Tema 4. Estimación de errores estándar mediante remuestreo

Limitaciones del TCL

Se toma el ejemplo de una distribución binomial con $n = 25$ en los casos de $p = 0.25$ y $p = 0.9$.

```
n = 25

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

```
head(p09)
```

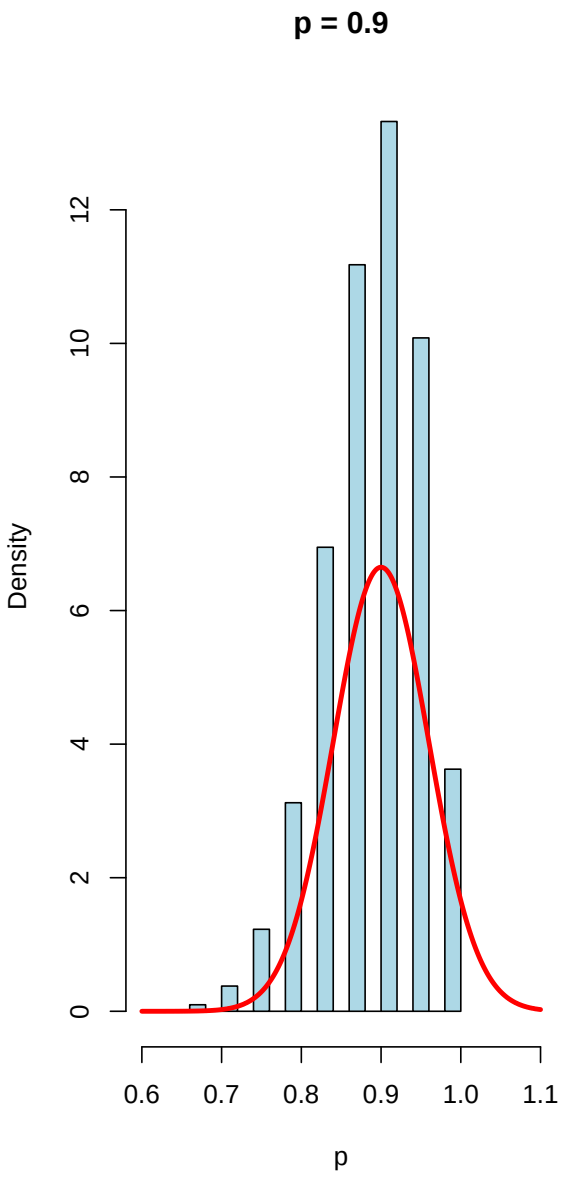
```
[1] 0.96 0.84 0.96 0.96 0.92 0.80
```

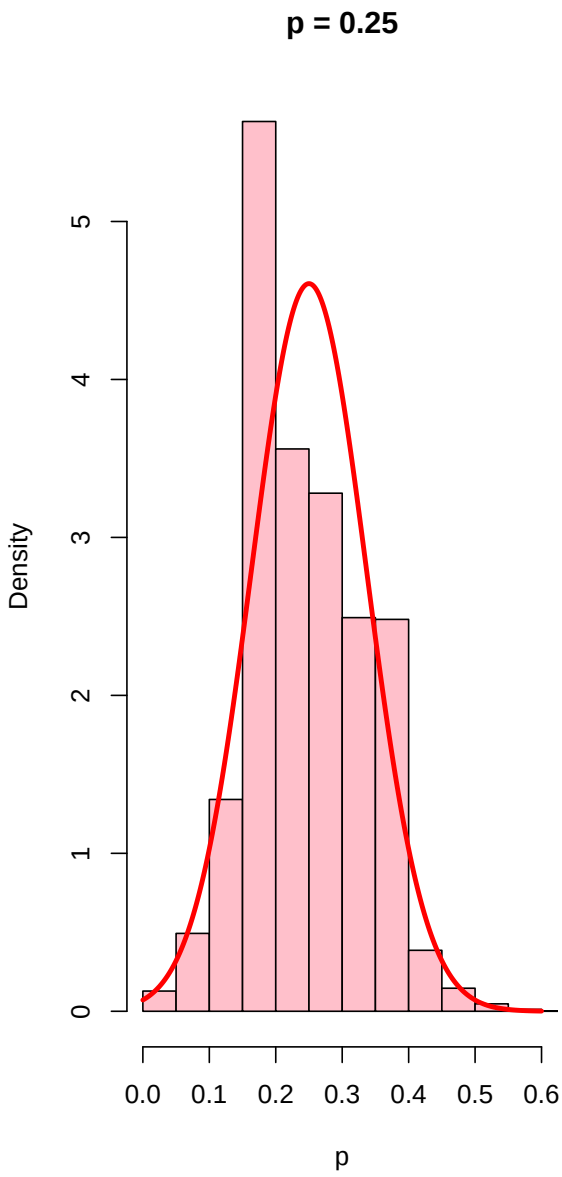
```
head(p025)
```

```
[1] 0.12 0.36 0.28 0.20 0.12 0.16
```

Para el caso de $p = 0.9$ la aproximación a la normal por el TCL **no** es muy buena.

```
par(mfrow = c(1, 2))

hist(p09, prob = T, xlim = c(0.6, 1.1), col = "lightblue", xlab = "p", main = "p = 0.9")
xs1 = seq(0.6, 1.1, 1e-04)
ys1 = dnorm(xs1, 0.9, sqrt((0.9 * 0.1)/n))
lines(xs1, ys1, lwd = 3, col = "red")

hist(p025, prob = T, xlim = c(0, 0.6), col = "pink", xlab = "p", main = "p = 0.25")
xs2 = seq(0, 0.6, 1e-04)
ys2 = dnorm(xs2, 0.25, sqrt((0.25 * 0.75)/n))
lines(xs2, ys2, lwd = 3, col = "red")
```





La correlación entre GPA y LSAT es

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

```
[1] 0.7763745
```

Siguiendo a Efron y Tibshirani (1993) tiene un error estándar igual a

```
(se = (1 - lawCor^2)/sqrt(dim(law)[1] - 3))
```

```
[1] 0.1146741
```

```
c((lawCor - 1.96 * se), min(1, (lawCor + 1.96 * se)))
```

```
[1] 0.5516133 1.0000000
```

Alternativamente se puede usar la librería `psychometric`

```
psychometric::CIr(lawCor, dim(law)[1])
```

```
[1] 0.4385108 0.9219648
```

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

