Simulación

4.5 Técnicas de Remuestreo:
Jackknife
Boostrap
Pruebas de Permutación
Validación Cruzada

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Sábado 22 de abril de 2017

Introducción

Introducción

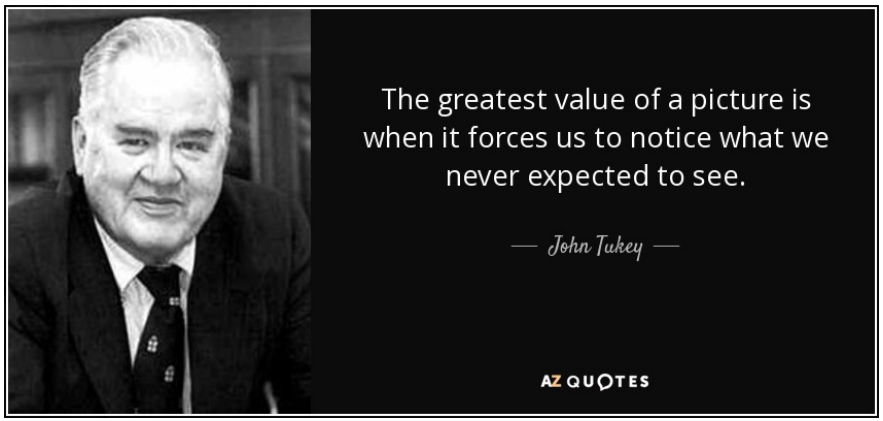


Figura: John Tukey, 1915-2000.

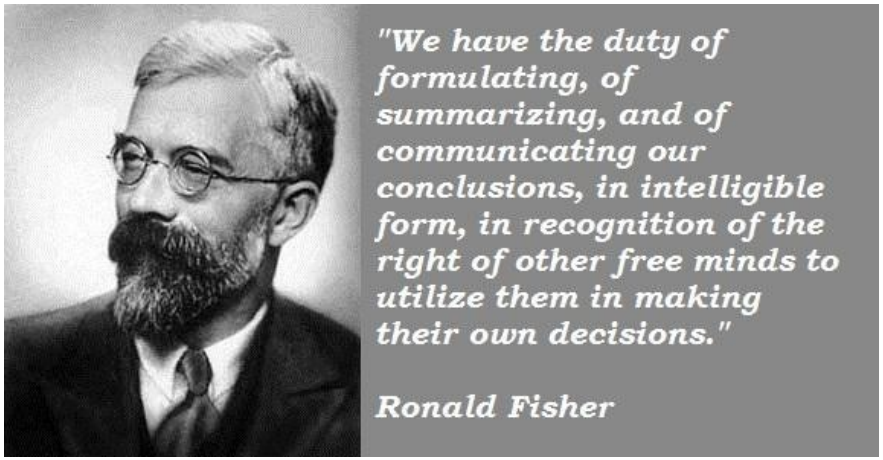


Figura: Ronald Avery Fisher, 1890-1962.

- Remuestreo es un conjunto de técnicas estadisticas, computacionalmente intensivas y no paramétricas, en donde las pruebas o métricas se basan en muestreo aleatorio con reemplazo.
- Las técnicas de remuestreo permiten asignar medidas de ajuste (en términos de sesgo, varianza, intervalos de confianza, errores de predicción o de algunas otras medidas) a los estimados basados en muestras.
- Usualmente no paramétricas, y varias son tan antiguas como la estadística misma. Por ejemplo, las técnicas de permutación son de Fisher (1935) y Pitmann (1937); la validación cruzadas fue propuesta por Kurtz en 1948, y el jackknife fue propuesto por Maurice Quenouille en 1949 y John Tukey en 1958 fue quien le dió nombre.

Introducción

- Bradley Efron introdujo el Bootstrap en 1979, y sus estudiantes Rob Tibshirani y Trevor Hastie han aportado mucho a la ciencia estadística. Ofrecen un curso en Statistical Learning en la plataforma MOOC de la Universidad de Stanford.
- El término 'bootsrapping' se refiere al concepto de "pulling oneself up by one's bootstraps", frase que aparentemente se usó por primera vez en *The Singular Travels, Campaigns and Adventures of Baron Munchausen* de Rudolph Erich Raspe en 1786.

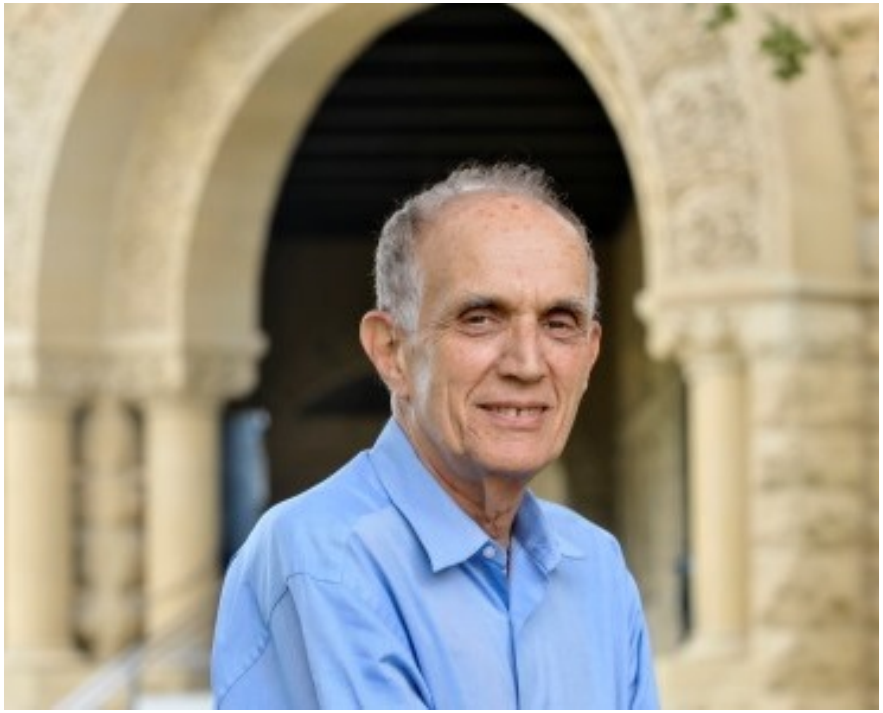


Figura: Bradley Efron en 2014.

Introducción I

- El objetivo del Remuestreo es estiar un parámetro (tal como media, mediana, desviación estándar, coeficientes de regresión, etc.) basado en los datos. También interesan las propiedades de la distribución de estimador, sin hacer supuestos restrictivos sobre la forma de la distribución de los datos originales.
- Para una muestra aleatoria, la distribución de remuestreo es la distribución empírica $\hat{F}_n$, que asigna probabilidad $1/n$ a cada una de las observaciones de la muestra.
- Se consideran como remuestreo las siguientes técnicas:
  - Jackknife
  - Bootstrap
  - Pruebas de permutación
  - Validación cruzada

  y los usos asociados a estas técnicas son los siguientes:
  - Estimación de sesgo de un estimador (jackknife, boostrap)
  - Reducción de sesgo de un estimador (jackknife, boostrap)
  - Pruebas estadísticas exactas (pruebas de permutación)
  - Validación de modelos (validación cruzada)

Motivación I

Consideremos una muestra de 4 parejas. La variable de interés es la diferencia del ingreso de cada pareja (en miles de pesos, al mes)

i	P1	P2	d_i
1	24	18	6
2	14	17	-3
3	40	35	5
4	44	41	3

- $\theta =$ promedio de diferencial de ingreso poblacional. Podemos estimar θ con $\hat{\theta} = \bar{d} = 2.75$. La desviación estándar de $\hat{\theta}$ es $sd(d) = \sigma/\sqrt{n}$, donde σ^2 es la varianza poblacional.
- Si σ^2 es conocida, y si d_i tuviera distribución normal, un intervalo de confianza del 95 % para μ sería $\bar{d} \pm z_{0.975}\sigma/\sqrt{n}$.
- si d_i no es normal, el resultado aún se cumple asintóticamente, pero en este problema, $n = 4$.
- Entonces, ¿cómo podemos concluir sobre $\bar{d}$ sin conocer la distribución de las observaciones?

Motivación II

- Como en la vida real usualmente no conocemos σ, tenemos que estimar también este parámetro a través del estimador de la desviación estándar, que es el *error estándar*:

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$$

  y el intervalo de confianza para μ cambia a $\bar{d} \pm t_{n-1,0.975}\, s/\sqrt{n}$.

- Ahora, supongamos que queremos estimar un parámetro diferente, por ejemplo, la mediana, $q_{.5} = F^{-1}(1/2)$. Un estimador de $q_{.5}$ es por ejemplo

$$\hat{q_{.5}} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ es impar} \\ \frac{X_{(n/2)} + X_{(n/2+1)}}{2} & \text{si } n \text{ es par} \end{cases}$$

  Aquí no es evidente cómo obtener un estimador de la variabilidad de $\hat{q_{.5}}$. Para contar con estimadores de variabilidad de parámetros como éste es para lo que se usan las técnicas de remuestreo.

El jackknife y el bootstrap nos ofrecen opciones para hacer la extensión. Además de la desviación estándar, el jackknife y el bootstrap permiten estimar otras medidas de error estadístico, como sesgo, error de predicción e intervalos de confianza.

Calculos bootstrap I

Una idea para obtener la distribución de la media es la siguiente:

- Extrae muestras con reemplazo de la muestra original (digamos $B = 100$). Hay un total de n^n posibles muestras (en este caso $4^4 = 256$ posibles muestras diferentes).

```
x <- c(6,-3,5,3) #muestra original
xb <- matrix(nrow=500,ncol=4) #contenedor para las muestras
for(i in 1:100) xb[i,] <- sample(x,replace=T) #muestra boostrap
head(xb) #ejemplo de muestras bootstrap
```

```
     [,1] [,2] [,3] [,4]
[1,]   -3    6    5    3
[2,]    5    6    6   -3
[3,]    5    5    6    6
[4,]    6    3    6    5
[5,]   -3    6   -3   -3
[6,]    5    3   -3    6
```

- Para cada muestra, calculamos la diferencia promedio $\bar{d}^*$ y obtenemos su histograma
- A partir del histograma, podemos obtener características de $\bar{d}$, como su variabilidad, intervalos de confianza, etc.

