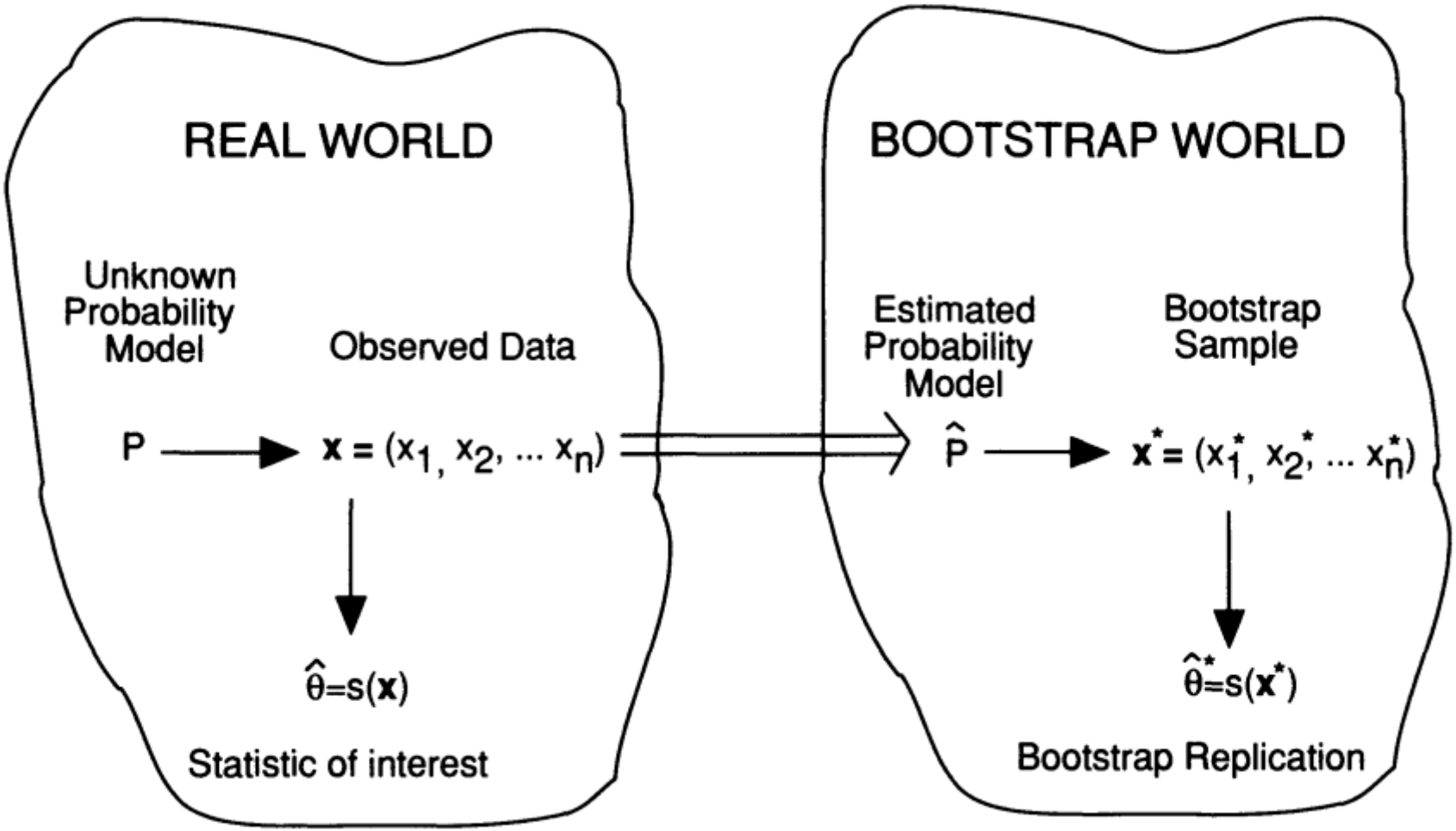
- En el caso del problema de inferencia de dos muestras, el modelo de probabilidad se puede considerar como $P = (F, G)$ donde F es la distribución de probabilidad del primer grupo y G la del segundo grupo.

- Se obtienen dos muestras aleatorias independientes $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ de modo que la aplicación $P \rightarrow \mathbf{z}$ se decribe como
  $F \rightarrow \mathbf{x}$ e independientemente $G \rightarrow \mathbf{y}$

- En este caso se toman las respectivas funciones de distribución empíricas y el estimador natural de P se construye como $\widehat{P} = (\widehat{F}, \widehat{G})$ y se obtiene una muestra bootstrap $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$ como
  $\widehat{F} \rightarrow \mathbf{x}^*$ e independientemente $\widehat{G} \rightarrow \mathbf{y}^*$

Estructuras de datos generales

- El esquema siguiente se aplica a estructuras generales $P \rightarrow \mathbf{x}$
- En el mundo real se tiene un distribución desconocida P que da lugar al conjunto de datos $\mathbf{x}$
- El paso principal es el indicado por $\Rightarrow$ que da lugar un estimador $\widehat{P}$ de la distribución original P.
  De este modo $\widehat{P} \rightarrow \mathbf{x}^*$ es equivalente a $P \rightarrow \mathbf{x}$
- Y así $\mathbf{x}^* \rightarrow \widehat{\theta}^* = s(\mathbf{x}^*)$ es la misma función que $\mathbf{x} \rightarrow \widehat{\theta} = s(\mathbf{x})$.
- Generalmente la generación de muestras bootstrap $\widehat{P} \rightarrow \mathbf{x}^*$ requiere menos tiempo de computación que el cálculo de $\widehat{\theta}^* = s(\mathbf{x}^*)$.


REAL WORLD
Unknown Probability Model
Observed Data
P
x = (x1, x2, ... xn)
θ̂=s(x)
Statistic of interest
BOOTSTRAP WORLD
Estimated Probability Model
Bootstrap Sample
P̂
x* = (x1*, x2*, ... xn*)
θ̂*=s(x*)
Bootstrap Replication