```
samplesize = 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] 0.1340337
```

Desde el punto de vista clásico, el error estándar del estimador de la correlación (asumiendo aproximadamente normalidad) es igual a

$$SE_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

se puede estimar así:

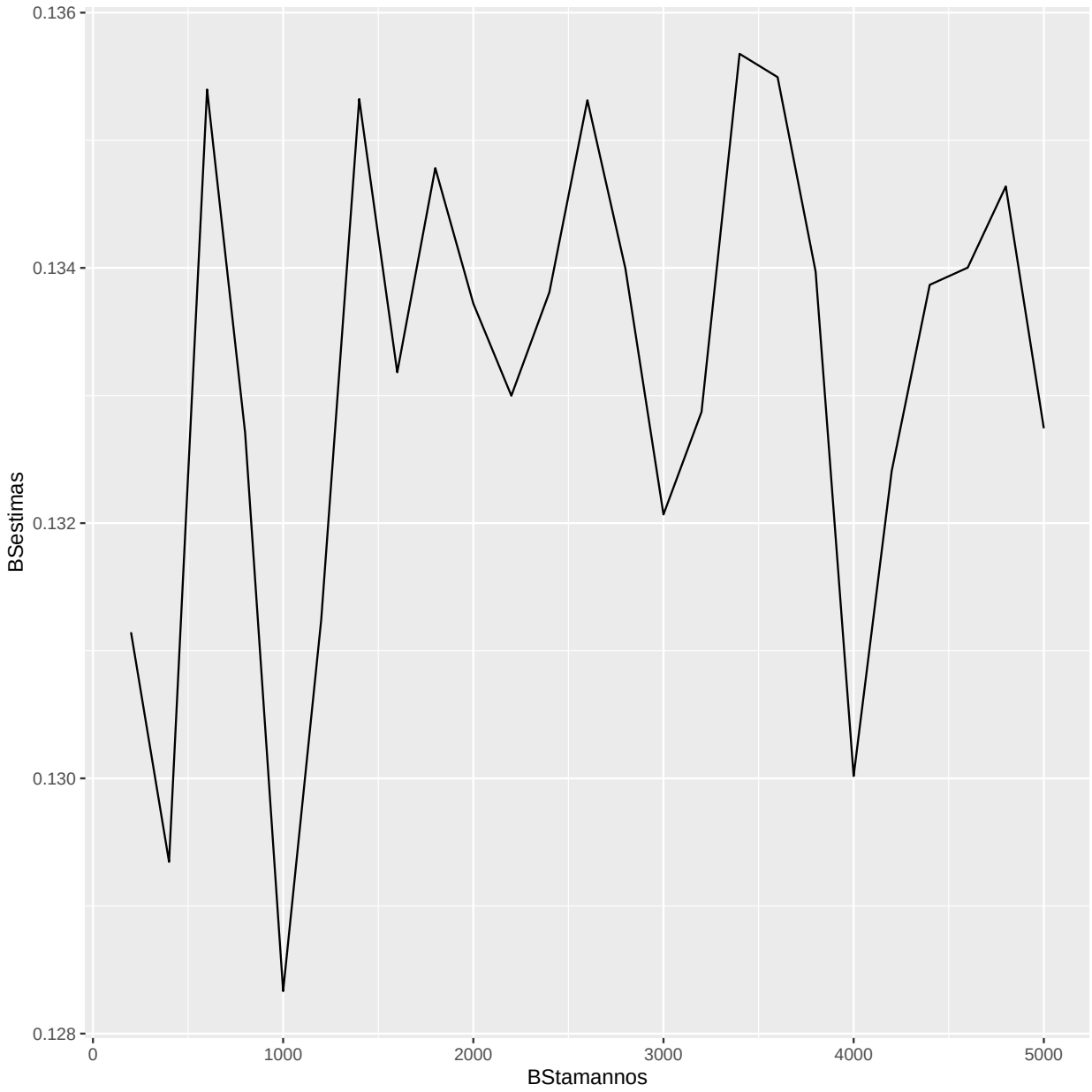
```r
# Asumes SE_r = sqrt((1-r^2)/(n-2))

cor.test.plus = function(x) {
    list(x, Standard.Error = unname(sqrt((1 - x$estimate^2)/x$parameter)))
}
```

```r
library(bootstrap)
cor.test.plus(cor.test(law$GPA, law$LSAT))
```

```
[[1]]

    Pearson's product-moment correlation

data:  law$GPA and law$LSAT
t = 4.4413, df = 13, p-value = 0.0006651
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4385108 0.9219648
sample estimates:
      cor
0.7763745


$Standard.Error
[1] 0.174806
```

¿Cómo converge de rápido el estimador bootstrap?

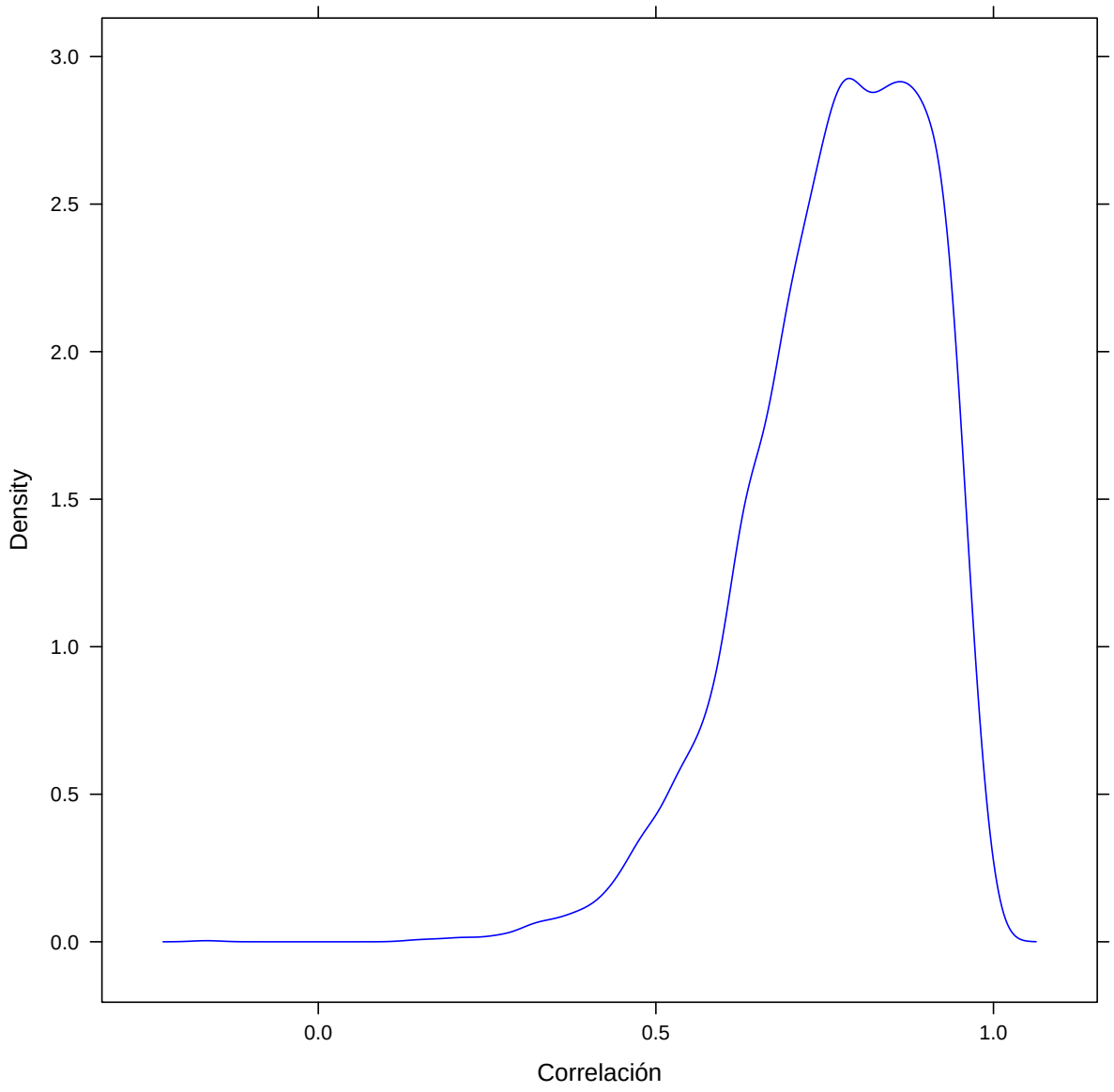
```r
samplesize = dim(law)[1]
ind = 1:samplesize
lawBS = function(B) sd(replicate(B, {
    indB = sample(ind, replace = TRUE)
    with(law[indB, ], cor(GPA, LSAT))
}))