Calculos bootstrap II

```
db <- apply(xb,1,mean) #media por renglon
db[1:100] #muestras del promedio de la diferencia
```

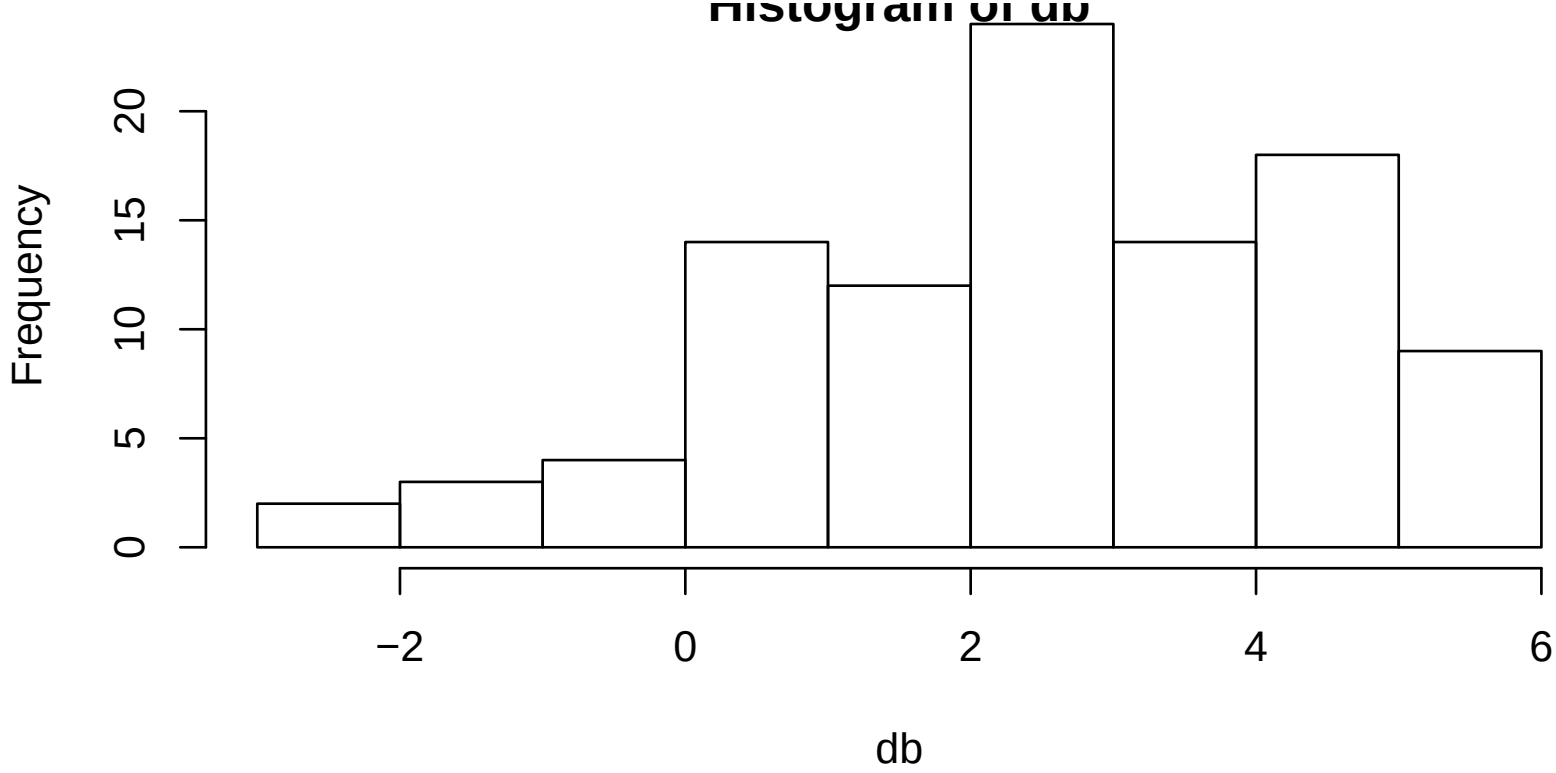
```
  [1]  2.75  3.50  5.50  5.00 -0.75  2.75  3.00  3.50  5.00  1.25  2.00
 [12]  5.25 -1.00  3.75  1.50  3.00  2.75  2.50  4.50  3.75  0.75  5.50
 [23]  4.50  4.25  1.25  4.25  3.50  2.00  2.25  3.25  2.75  1.25  0.00
 [34]  2.75  3.00  5.25  3.50 -1.50  0.75  5.50  3.00  2.75  2.75  2.75
 [45]  2.75 -1.00  4.25  3.75  0.50  1.25  0.00  0.50  3.50  5.50  0.00
 [56]  2.00  2.75  0.50  3.50  3.50  0.50  4.75  2.00  1.25  2.25  5.00
 [67]  0.50  1.25  2.25  0.50  0.75  0.75  5.00  4.50  2.50  4.50  5.00
 [78] -3.00  5.00  4.75  2.50  0.75  4.00  3.00  2.00  0.50  5.25  0.75
 [89]  4.25  2.50  3.50  4.25  0.75  4.50  3.00  5.50  3.75  2.75 -3.00
[100]  6.00
```

```
summary(db)  #características de las muestras bootstrap
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -3.000   1.250   2.750   2.652   4.250   6.000     400
```

```
hist(db)
```

Calculos bootstrap III



Jackknife

Estimador Jackknife del Sesgo I

Si θ es un parámetro y $\hat{\theta} = T(X_1, \ldots, X_n)$ es un estimador de θ, se define el sesgo del estimador como $sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$. Si $\hat{\theta}_{(i)}$ es la estadística T calculada sin la observación i (cuidado: no confundir la notación con las estadísticas de orden), entonces

Estimador Jackknife del sesgo

El estimador jackknife del sesgo es

$$sesgo_{jack} = (n-1)\left(\frac{\sum_{i=1}^{n}\hat{\theta}_{(i)}}{n} - \hat{\theta}\right)$$

y el estimador corregido por sesgo es:

$$\hat{\theta}_{jack} = \hat{\theta} - sesgo_{jack} = n\hat{\theta} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$$

Ejamplo

En nuestro ejemplo inicial, $\hat{\theta} = \bar{d} = 2.75$, y el sesgo está dado por:

Estimador Jackknife del Sesgo II

```
d <- c(6, -3, 5, 3)
dadj <- NULL #guarda los valores ajustados
for(i in 1:4) dadj[i] <- mean(d[-i])
sesgo <- 3*(mean(dadj)-mean(d))
sesgo

[1] 0

dadj

[1] 1.666667 4.666667 2.000000 2.666667
```

Noten que si definimos $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$ como *pseudovalores*, entonces $\hat{\theta}_{jack}$ es el promedio de estos pseudovalores:

$$\hat{\theta}_{jack} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i$$

En el ejemplo considerado, los pseudovalores son iguales a los valores originales.

Estimador Jackknife del Sesgo III

Estimador Jackknife de varianza

El estimador jackknife de varianza del estimador, $var(\hat{\theta})$ es

$$\hat{var}(\hat{\theta}) = v_{jack} = \frac{\tilde{s}^2}{n}$$

donde

$$\tilde{s}^2 = \frac{\sum_{i=1}^{n} \left(\tilde{\theta}_i - \hat{\theta}_{jack}\right)^2}{n-1}$$

Ejemplo

En nuestro ejemplo, el error estándar coincide con el error estándar del parámetro original.

Ejemplo I

Si $\theta = E(X)$, entonces basado en una muestra $X_1, \ldots, X_n$, tenemos $\hat{\theta} = \bar{X}$, y sea

$$\bar{X}_{(i)} = \frac{n\bar{X} - X_i}{n-1} = \frac{1}{n-1}\sum_{j \neq i} X_j$$

la media de $n-1$ observaciones sin considerar la i-ésima observación. Definir

$$\bar{X}_{(\cdot)} = \frac{\sum_{i=1}^{n} \bar{X}_{(i)}}{n}$$

Noten que $\bar{X}_{(\cdot)} = \bar{X}$. Pero al hacer esta descomposición, podemos calcular un estimador de variabilidad. El estimador jackknife para la varianza de estimador es

$$v_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\bar{X}_{(i)} - \bar{X}_{(\cdot)})^2$$

Este estimador se puede aplicar no sólo para la media.

Ejemplo Jackknife I

Consideren una muestra aleatoria $X_1, \ldots, X_n \sim$ **Bernoulli** (θ). Queremos estimar θ^2.

- El estimador máximo verosimil de θ^2 es $\bar{x}^2$. Sin embargo, este estimador tiene sesgo (el cuál es fácil de calcular considerando que $Y = \sum_{i=1}^n X_i \sim$ **Bin** (n, θ) y $E(Y^2) = n\theta(1-\theta) + n^2\theta^2$:

$$E(\hat{\theta}^2) = \theta + \frac{\theta(1-\theta)}{n}$$

- De acuerdo a las definiciones previas, el estimador jackknife está dado por:

$$\hat{\theta}^2_{jack} = \frac{n}{n-1}\bar{x}^2 - \frac{\sum_{i=1}^n X_i^2}{n(n-1)}$$

y el sesgo de este estimador es 0:

$$\begin{aligned} E(\hat{\theta}^2_{jack}) &= \frac{n}{n-1}E(\bar{x}^2) - \frac{E(\sum_{i=1}^n X_i^2)}{n(n-1)} \\ &= \frac{n}{n-1}\left(\theta^2 + \frac{\theta(1-\theta)}{n}\right) - \frac{n\theta}{n(n-1)} \\ &= \frac{n\theta^2 + \theta - \theta^2 - \theta}{n-1} \\ &= \theta^2 \end{aligned}$$

Ejemplo Jackknife II

- numéricamente:

```
set.seed(1)
   n <- 100  #tamaño de muestra
   x <- rbinom(n,size=1,p=0.3)
   x

  [1] 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 1 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1
 [36] 0 1 0 1 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1
 [71] 0 1 0 0 0 1 1 0 1 1 0 1 0 0 1 0 1 0 0 0 0 0 0 1 1 1 0 0 1 0

   thetahat2 <- mean(x)^2  #estimador máximo verosímil
   thetahat2

[1] 0.1024

   thetahat2_jn <- (n/(n-1))*thetahat2-sum(x^2)/(n*(n-1))
   thetahat2_jn

[1] 0.100202

   sesgo <- thetahat2 - thetahat2_jn
   sesgo

[1] 0.00219798

   #Ahora calculamos los pseudovalores:
   pseudovals <- NULL
   for (i in 1:n) pseudovals[i] <- n*thetahat2 - (n-1)*mean(x[-i])^2
   mean(pseudovals)

[1] 0.100202

   stilde2 <- sum((pseudovals-thetahat2_jn)^2)/(n-1)
   stilde2

[1] 0.0890091
```

Bootstrap

Boostrap I

El bootstrap es un método para estimar la varianza y la distribución de la estadística $\hat{\theta} = T(X_1, \ldots, X_n)$, así como intervalos de confianza. En el caso del bootstrap, la distribución empírica juega un rol muy importante.

- Los parámetros se pueden escribir como funcionales de F: $\theta = T(F)$. Por ejemplo
  - $\mu = E_F(X) = \int x dF$
  - $\sigma^2 = E_F((X - \mu)^2) = \int (x - \mu)^2 dF$
  - $\eta = P_F(x \in A) = \int_A dF$

Principio del plug-in:

Estima θ del siguiente modo: si $\theta = T(F)$, entonces $\hat{\theta} = T(\hat{F})$.

- Por ejemplo $\bar{X}$, s^2 y $\rho_{\hat{X}Y}$ son estimados plug-in.
- Cuando hay información adicional acerca de F que no viene de la muestra, el principio plug-in es menos bueno en general (por ejemplo, si se asume que F viene de una familia paramétrica). Pero aun en estos casos el principio plug-in puede ser adoptado.