BStamannos = seq(200, 5000, 200)
BSestimas = sapply(BStamannos, lawBS)

library(ggplot2)
qplot(BStamannos, BSestimas, geom = "path")
```



Se obtiene la distribución empírica muestral de la población $\hat{F}(\hat{\theta}^*)$

```
ind1 = 1:dim(law)[1]
law.boot = replicate(5000, {
    indB = sample(ind1, size = length(ind1), replace = TRUE)
    with(law[indB, ], cor(GPA, LSAT))
})
```

```
library(latticeExtra)

densityplot(~law.boot, plot.points = FALSE, xlab = "Correlación", col = "blue")
```



El estimador bootstrap del error estándar a partir de la muestra $se_{\hat{F}}(\hat{\rho})$ es

```
sd(law.boot)
```

```
[1] 0.1323102
```

Ejemplo de los institutos de máster en leyes

```
library(bootstrap)

paraBoot = function(datos) {
    ndatos = dim(datos)[1]
    sigma = cov(datos)
    mu = rapply(datos, mean)

    c = sqrt(prod(diag(sigma))/sigma[1, 2]^2 - 1)
    z1 = rnorm(ndatos)
    z2 = rnorm(ndatos)
    x = mu[1] + sqrt(sigma[1, 1]) * z1
    y = mu[2] + sqrt(sigma[2, 2]) * (z1 + c * z2)/sqrt(1 + c^2)
    cbind(x, y)
}
```

Alternativamente, se puede programar con una librería específica que trata la normal multivariante: `mvrnorm`

```
paraBoot = function(datos) {
    ndatos = dim(datos)[1]
    sigma = cov(datos)
    mu = rapply(datos, mean)

    x = MASS::mvrnorm(ndatos, mu = mu, Sigma = sigma)
    return(x)
}
```

Aproximación con bootstrap paramétrico

```
pBoot = replicate(5000, cor(paraBoot(law))[2, 1])

sd(pBoot)
```

```
[1] 0.1185058
```

Con la librería `boot`:

```r
library(boot)

simulaBoot = function(datos, ...) {
    ndatos = dim(datos)[1]
    x = MASS::mvrnorm(ndatos, mu = mu0, Sigma = sigma0)
    return(x)
}

sigma0 = Rfast::mvnorm.mle(as.matrix(law))$sigma
mu0 = Rfast::mvnorm.mle(as.matrix(law))$mu

estadistico = function(x, i) {
    cor(x)[2, 1]
}


saleB = boot(data = law, sim = "parametric", ran.gen = simulaBoot, mle = list(mu0,
    sigma0), statistic = estadistico, R = 1000)
saleB
```

```
PARAMETRIC BOOTSTRAP


Call:
boot(data = law, statistic = estadistico, R = 1000, sim = "parametric",
    ran.gen = simulaBoot, mle = list(mu0, sigma0))


Bootstrap Statistics :
     original       bias    std. error
t1* 0.7763745 -0.006117189   0.1181514
```

```r
boot.ci(saleB, type = "perc")
```

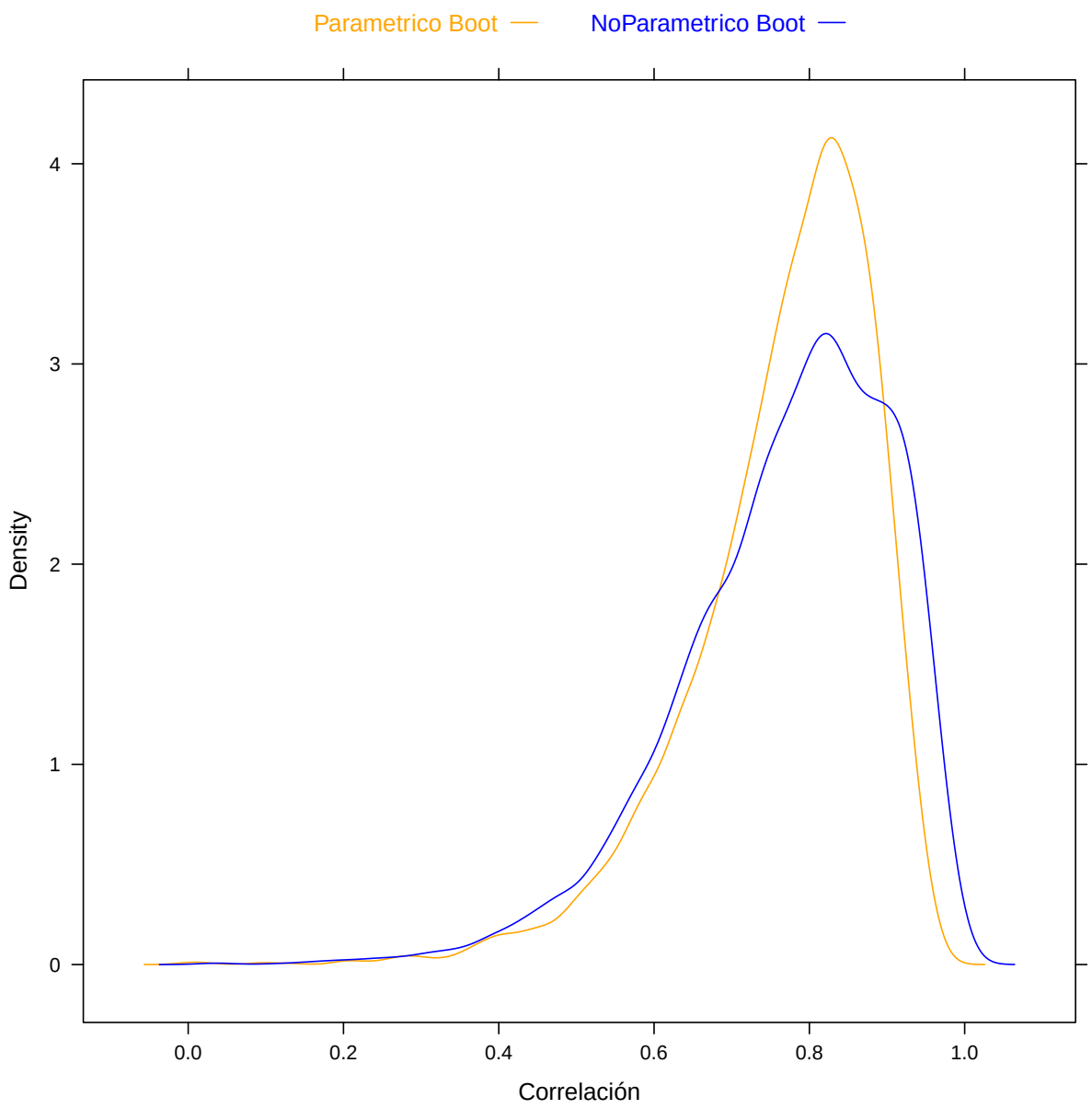
```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = saleB, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.4583,  0.9364 )
Calculations and Intervals on Original Scale
```

Aproximación asintótica

```r
samplesize = dim(law)[1]

(1 - cor(law)[2, 1]^2)/sqrt(samplesize - 3)
```

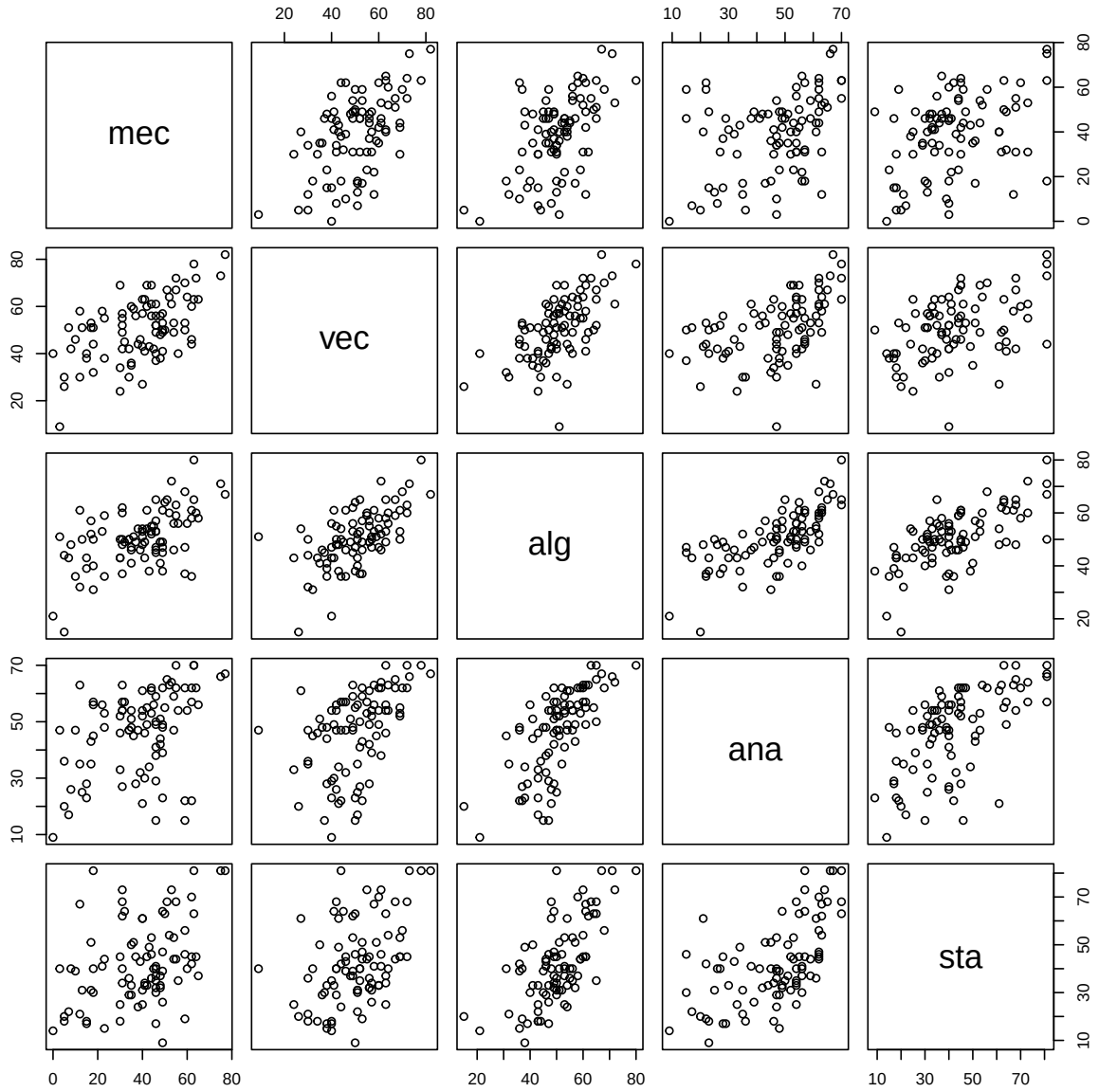
```
[1] 0.1146741
```

```r
library(latticeExtra)

ind1 = 1:dim(law)[1]
law.boot = replicate(5000, {
    indB = sample(ind1, size = dim(law)[1], replace = TRUE)
    with(law[indB, ], cor(GPA, LSAT))
})

densityplot(~pBoot + law.boot, plot.points = FALSE, auto.key = list(columns = 2,
    size = 2, between = 1, col = c("orange", "blue"), text = c("Parametrico Boot",
        "NoParametrico Boot")), xlab = "Correlación", par.settings = list(superpose.line
        ↪  = list(col = c("orange",
    "blue"))))
```



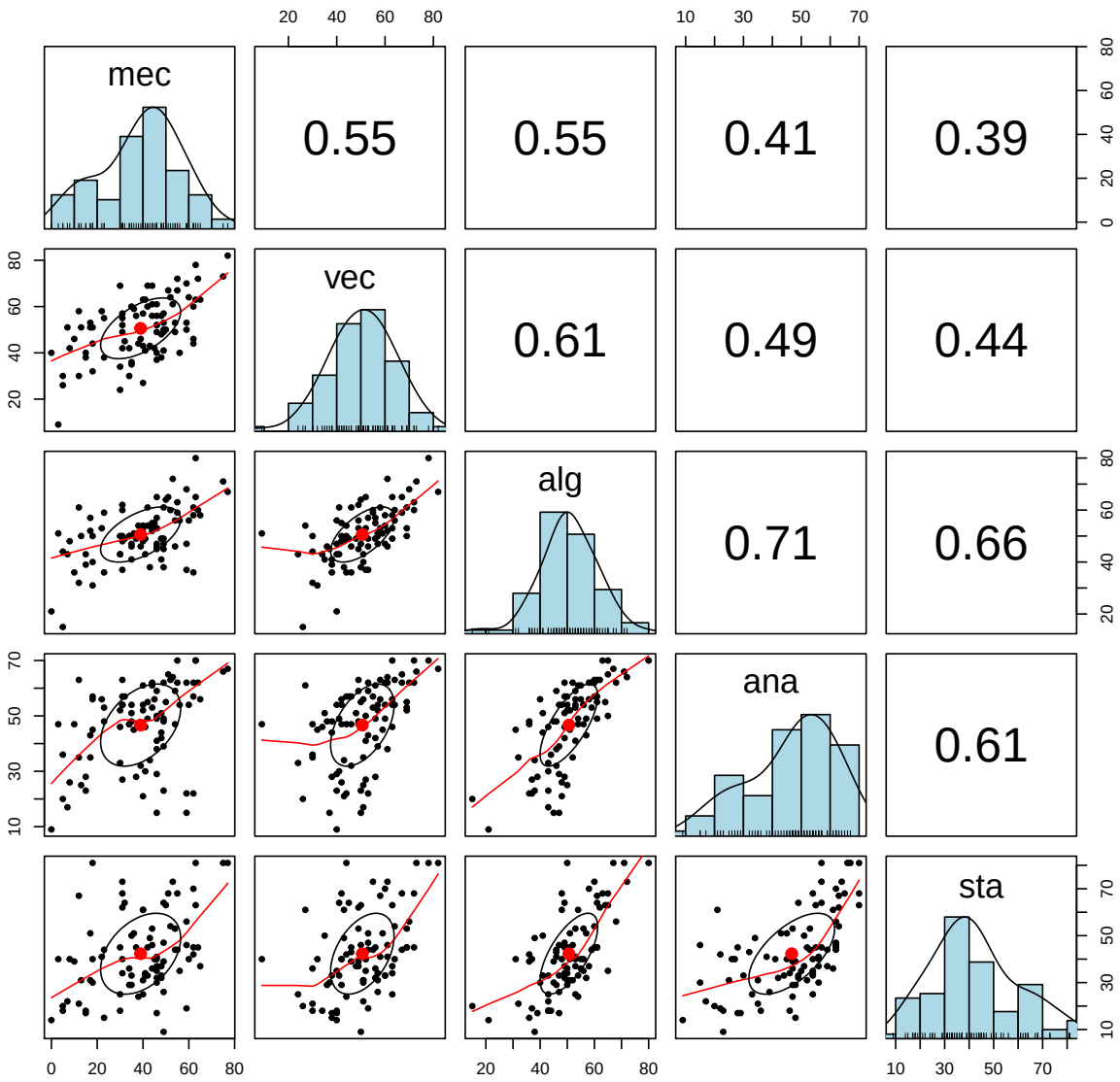
Ejemplo sobre calificaciones de alumnos

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



Alternativamente