Algoritmo bootstrap para estimar varianza de un estimador I

Algoritmo para estimar varianza de un estimador

Si $\theta = T(F)$ y $\hat{\theta} = g(X_1, \ldots, X_n)$, entonces

1. Extraer una muestra $X_1^*, \ldots, X_n^* \sim \hat{F}$ (equivalentemente, extraer $X_1^*, \ldots, X_n^*$ con reemplazo de $X_1, \ldots X_n$).
2. Calcular $\hat{\theta}^* = g(X_1^*, \ldots, X_n^*)$.
3. Repetir los pasos 1 y 2, B veces para obtener $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$
4. Definir

$$v_{boot} = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \frac{1}{B} \sum_{r=1}^{B} \hat{\theta}_r^* \right)^2$$

Por la ley de los grandes números, $v_{boot} \overset{cs}{\to} Var_{\hat{F}}(\hat{\theta})$ conforme $B \to \infty$. El siguiente diagrama describe la idea bootstrap:

- Mundo real: $F \implies X_1, \ldots, X_n \implies \theta = T(F), \hat{\theta} = g(X_1, \ldots, X_n)$
- Mundo bootstrap: $\hat{F} \implies X_1^*, \ldots, X_n^* \implies \hat{\theta}^* = g(X_1^*, \ldots, X_n^*) = T(\hat{F})$

$$Var_F(\hat{\theta}) \approx \text{Var}_{\hat{F}}(\hat{\theta}) \approx v_{boot}$$

Ejemplo de las diferencias I

Siguiendo con el ejemplo de las diferencias, podemos encontrar la distribución de las muestras bootstrap:

```
d <- c(3,5,-3,6)
Boot <- NULL
B <- 200
Boot <- matrix(0, nrow=B, ncol=4)
for(i in 1:B) Boot[i,] <- sample(d,replace = T)
head(Boot)
```

```
     [,1] [,2] [,3] [,4]
[1,]   -3    5    5    6
[2,]   -3    3    3    5
[3,]    6   -3    6   -3
[4,]    5    5    3    3
[5,]   -3    3    5   -3
[6,]    6    5    5    3
```

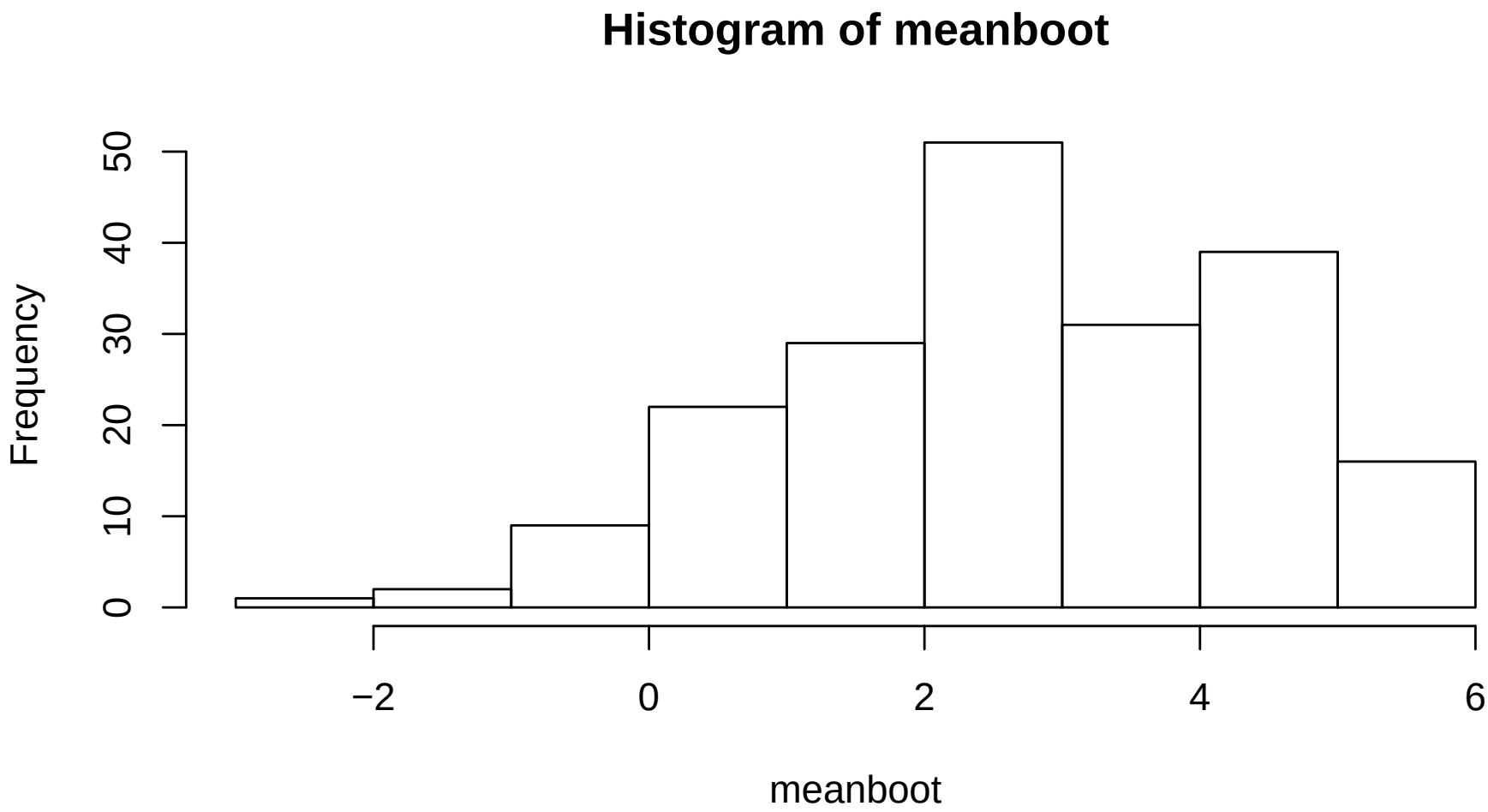
Ejemplo de las diferencias II

```
meanboot <- apply(Boot, 1, mean)
meanboot
```

```
  [1]  3.25  2.00  1.50  4.00  0.50  4.75  2.75  0.00  0.75 -0.75  0.50
 [12]  0.50  0.75  2.50  2.00  5.50  4.75  1.25  5.50 -0.75  2.75  3.00
 [23]  1.25  2.75  5.00  2.50  0.00  2.75  5.75  3.50  2.50  2.75  4.75
 [34]  1.00  1.00  2.75  4.00  5.75  2.50  4.75  3.25  2.25  1.50  4.00
 [45]  4.75  2.75  3.50  3.00  1.25  3.75  2.50  5.00  3.50  3.25  2.75
 [56]  3.25  6.00  5.50  4.75  2.75  2.75  2.50  2.50  5.50  5.00  4.75
 [67]  0.75  1.00  4.25  4.75  1.25  4.50  3.75  2.75  2.50  3.00  3.50
 [78]  5.00  1.25  3.25  0.50  4.75  5.25  2.75  2.75  0.50  5.25  4.25
 [89]  2.00  0.75  4.00  2.75  3.25  2.00  3.25  5.25  0.00  4.50  4.50
[100]  2.75 -3.00  5.50  3.50  0.00 -1.00  2.75  5.75  3.00  4.00  2.50
[111]  3.25  4.25 -0.75  5.00  4.50  0.50  0.75  4.75  2.75  1.25  2.75
[122]  0.50  0.75  1.50  5.00  3.00  5.50  5.00 -1.00  2.00  2.25  1.00
[133]  5.50  5.25  0.00  5.00  4.50  1.50  2.50  3.00  2.75  2.75  1.25
[144]  1.25  4.50  3.00  1.00  2.25  1.25  5.00  3.75  0.50  1.25  4.75
[155]  4.25  0.75  3.25  0.50  2.75  4.25  1.00  4.75  2.75  2.00  4.75
[166]  4.50  0.00  3.50  2.75  4.50  1.25  2.75  3.50  1.25  2.75  3.00
[177]  3.25  4.50  3.25  3.00  1.25  4.25  1.25  3.50  3.00  4.25  2.50
[188]  1.25  3.75  2.00  2.00  2.50  2.75  4.25  3.25  5.75  3.50  2.00
[199]  3.50  1.50
```

```
hist(meanboot)
```

Ejemplo de las diferencias III




```
mean(meanboot)
```

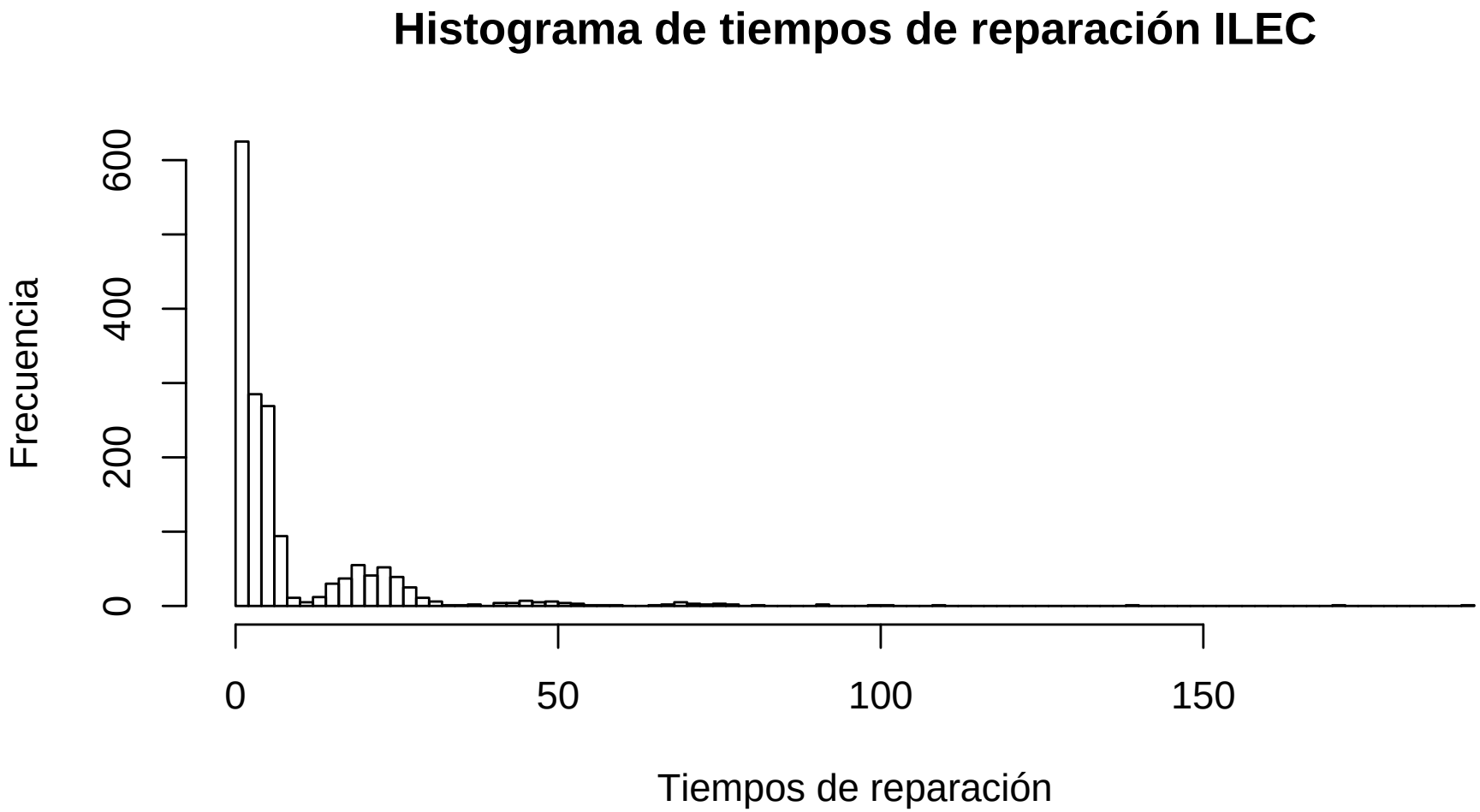
```
[1] 2.84875
```

Ejemplo 1. Media poblacional I

A continuación veremos la aplicación de Bootstrap, Jackknife, valiación cruzada y pruebas de permutación con varios ejemplos.
Los siguientes datos corresponden a los tiempos de reparación de Verizon, compañía telefónica que actua en dos modos: como compañia primaria local telefónica (Incumbent Local Exchange Carrier ILEC) o como competidor en otras regiones (Competing Local Exchange Carrier, CLEC). Como ILEC Verizon debe proveer servicio de reparación para los clientes de las CLEC en su región. está sujeta a multas y la autoridad requiere el uso de pruebas de significancia para comparar los tiempos de reparación de los dos grupos de clientes.

```
library(resample) #paquete con los datos.
data(Verizon)
ilec <- Verizon$Time[Verizon$Group=="ILEC"] #Toma los tiempos de reparación para la competencia
hist(ilec,breaks = 100, main = "Histograma de tiempos de reparación ILEC",
                ylab = "Frecuencia",
                xlab = "Tiempos de reparación")
```

Ejemplo 1. Media poblacional II



Los datos no son normales. Sin embargo, sabemos que la media tiende a distribuirse como una normal.