```
library(psych)

pairs.panels(scor, method = "pearson", hist.col = "lightblue", density = TRUE,
    ellipses = TRUE, main = "Matriz de variables")
```



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
```

Se calculan los autovalores y autovectores de la matriz de covarianzas.

```
eigen(cov(scor))$values  # Autovalores
```

```
[1] 686.98981 202.11107 103.74731  84.63044  32.15329
```

```
eigen(cov(scor))$vectors  # Autovectores
```

```
           [,1]        [,2]       [,3]         [,4]        [,5]
[1,] -0.5054457  0.74874751 -0.2997888  0.296184264 -0.07939388
[2,] -0.3683486  0.20740314  0.4155900 -0.782888173 -0.18887639
[3,] -0.3456612 -0.07590813  0.1453182 -0.003236339  0.92392015
[4,] -0.4511226 -0.30088849  0.5966265  0.518139724 -0.28552169
[5,] -0.5346501 -0.54778205 -0.6002758 -0.175732020 -0.15123239
```

En componente principales svd es numéricamente más estable que la descomposición por autovectores y autovalores, pero para aplicar bootstrap esta última es mas rápida

```
library(bootstrap)

autovals = eigen(var(scor), symmetric = TRUE, only.values = TRUE)$values
(teta = autovals[1]/sum(autovals))
```

```
[1] 0.619115
```

```
theta = function(ind) {
    vals = eigen(var(scor[ind, ]), symmetric = TRUE, only.values = TRUE)$values
    vals[1]/sum(vals)
}

scor.boot = bootstrap(1:dim(scor)[1], 500, theta)
```

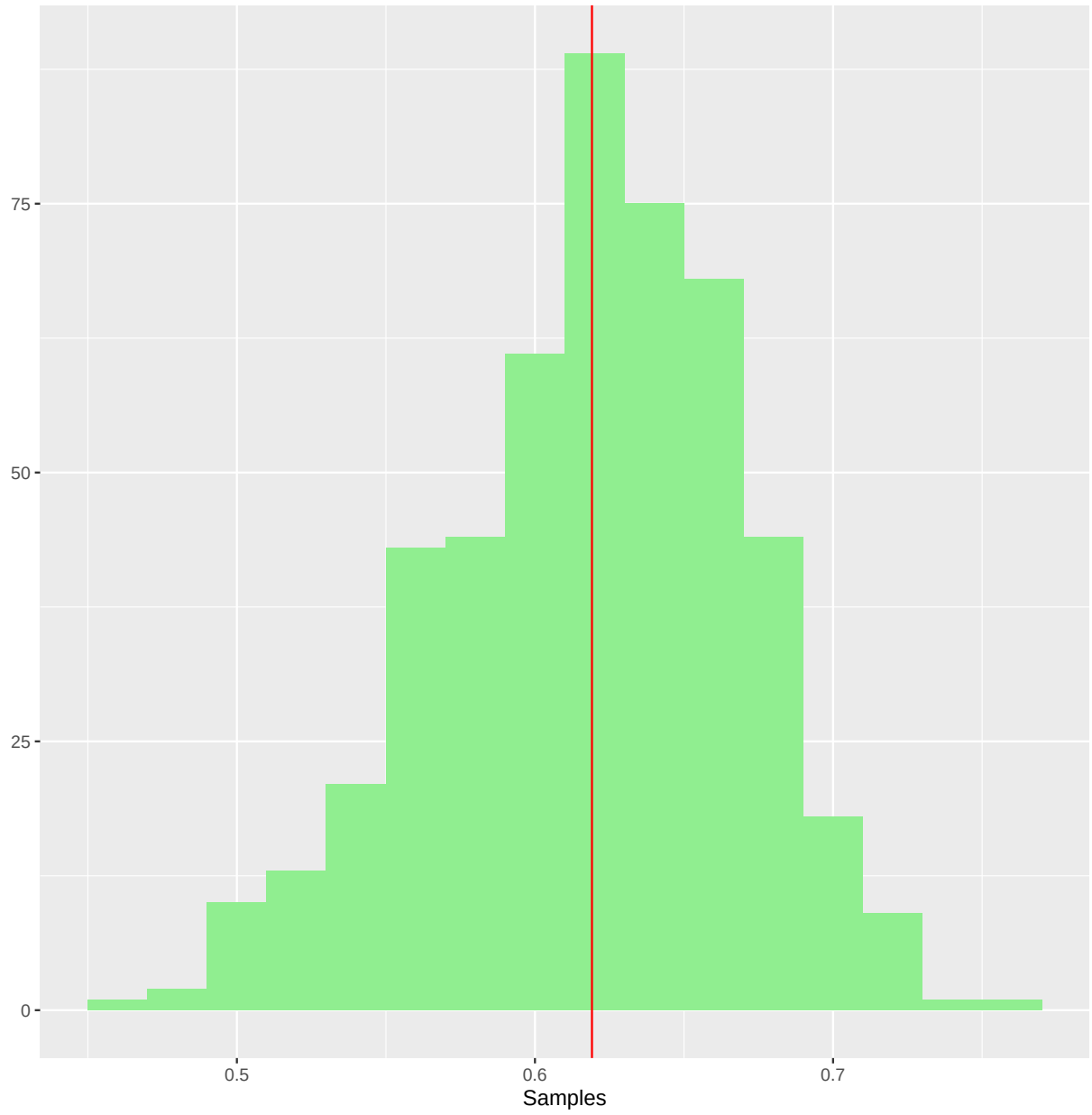
Error estándar del bootstrap

```
sd(scor.boot$thetastar)
```

```
[1] 0.04993071
```

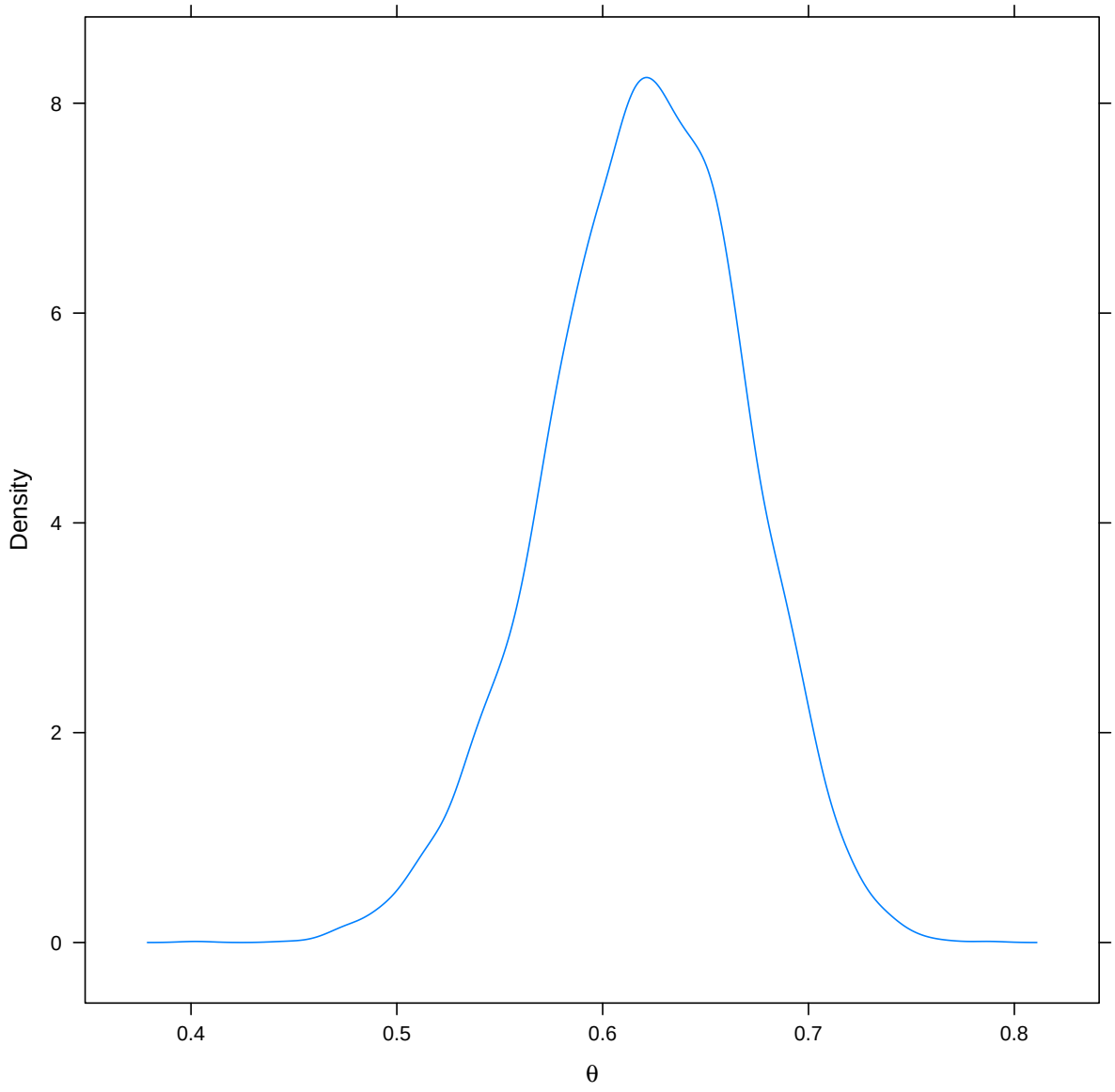
Distribución bootstrap

```
library(ggplot2)

qplot(scor.boot$thetastar, geom = "histogram", binwidth = 0.02, fill = I("lightgreen"),
    xlab = "Samples") + geom_vline(xintercept = teta, col = "red")
```




```
library(bootstrap)
data(scor)
X = scor

eigenTeta = function(X) {
    ee = eigen(cov(X))[["values"]]
    ee[1]/sum(ee)
}

ind = 1:dim(X)[[1]]
```

```
eigendist = replicate(5000, eigenTeta(X[sample(ind, replace = TRUE), ]))

library(latticeExtra)

densityplot(eigendist, plot.points = FALSE, xlab = expression(theta))
```



Tanto $\hat{\mathbf{v}}_1$ como $\hat{\mathbf{v}}_2$ son estadísticos del mismo modo que lo es $\hat{\theta}$, y de este modo se puede aplicar el bootstrap para calcular su variabilidad.

```
library(bootstrap)
data(scor)

X = scor

eigenVec = function(X) {
    ee = eigen(cov(X))[["vectors"]]
    return(cbind(ee[, 1], ee[, 2]))
}
```

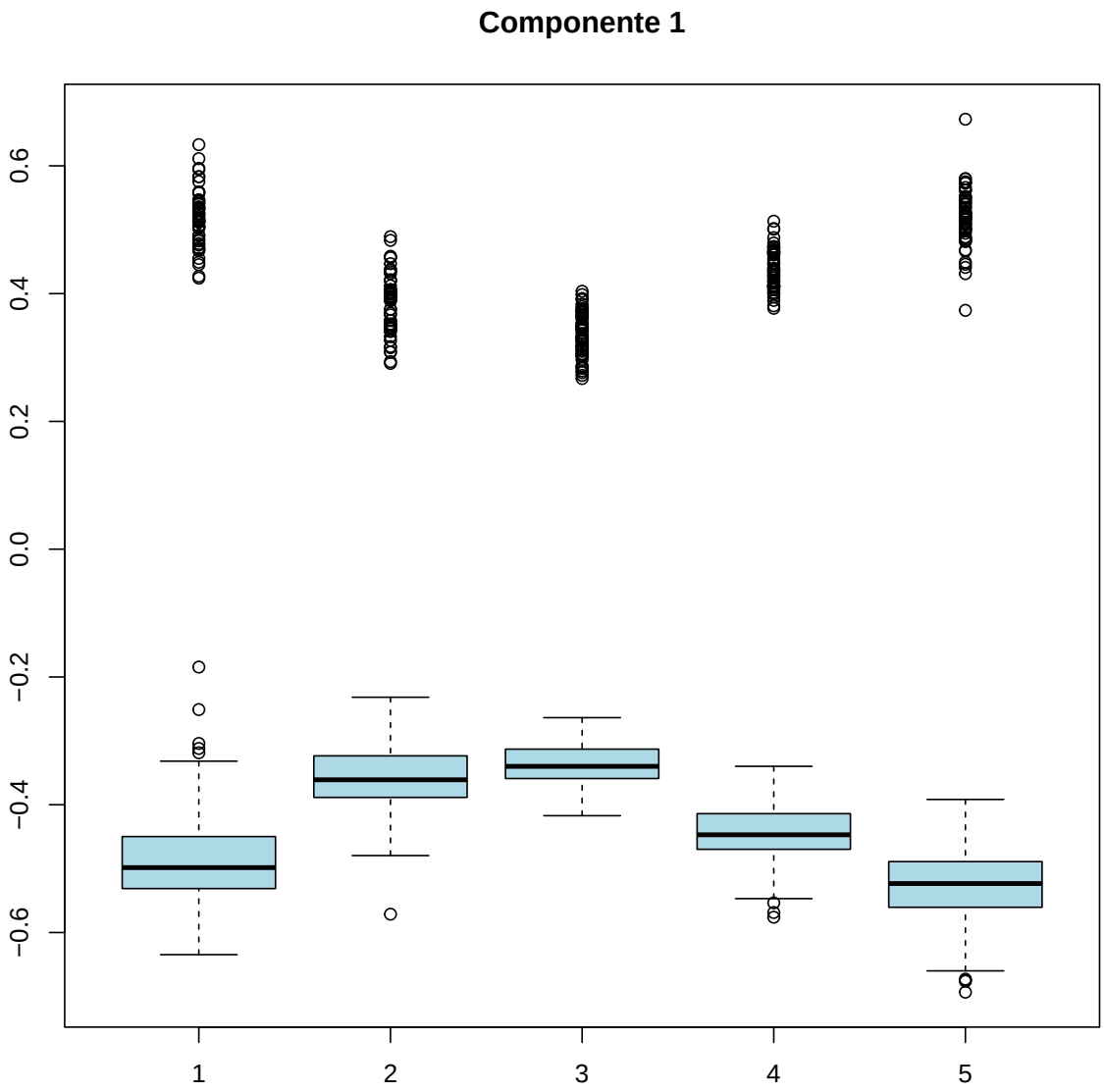
```
ind = 1:dim(X)[[1]]
eigendist = replicate(500, eigenVec(X[sample(ind, replace = TRUE), ]))

apply(eigendist[1:5, 1, ], 1, sd)
```

```
[1] 0.3211814 0.2371940 0.2132315 0.2793434 0.3303400
```

```
apply(eigendist[1:5, 2, ], 1, sd)
```

```
[1] 0.50234297 0.21266747 0.07540386 0.23636800 0.42315190
```

```
boxplot(cbind(eigendist[1, 1, ], eigendist[2, 1, ], eigendist[3, 1, ], eigendist[4,
    1, ], eigendist[5, 1, ]), main = "Componente 1", col = "lightblue")
```