Ejemplo 1. Media poblacional III

```
n <- length(ilec)
summary(ilec)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.730   3.590   8.412   7.080 191.600
```

```
#Intervalo para la media:
mean(ilec) + c(-1,1)*qt(.975,df = n-1 ,lower.tail = T)*sd(ilec)/sqrt(n)
```

```
[1] 7.705276 9.117945
```

Con una muestra bootstrap, se puede reproducir la forma y dispersión de la distribución. En el caso de la media es obvio, pero esto no es evidente para otras estadísticas. Ese es justamente la fortaleza del bootstrap, que nos permite obtener una estimación de la distribución de la estadística, no sólo para la media, sino para estadísticas mucho más complejas.
La media de la distribución boostrap está centrada cerca de la media de la muestra obtenida, no la media poblacional. Así que la media de muestra boostrap tiene sesgo como estimador de la media poblacional. Sin embargo, el tamaño del sesgo de la estimación es parecido al sesgo que puede tener la media muestral respecto a la media poblacional.

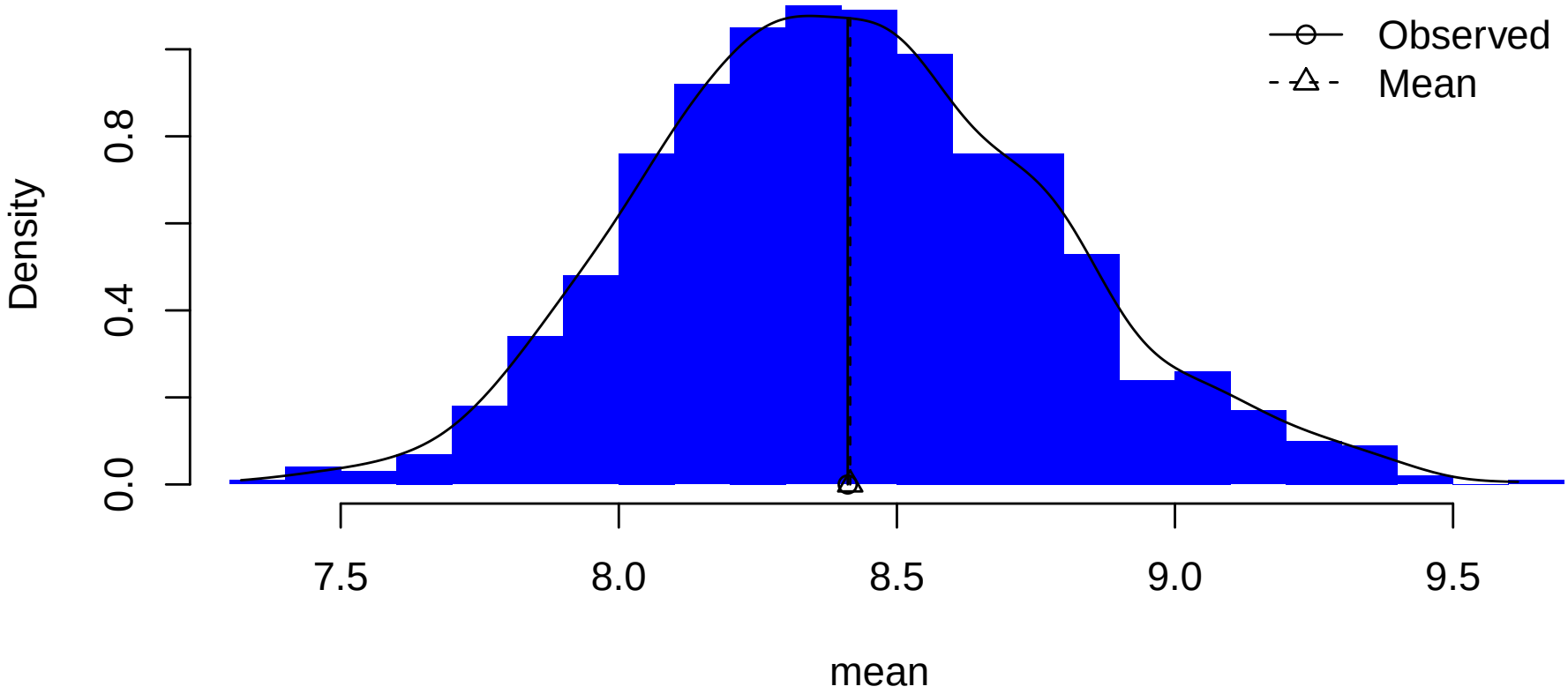
Ejemplo 1. Media poblacional IV

```
z <- bootstrap(ilec, mean, R = 1000)
summary(z$replicates)
```

```
      mean
 Min.   :7.321
 1st Qu.:8.168
 Median :8.401
 Mean   :8.416
 3rd Qu.:8.652
 Max.   :9.617
```

```
hist(z)
```

Ejemplo 1. Media poblacional V



Intervalos de confianza bootstrap I

Hay diferentes maneras de calcular intervalos de confianza bootstrap. Varian en su facilidad de cómputo y en su exactitud. Estos se basan en el error estándar bootstrap $\hat{se}_{boot} = \sqrt{v_{boot}}$.

- Usualmente este intervalo no es adecuado a menos que $\hat{\theta}$ tenga una distribución parecida a la normal.

Intervalo Normal

$$\hat{\theta} \pm z_{\alpha/2} \cdot \hat{se}_{boot}$$

- Si la distribución bootstrap de una estadística tiene forma normal y sesgo pequeño, entonces se puede obtener un intervalo de confianza para el parámetro usando el error estándar bootstrap y la distribución t.

Intervalo t

$$\hat{\theta} \pm t_{n-1,\alpha/2} \cdot \hat{se}_{boot}$$

Intervalos de confianza bootstrap II

- Los intervalos percentíl se basan en la distribución de las réplicas bootstrap.

Intervalo percentil

A partir de la distribución bootstrap de $\hat{\theta}^*$ se puede calcular un intervalo de confianza no paramétrico. El $(1-\alpha)\times 100\,\%$ intervalo basado en percentiles es

$$(\hat{\theta}^*_{(B\cdot\alpha/2)}, \hat{\theta}^*_{(B\cdot(1-\alpha/2))})$$

basado en los cuantiles $B\cdot\alpha/2$ y $B\cdot(1-\alpha/2)$ de la muestra bootstrap.

- Los intervalos percentil pueden ser mejorados con ciertos métodos de ajuste de percentiles. El más popular es el método acelerado con corrección de sesgo, BC_a (*bias-corrected, accelerated* en inglés)

Intervalos de confianza bootstrap III

Intervalos BC_a

Para calcular un intervalo de confianza de $100(1-\alpha)\,\%$ para θ, realizamos los siguientes ajustes:

1. Calcular $z_B = \Phi^{-1}\left(\frac{\#\{\hat{\theta}_i^* < \hat{\theta}\}}{B}\right)$
2. Calcular
$$A = \frac{\sum_{b=1}^{B}(\hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)})^3}{6\left[\sum_{b=1}^{B}(\hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)})^2\right]^{3/2}}$$
Este cálculo corresponde al cociente del estimador de sesgo entre la varianza.
3. Define $A_1 = \Phi\left[z + \frac{z - z_{\alpha/2}}{1 - A(z - z_{\alpha/2})}\right]$ y $A_2 = \Phi\left[z + \frac{z + z_{\alpha/2}}{1 - A(z + z_{\alpha/2})}\right]$. El intervalo está dado por $(\hat{\theta}^*_{(l^*)}, \hat{\theta}^*_{(u^*)})$, donde $l^* = BA_1$ y $u^* = BA_2$.

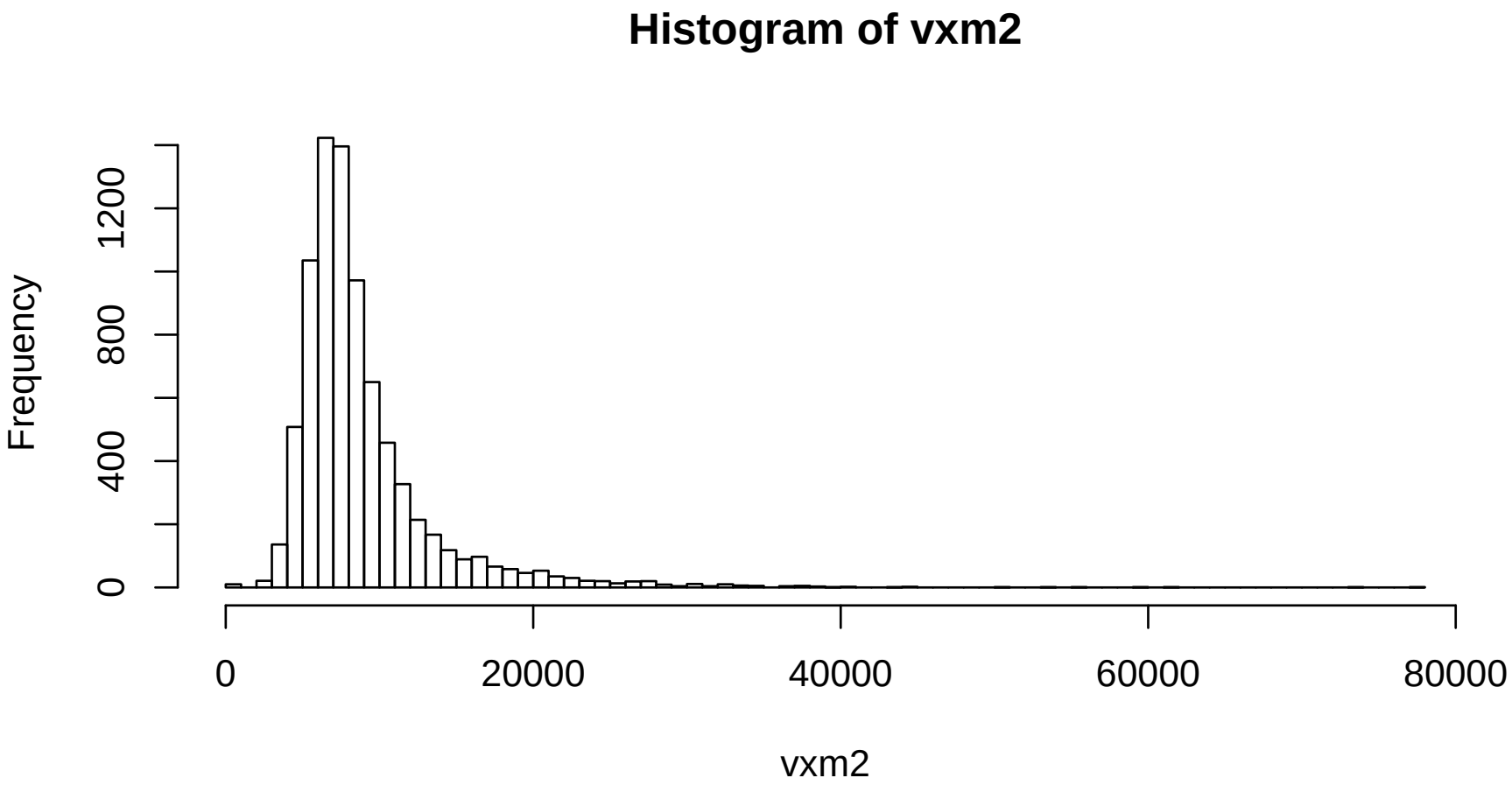
Otros intervalos son posibles. Por ejemplo, intervalos pivotales, intervalos estudentizados, etc. No serán considerados en lo que sigue. Sin embargo, en términos de exactitud, se puede ver que los más exactos son los ajustados, los estudentizados y los que son menos exactos son el normal, el pivotal y el percentil.