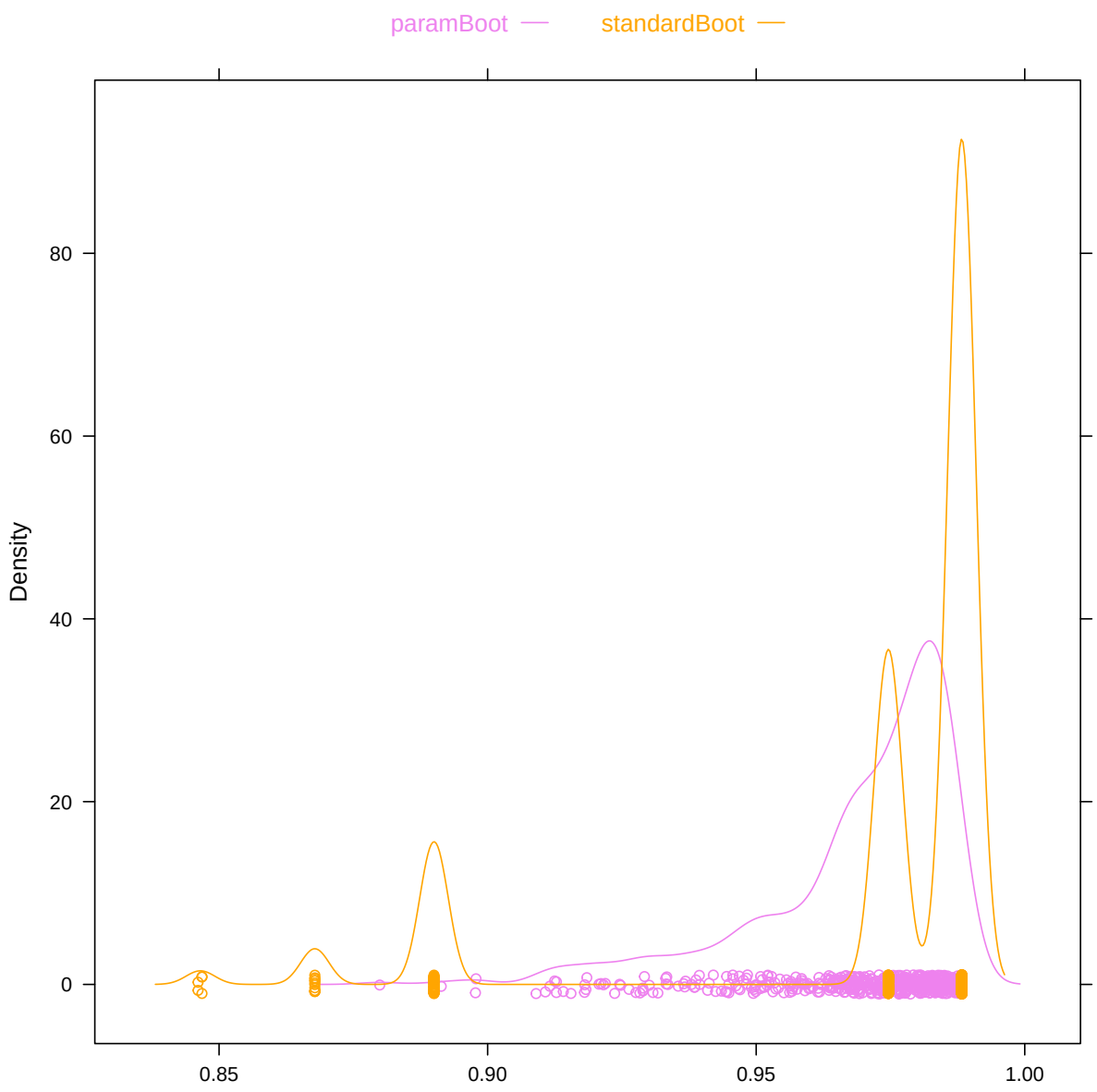


Cuando puede fallar el bootstrap

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

```
[1] 0.9882886
```

```
standardBoot = replicate(500, max(sample(X, N, replace = TRUE)))

paramBoot = replicate(500, max(runif(N, min = 0, max = thetaHat)))

library(latticeExtra)

densityplot(~paramBoot + standardBoot, xlab = "", auto.key = list(columns = 2,
    size = 2, between = 1, col = c("violet", "orange")), par.settings =
    ↪  list(superpose.line = list(col = c("violet",
    "orange"))), col = c("violet", "orange"))
```



Ejemplo de los ratones

Se toma el ejemplo de las diferencias de medias entre ratones según son tratamiento o control

```
Trata = c(94, 197, 16, 38, 99, 141, 23)
Cont = c(52, 104, 146, 10, 51, 30, 40, 27, 46)
```

```r
mean(Trata) - mean(Cont)
```

```
[1] 30.63492
```

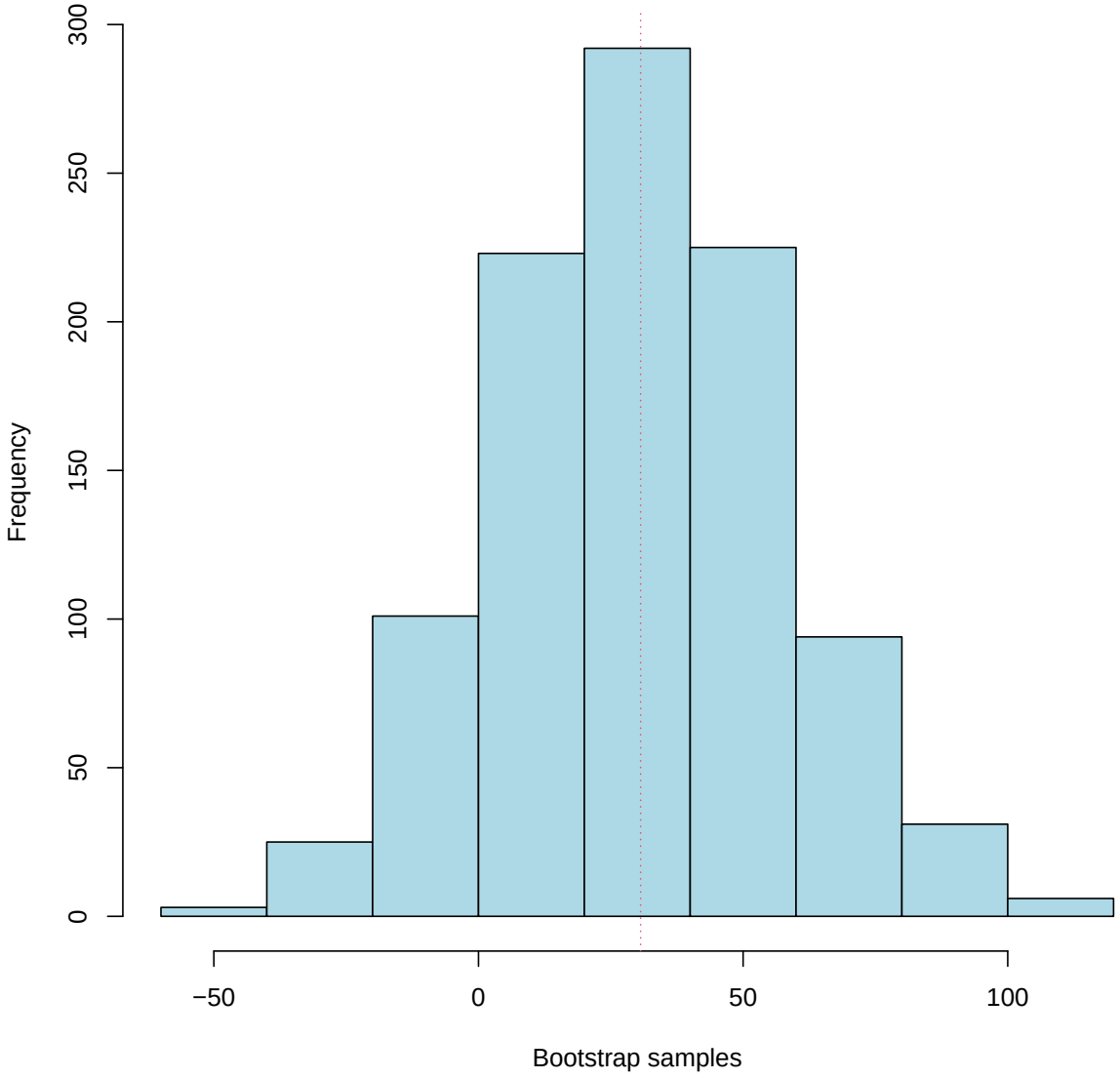
```r
B = 1000

sd(replicate(B, mean(sample(Trata, replace = TRUE)) - mean(sample(Cont, replace =
↪  TRUE))))
```

```
[1] 26.74265
```

```r
library(simpleboot)

b = two.boot(Trata, Cont, mean, R = B)

sd(b$t)
```

```
[1] 26.74637
```

```r
hist(b, col = "lightblue")
```