Ejemplo 2: Intervalos de confianza boostrap-t I

- Consideren los costos por metro cuadrado de vivienda en México. Los datos corresponden al Promedio de valor del mercado por metro cuadrado de construcción, que se puede obtener de la Sociedad Hipotecaria Federal (avaluos) a partir de avalúos.Los datos corresponden a 8,076 registros de todos los municipios de México registrados en 2015, y los datos consideran todo tipo de vivienda, desde aquella de interés social como la residencial.
- La variable de interés es `vxm2`, el valor de la vivienda por metro cuadrado de construcción, en pesos. Los datos están en el archivo `SHF-PromedioValorMercadoxMetroCuadradoConstruccion.csv` y se encuentran en el blog.

```
vxm2  <- read.csv("https://github.com/jvega68/Simulacion/blob/master/SHF-PromedioValorMercadoxMetroCuadradoConstruccion.csv?raw
hist(vxm2,breaks=100)
```

Ejemplo 2: Intervalos de confianza boostrap-t II



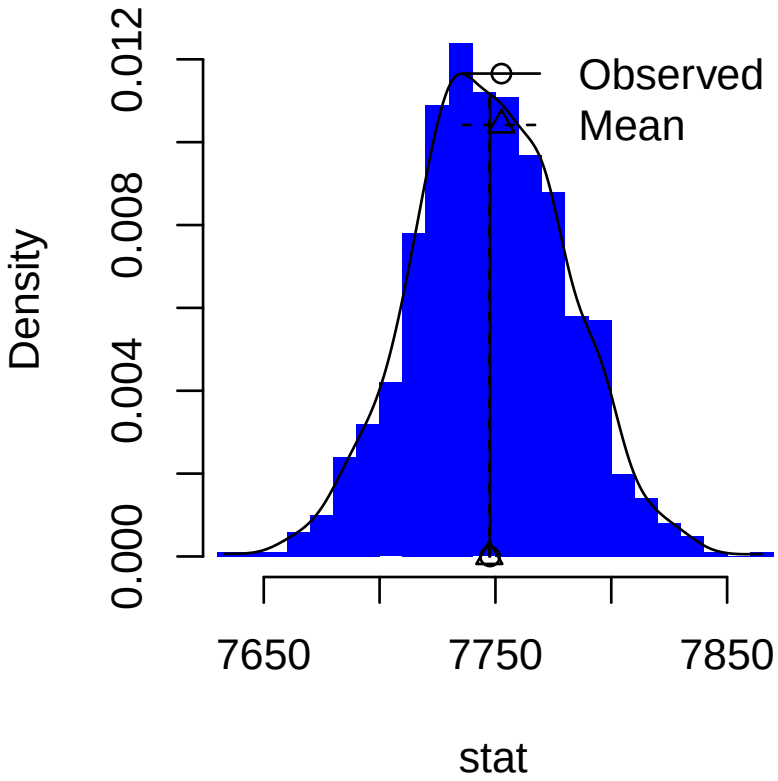

```
summary(vxm2)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    6242    7616    8861    9848   77161
```

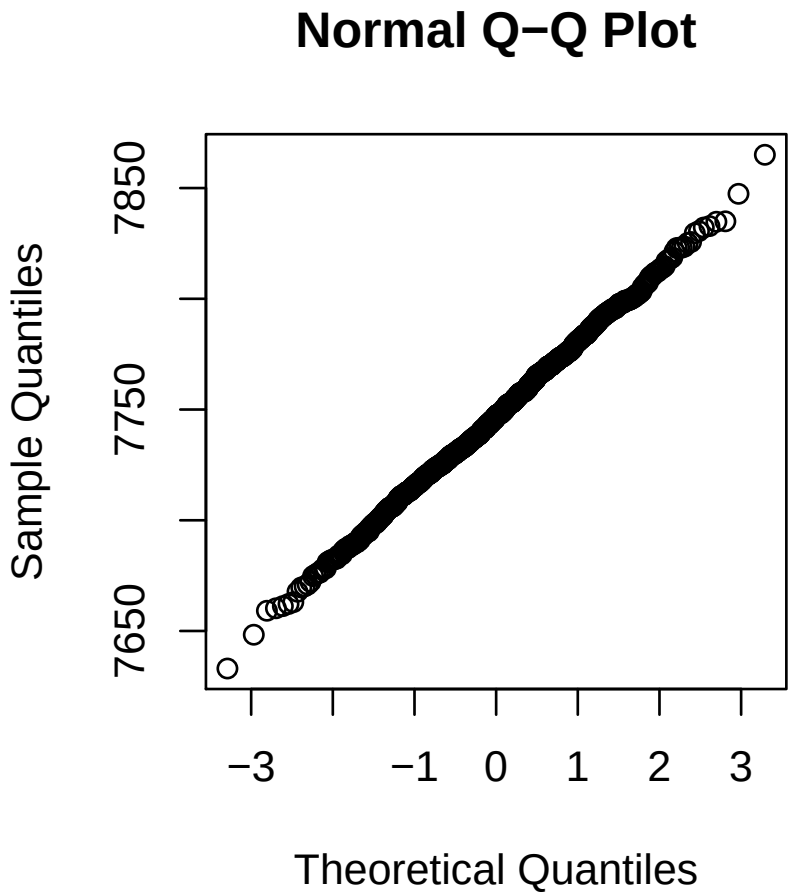
Ejemplo 2: Intervalos de confianza boostrap-t III

- La distribución de los precios no es normal, y está 'contaminada' por diferentes tipos de propiedades. Para intentar corregir este efecto, consideremos un estimador robusto del parámetro de localización, que puede ser la media recortada del 25 % que es la media del 50 % de los datos que están en el centro de la distribución.Con bootstrap estimemos su distribución.

```
n <- length(vxm2)
mean(vxm2,trim = .25)

[1] 7747.663

z <- bootstrap(data = vxm2, statistic = function(x)mean(x,trim=.25),R = 1000)
z

Call:
bootstrap(data = vxm2, statistic = function(x) mean(x, trim = 0.25),
    R = 1000)
Replications: 1000

Summary Statistics:
     Observed       SE    Mean       Bias
stat 7747.663 32.99285 7747.37 -0.2930763

par(mfrow=c(1,2))
hist(z)
qqnorm(z$replicates)
```

Ejemplo 2: Intervalos de confianza boostrap-t IV





Ejemplo 2: Intervalos de confianza boostrap-t V

- La distribución boostrap es aproximadamente normal, tiene sesgo pequeño relativo al valor de la media, y tiene un error estándar de $zstats$SE. Entonces podemos calcular un intervalo de confianza con la expresión $estadistica \pm t * SE_{boot}$. Podemos usar este intervalo porque la estadística muestral es aproximadamente normal.

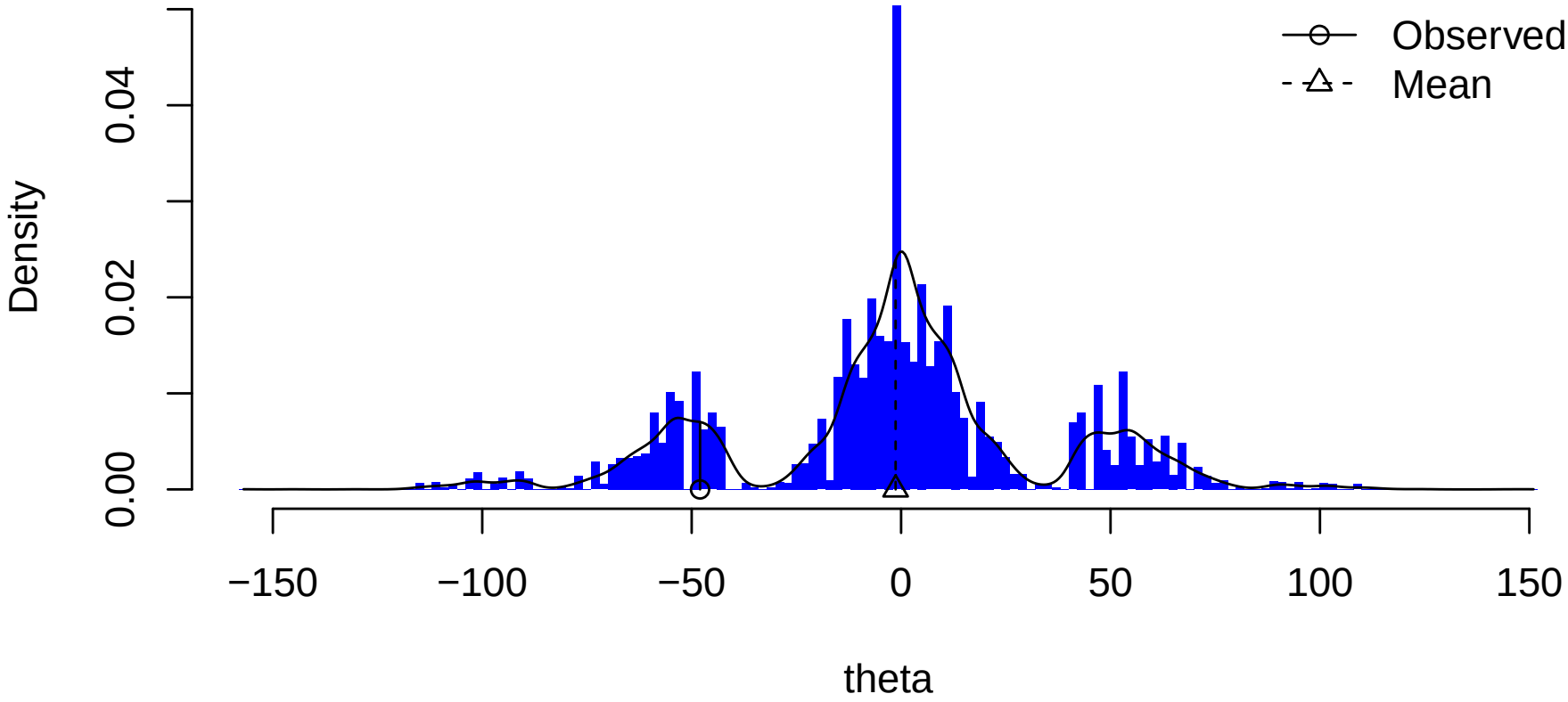
```
z$stats$Mean + c(-1,1)*qt(.975,n-1,lower.tail=T)*z$stats$SE

[1] 7682.696 7812.045

CI.t(z,probs=c(0.025,0.975))

         2.5%    97.5%
stat 7682.989 7812.338
```

Ejemplo 3 Comparación de medias (Efron & Tibshirani) I

Consideren el siguiente experimento: 7 de 16 ratones fueron seleccionados aleatoriamente para recibir un nuevo tratamiento médico, mientras que los 9 restantes fueron asignados a un grupo de control El tratamiento tiene la intención de prolongar sobrevivencia después de una cirugía de prueba. Los tiempos de sobrevivencia después de la cirugía son los que se reportan aquí.
El objetivo es analizar si el tratamiento prolonga la supervivencia.Para ver esto, podemos calcular un intervalo de confianza para la diferencia de medias entre los tratamientos, y para efectos interesantes, hagámoslo con las medianas, que es una medida de localización más robusta.

```
library(bootstrap) #paquete de Efron y Tibshirani con los datos


Attaching package: 'bootstrap'
The following objects are masked from 'package:resample':
    bootstrap, jackknife

#Función a remuestrar:
theta <- function(x){
median(x[1:9]) - median(x[10:16])
}
z <- resample::bootstrap(c(mouse.c,mouse.t),statistic = theta,R = 10000)
hist(z)
```

Ejemplo 3 Comparación de medias (Efron & Tibshirani) II



Ejemplo 3 Comparación de medias (Efron & Tibshirani) III

```
z

Call:
resample::bootstrap(data = c(mouse.c, mouse.t), statistic = theta,
    R = 10000)
Replications: 10000

Summary Statistics:
      Observed      SE   Mean   Bias
theta      -48 37.3518 -1.309 46.691
```

Vemos que en este caso la distribucion muestral de la diferencia de medianas no es aproximadamente normal. Por lo tanto, no podemos confiar en el intervalo bootstrap-t, se necesita hacer una corrección para tomar en cuenta la no-normalidad de los datos.

```
CI.percentile(z)

      2.5% 97.5%
theta  -94    76
```

Ejemplo 4 Correlación (Efron & Tibshirani) I

El siguiente ejemplo considera una muestra de $n = 15$ escuelas de una población de $N = 82$ escuelas de leyes americanas. Muestra los resultados promedios de dos scores: LSAT y GPA. Los datos están en la variable `law` que se obtiene al cargar el paquete `boostrap`. En este ejercicio el parámetro de interés es el coeficiente de correlación ρ. A partir de la muestra obtenemos el coeficiente de correlación muestral, $\hat{\rho} = 0.776$. Noten también que la correlación muestral es un estimador plug-in.

```
cor(law)
```

```
          LSAT       GPA
LSAT 1.0000000 0.7763745
GPA  0.7763745 1.0000000
```

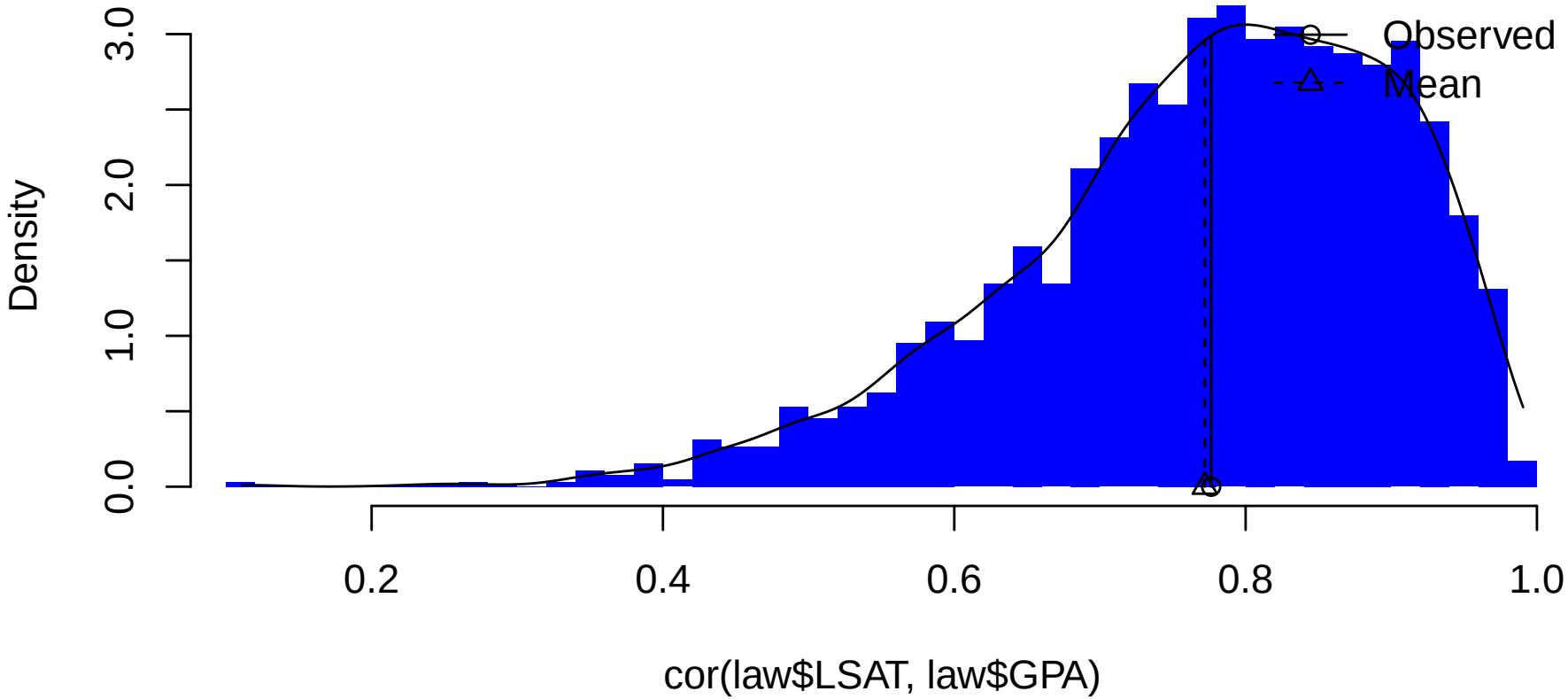
Ahora obtengamos un estimador bootstrap de la desviación estándar del estimador del coeficiente de correlación. Haremos el ejercicio para diferentes tamaños de B

```
z <- list(NULL)
B <- c(25,50,100, 200, 400, 800, 1600, 3200) #tamaños de B
for (i in B) z[[match(i,B)]] <- resample::bootstrap(law, R = i, statistic = cor(law$LSAT,law$GPA))
rbind(B,SE=unlist(lapply(z,function(x)x$stats$SE)))
```

```
         [,1]       [,2]        [,3]        [,4]        [,5]        [,6]
B  25.0000000 50.0000000 100.0000000 200.0000000 400.0000000 800.0000000
SE  0.1503274  0.1192803   0.1406207   0.1193588   0.1305538   0.1257765
           [,7]         [,8]
B  1600.0000000 3200.0000000
SE    0.1311334    0.1304357
```

```
hist(z[[8]])
```

Ejemplo 4 Correlación (Efron & Tibshirani) II



Ejemplo 4 Correlación (Efron & Tibshirani) III

```
z[[8]] #Considerando la versión con mayor tamaño de muestra

Call:
resample::bootstrap(data = law, statistic = cor(law$LSAT, law$GPA),
    R = i)
Replications: 3200

Summary Statistics:
                        Observed        SE      Mean         Bias
cor(law$LSAT, law$GPA) 0.7763745 0.1304357 0.7719992 -0.004375252
```

La distribución muestral de $\hat{\rho}$ no es normal.Podemos calcular intervalos de confianza con los cuantiles de la distribución bootstrap o $t-$boostrap

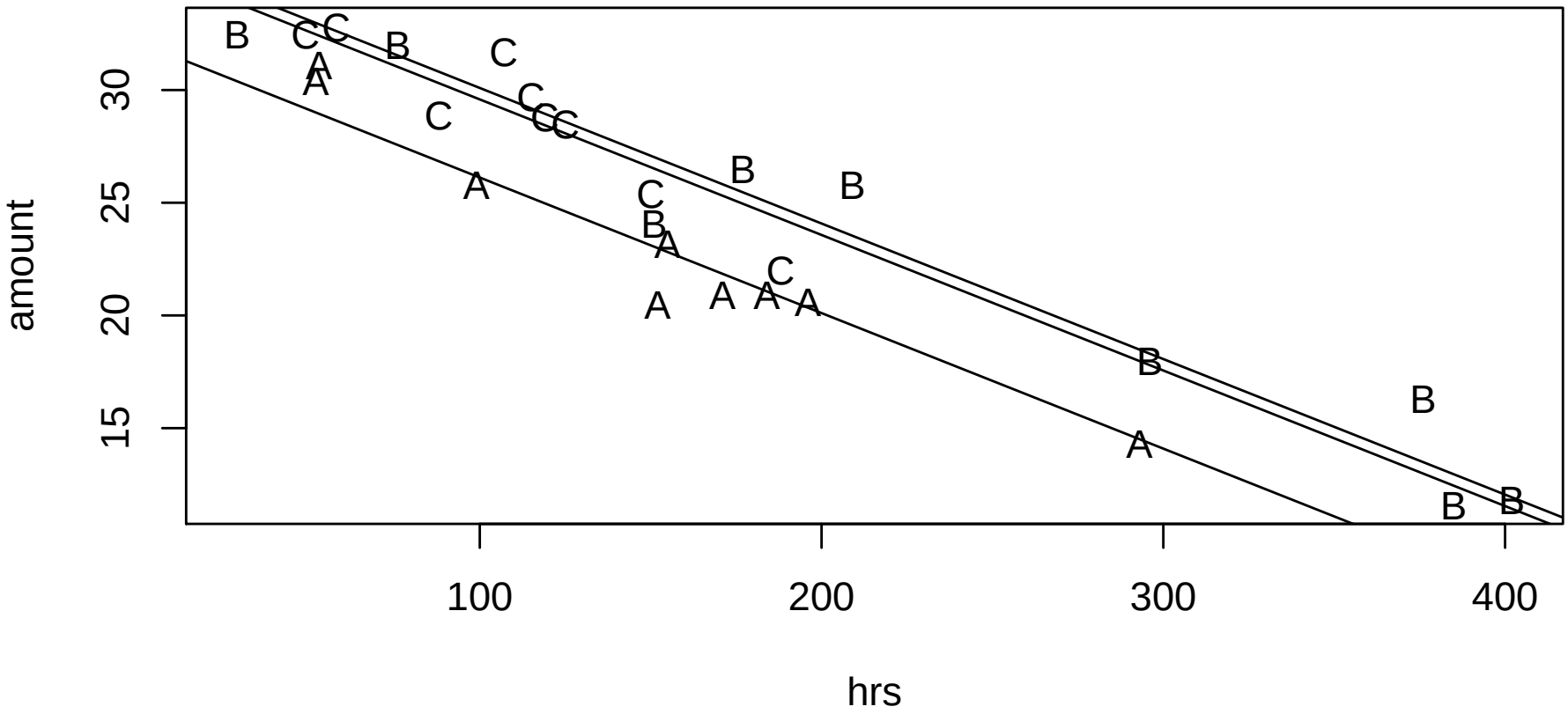
```
resample::CI.percentile(z[[8]])

                           2.5%    97.5%
cor(law$LSAT, law$GPA) 0.429496 0.969644

resample::CI.t(z[[8]])

                            2.5%    97.5%
cor(law$LSAT, law$GPA) 0.4966178 1.056131
```

Ejemplo 5 Regresión (Efron & Tibshirani) I

En este ejemplo se considera el siguiente problema. Un dispensador médico que libera de manera continua una hormona anti-inflamatoria se prueba en 27 sujetos. La variable de respuesta y es la cantidad de hormona que permanece en el dispensador cuando se desgasta. Los predictores son $lot =$ lote de manufactura del dispensador, y $hrs =$ número de horas de uso del dispositivo antes del desgaste.
Los datos `hormone` son del paquete `bootstrap`

```
plot(amount ~ hrs, data=hormone, pch = Lot) #Gráfica de los datos
m1 <- lm(amount ~ hrs + factor(Lot), data=hormone)
abline(a = m1$coefficients[1], b = m1$coefficients[2]) #linea para A
abline(a = m1$coefficients[1] + m1$coefficients[3], b = m1$coefficients[2]) #linea para B
abline(a = m1$coefficients[1] + m1$coefficients[4], b = m1$coefficients[2]) #linea para C
```

Ejemplo 5 Regresión (Efron & Tibshirani) II



Ejemplo 5 Regresión (Efron & Tibshirani) III

```
summary(m1)


Call:
lm(formula = amount ~ hrs + factor(Lot), data = hormone)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9245 -1.0626 -0.1304  0.8544  2.8061

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.131595   0.748277  42.941  < 2e-16 ***
hrs          -0.060136   0.003474 -17.310 1.10e-14 ***
factor(Lot)B  3.973500   0.809686   4.907 5.87e-05 ***
factor(Lot)C  3.465729   0.769123   4.506 0.000159 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.605 on 23 degrees of freedom
Multiple R-squared:  0.945,Adjusted R-squared:  0.9378
F-statistic: 131.8 on 3 and 23 DF,  p-value: 1.254e-14
```

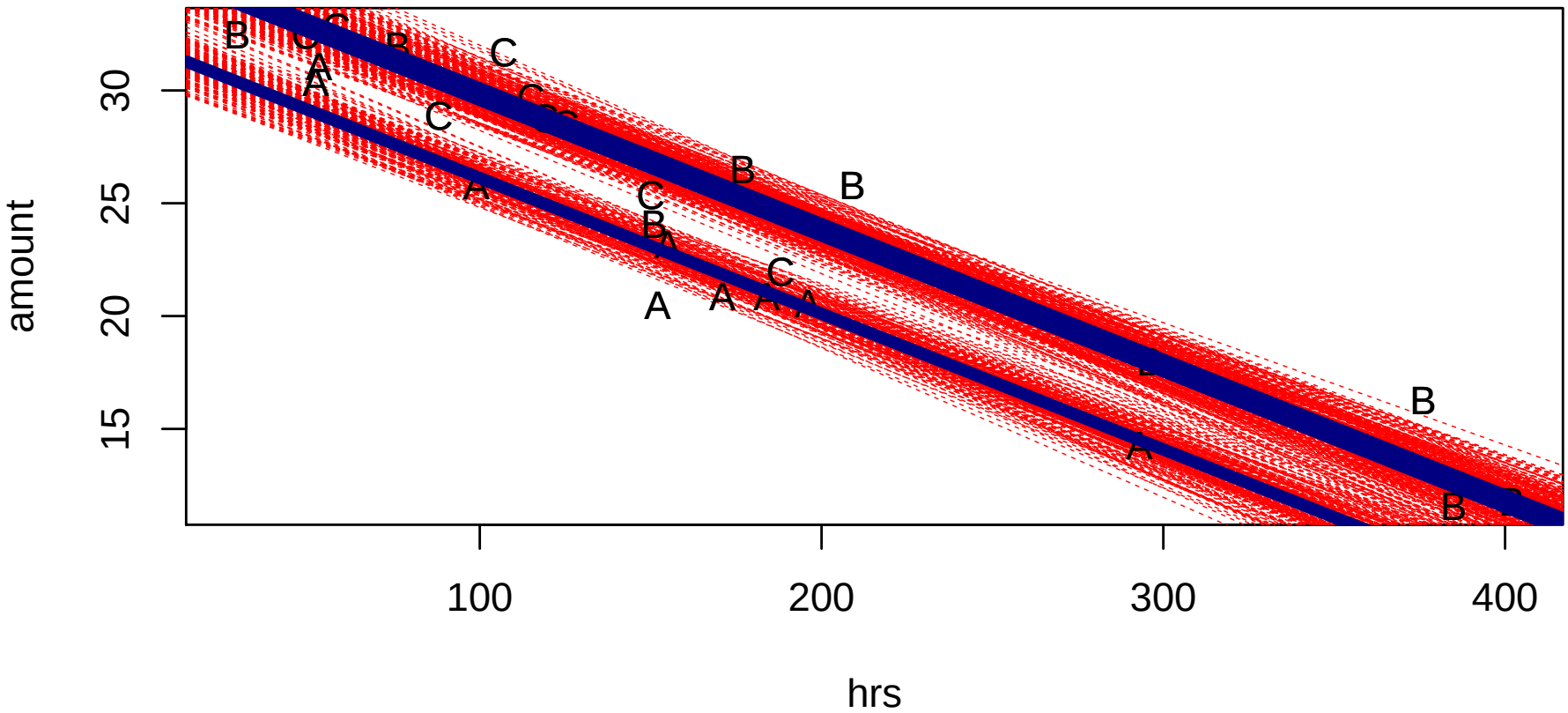
Primera versión de bootstrap: sobre los residuales

Una forma de hacer boostrap es en los residuales y con ellos calcular valores boostrap de la respuesta. Con el estimador de β. Noten que aquí se deja fijo los valores de los predictores y sólo se cambia la respuesta. Esto es porque se considera que los predictores no son aleatorios.

Ejemplo 5 Regresión (Efron & Tibshirani) IV

```
regres <- function(residuos,x,beta){
y <- x %*% beta + residuos
lm(y ~ x - 1)$coefficients
}
z <- resample::bootstrap(m1$residuals,statistic = regres,
     args.stat = list(x=model.matrix(~ hrs + factor(Lot) ,
     data=hormone), beta=m1$coefficients))
z

Call:
resample::bootstrap(data = m1$residuals, statistic = regres,
    args.stat = list(x = model.matrix(~hrs + factor(Lot), data = hormone),
        beta = m1$coefficients))
Replications: 10000

Summary Statistics:
                 Observed          SE        Mean          Bias
x(Intercept)  32.13159495 0.700060089 32.12169986 -9.895089e-03
xhrs          -0.06013605 0.003243465 -0.06009957  3.648416e-05
xfactor(Lot)B  3.97349967 0.757722098  3.98295013  9.450463e-03
xfactor(Lot)C  3.46572938 0.714828274  3.46933223  3.602852e-03

z$stats$SE*sqrt(27/25) #corrigiendo el factor plug-in

[1] 0.727523785 0.003370707 0.787447903 0.742871333

plot(amount ~ hrs, data=hormone, pch = Lot) #Gráfica de los datos
for(i in 1:200) abline(a = z$replicates[i,1], b = z$replicates[i,2],col="red",lwd=0.5,lty=2) #A
for(i in 1:200) abline(a = z$replicates[i,1] + z$replicates[i,3] , b = z$replicates[i,2],col="red",lwd=0.5,lty=2) #B
for(i in 1:200) abline(a = z$replicates[i,1] + z$replicates[i,4] , b = z$replicates[i,2],col="red",lwd=0.5,lty=2) #C
points(hormone$hrs,hormone$amount,pch=hormone$Lot)
abline(a = m1$coefficients[1],b = m1$coefficients[2], col="navy", lwd=5) #linea para A
abline(a = m1$coefficients[1] + m1$coefficients[3], b = m1$coefficients[2], col="navy", lwd=5) #linea para B
abline(a = m1$coefficients[1] + m1$coefficients[4], b = m1$coefficients[2], col="navy", lwd=5) #linea para C
```

Ejemplo 5 Regresión (Efron & Tibshirani) V



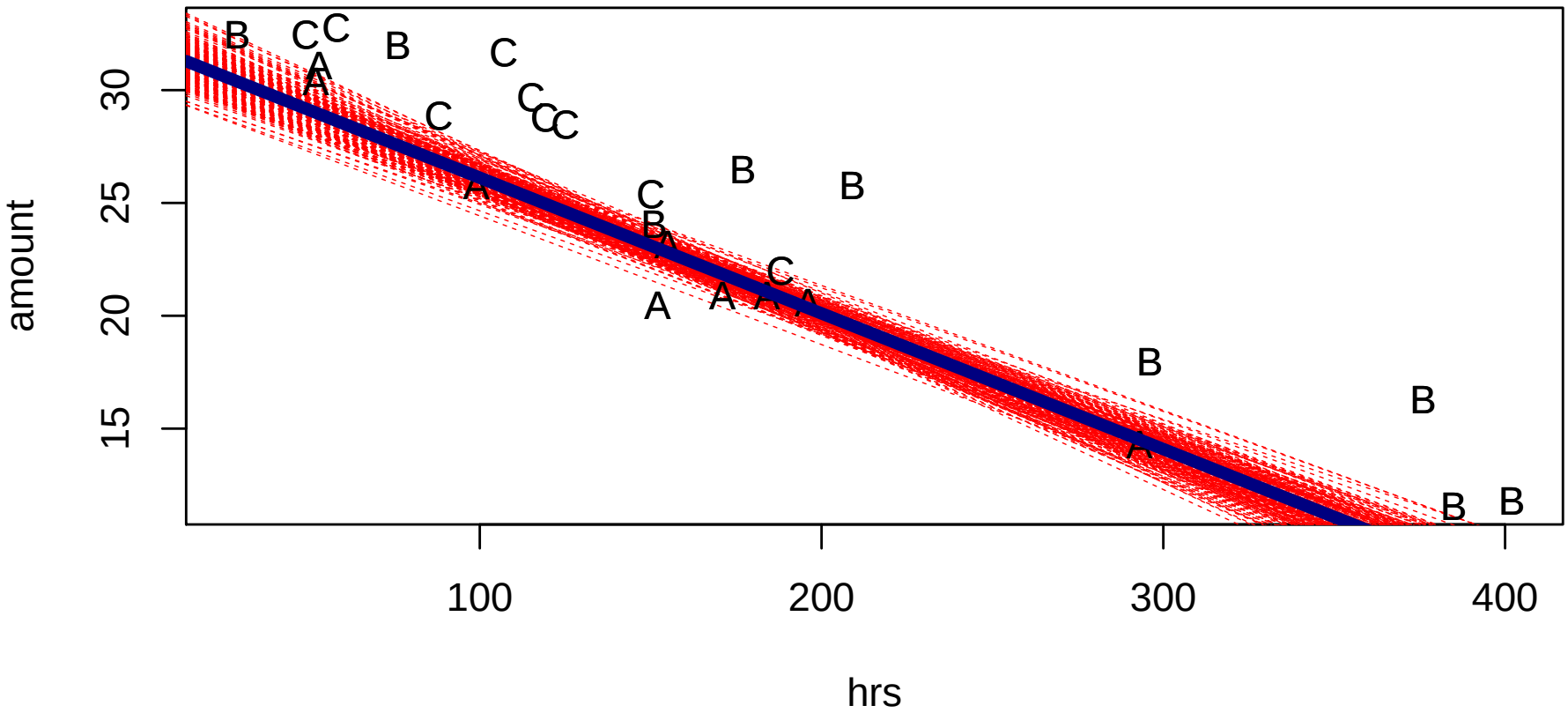
Segunda versión de bootstrap sobre las observaciones

Ejemplo 5 Regresión (Efron & Tibshirani) VI

La segunda forma de hacer boostrap es muestrear sobre las observaciones de los datos, variando la respuesta, pero al igual que en el caso anterior, se mantiene fija la matriz de predictores.

```
#función a aplicar el bootstrap:
theta <- function(x,datos){
         lm(amount ~ hrs + factor(Lot), data=datos[x,])$coef}
z2 <- bootstrap(1:dim(hormone)[1],200,theta,hormone)

z2[[1]][,1]


 (Intercept)          hrs factor(Lot)B factor(Lot)C
 32.11258838  -0.05925921   3.69768894   3.20079671


plot(amount ~ hrs, data = hormone, pch = Lot) #Gráfica de los datos
for(i in 1:200) abline(a = z2$thetastar[1,i], b =               z2$thetastar[2,i],col="red",lwd=0.5,lty=2)
points(hormone$hrs, hormone$amount, pch = hormone$Lot)
abline(a = m1$coefficients[1], b = m1$coefficients[2], col = "navy", lwd = 5)
```

Ejemplo 5 Regresión (Efron & Tibshirani) VII



Boostrap paramétrico I

Este tipo de bootstrap surge cuando en lugar de estimar F en $\hat{\theta} = T(F)$ usamos una distribución paramétrica. Si F_θ depende de un parámetro θ y $\hat{\theta}$ es un estimado de θ, entonces se muestrea de $F_{\hat{\theta}}$ en lugar de $\hat{F}$.

Pruebas de permutación

Pruebas de Permutación

o pruebas de aleatorización

- Enfoque no paramétrico a pruebas de hipótesis estadísticas
- El paradigma usual:

Pruebas de hipótesis

- Define la hipótesis a probar: $H_0 : \theta \in \Theta$ vs. $\theta \notin \Theta$
- Observar datos: $x_1, \ldots, x_n$
- Define la estadística de prueba: $T_0 \sim F_0$ bajo H_0
- Determina si el valor de la estadística observada da un valor dentro de la región de rechazo: $T \in R_\alpha \implies$ rechaza H_0.

- La hipótesis nula induce una distribución de probabilidad sobre la estadística de prueba. Esta distribución usualmente hace que las muestras sean *indistinguibles e intercambiables*.
- Bajo esta distribución, se puede construir la **distribución de permutación**.

Ejemplo Pruebas de permutación I

- Se tienen dos muestras y se quieren comparar poblaciones. La hipótesis nula es $H_0 : \mu_1 = \mu_2$. Los datos observados son los siguientes:

```
x <- c(4714,4601,4696,4896,4905,4870,4987,5144,3962,4066,4561,4626,4924,5096,4321)
y <- c(4295,4271,4326,4530,4618,4779,4752,4744,3764,3797,4401,4339,4700)
```

- La estadística de prueba se basa en la distribución t (usualmente suponiendo que las muestras provienen de una distribución normal) y se consideran dos casos:
  - Varianzas iguales pero desconocidas: $\sigma_1 = \sigma_2$. En este caso, la estadística de prueba es de la forma:

$$t_0 = \frac{\bar{x} - \bar{y}}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{(n_1+n_2-2)}$$

    donde $S_p = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ es la varianza combinada.
  - Varianzas diferentes y desconocidas: $\sigma_1 \neq \sigma_2$. En este caso, la estadística de prueba es de la forma (aproximada):

$$t_0 = \frac{\bar{x} - \bar{y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t_{(\nu)}$$

    y los grados de libertad se obtienen usando la aproximación de Satterthwaite:

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2$$

Ejemplo de pruebas de permutación I

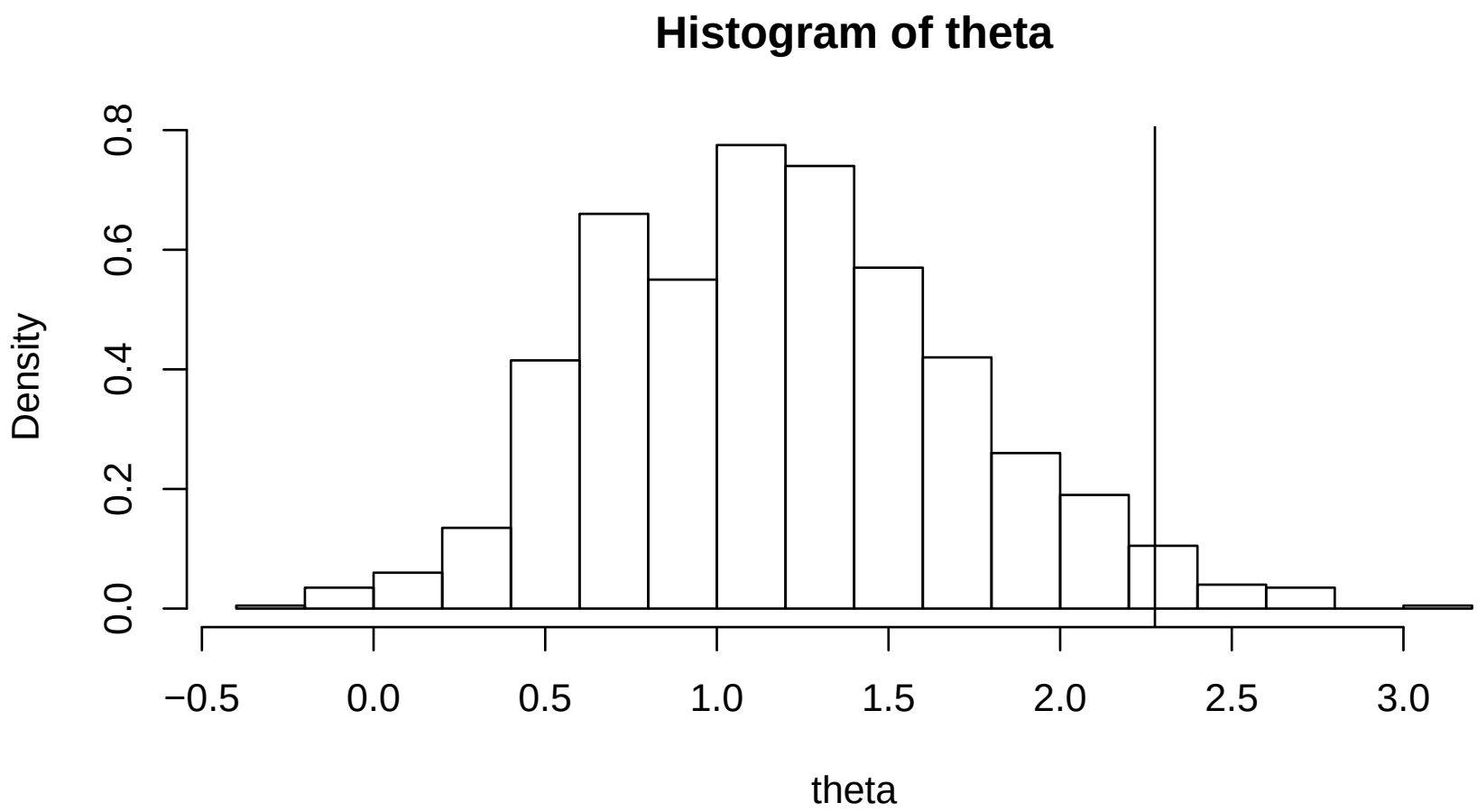
- Para aplicar una prueba de permutación, observamos que bajo la hipótesis nula, esperaríamos que las diferencias de medias muestrales $\hat{\theta} = \bar{x} - \bar{y}$ fueran muy pequeñas. Si H_0 no es cierta, entonces tenderíamos a observar valores "grandes" de $\hat{\theta}$. Habiendo observado $\hat{\theta}$, definimos el $p - value$ como $pv = P(\hat{\theta}^* \geq \hat{\theta}|H_0)$ (también se denota como ASL, *achieved significance level*).
- Si H_0 es cierta, cualquiera de las observaciones pudo haber venido de cualquiera de las poblaciones. Podemos entonces combinar *todas* las observaciones $n_1 + n_2$ observaciones de las dos muestras, y extraer *sin reemplazo* n_1 de ellas para crear el primer grupo, y las n_2 observaciones restantes forman el otro grupo. Con estas nuevas muestras calculamos la estadística de prueba y repetimos el proceso B veces.

```
estadística <- function(x,n1,n2){
        s1 <- sample(x,n1,replace=F)
        s2 <- x[-which(s1 %in% x)]
        return((mean(s1)-mean(s2))/sqrt(var(s1)/n1+var(s2)/n2))
}
theta <- NULL
B <- 1000
for(i in 1:B) theta[i] <- estadística(c(x,y),length(x),length(y))
ASL <-sum(theta > (mean(x)-mean(y))/sqrt(var(x)/length(x)+var(y)/length(y)))/B   #p-value
ASL

[1] 0.041
```

Con este procedimiento ya no tenemos que aproximar los grados de libertad.

Ejemplo de pruebas de permutación II

```r
hist(theta,breaks=20,probability=T)
abline(v=quantile(theta,.975)) #agrega una linea en el quantile del p-value.
```



Conceptos detrás de las pruebas de permutación

- Al considerar las permutaciones de las observaciones de los dos grupos hay $\binom{n_1+n_2}{n_1}$ posibles arreglos.
- Bajo la hipótesis nula, cada posible permutación tiene la misma probabilidad $1/\binom{n_1+n_2}{n_1}$ de ser seleccionada.
- Por lo tanto hay $\binom{n_1+n_2}{n_1}$ replicaciones de permutación $\hat{\theta}^*$. El p-value de permutación se define como la probabilidad de que una réplica $\hat{\theta}*$ exceda el valor observado $\hat{\theta}$:

$$p - val = \#\{\hat{\theta}^* \geq \hat{\theta}\} / \binom{n_1 + n_2}{n_1}$$

- En la práctica, usualmente el valor del p-value se aproxima por muestreo de Monte Carlo.

Validación Cruzada

Validación cruzada: ideas básicas I

La validación cruzada (cross-validation) es una forma de medir el desempeño predictivo de un modelo estadístico.

- Las estadísticas de ajuste de un modelo no son guía adecuada del poder predictivo del modelo. Por ejemplo, en regresión una R^2 alta no necesariamente indican que el modelo es bueno para predecir (se pueden incluir más términos para mejorar R^2 pero su poder predictivo empeora con el número de términos)

El enfoque de la validación cruzada es dividir los datos disponibles en dos conjuntos: un conjunto de *entrenamiento*, que se usa para estimar el modelo y un conjunto de *prueba*, en el que se evalúa el modelo y se obtiene un estimador del error de ajuste del modelo.
Hay diversas maneras de hacer este procedimiento:

- **uno-afuera**. se usan $n-1$ datos para estimar el modelo. El modelo se prueba en el dato que se dejó afuera. Esto se puede realizar n veces se utilizan los errores $e_i^* = y_i - \hat{y}_i$ para calcular el error cuadrático medio de validación cruzada: $MSE_{cv} = \frac{\sum_{i=1}^{n} e_i^2}{n}$
- **k-afuera**.
- **Muestreo aleatorio**

La validación cruzada se utiliza para:

- Seleccionar variables a incluir en el modelo.

Validación cruzada: ideas básicas II

- Seleccionar el tipo de función de predicción a utilizar.
- Seleccionar los parámetros en la función de predicción.
- Comparar diferentes predictores.