Programa original de la librería bootstrap

```
B = 1000
library(bootstrap)

mouse.boot.c = bootstrap(mouse.c, B, mean)
mouse.boot.t = bootstrap(mouse.t, B, mean)

mouse.boot.diff = mouse.boot.t$thetastar - mouse.boot.c$thetastar
```

```r
Trata = c(94, 197, 16, 38, 99, 141, 23)
Cont = c(52, 104, 146, 10, 51, 30, 40, 27, 46)
B = 1000
n = length(Trata)
Losratones = c(Trata, Cont)

(t.obs = mean(Trata) - mean(Cont))
```

```
[1] 30.63492
```

```r
library(boot)
t.fun = function(data, i, n) {
    bobo = data[i]
    mean(bobo[1:n]) - mean(bobo[-c(1:n)])
}

(mouse.boot = boot(Losratones, t.fun, R = 1000, n = n))
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = Losratones, statistic = t.fun, R = 1000, n = n)


Bootstrap Statistics :
    original    bias    std. error
t1* 30.63492 -29.75003    26.37164
```

Ejemplo sobre pastillas para dormir

Se toma el ejemplo de los datos correspondientes a 20 observaciones donde se mide el efecto de unas pastillas para dormir (el incremento de horas de sueño en relación a mediciones control).

Se usa primero el test de t-Student

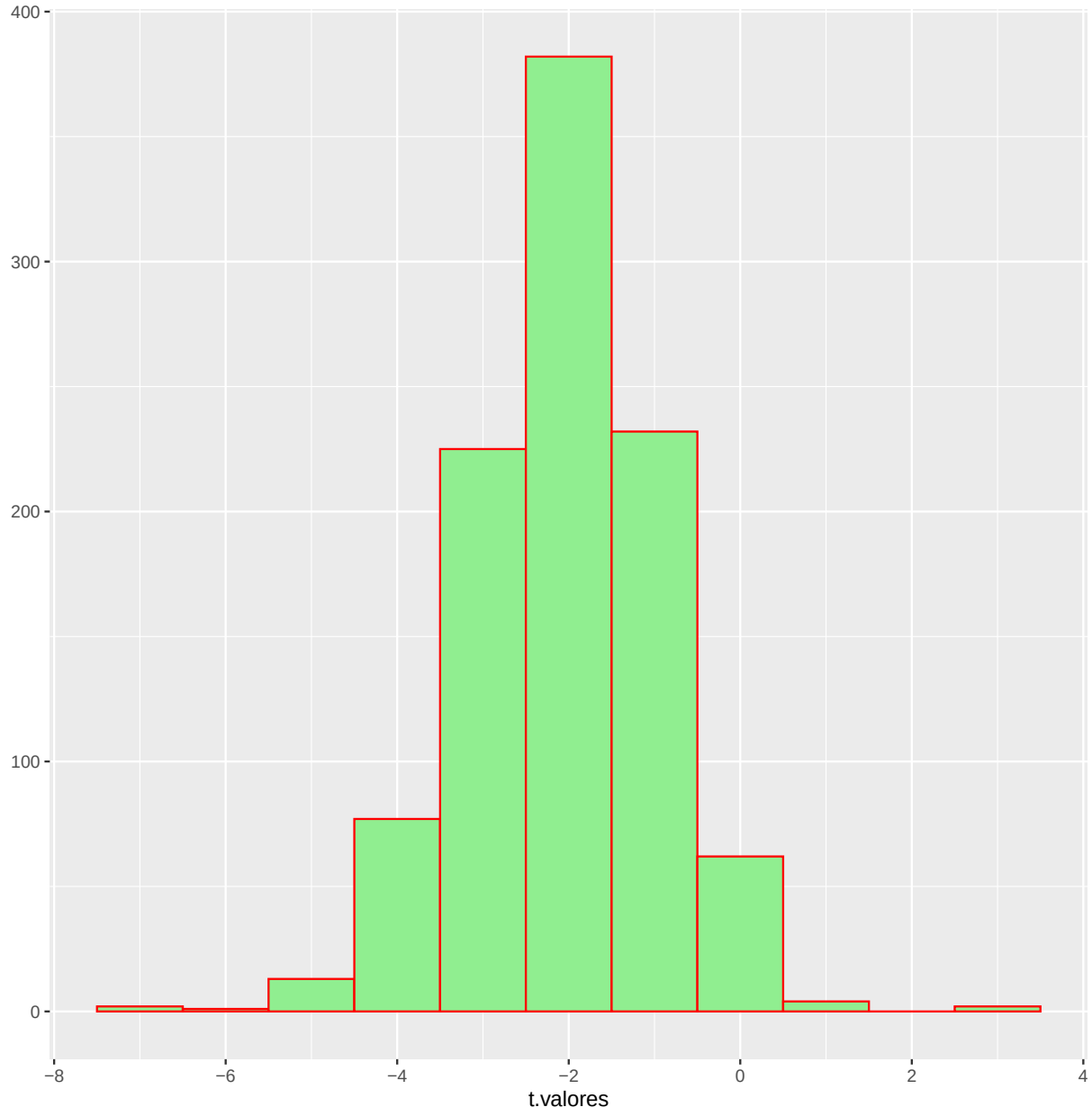
```r
data(sleep)

# test t Student
with(sleep, t.test(extra ~ group)$statistic)
```

```
        t
-1.860813
```

```
scores = sleep$extra
R = 1000
t.valores = numeric(R)

scoresG1 = subset(scores, sleep$group == 1)
scoresG2 = subset(scores, sleep$group == 2)

for (i in 1:R) {
    grupo1 = sample(scoresG1, size = 10, replace = T)
    grupo2 = sample(scoresG2, size = 10, replace = T)
    t.valores[i] = t.test(grupo1, grupo2)$statistic
}

sd(t.valores)
```

```
[1] 1.087016
```

```
ggplot2::qplot(t.valores, geom = "histogram", binwidth = 1, fill = I("lightgreen"),
    col = I("red"))
```


400
300
200
100
0
−8
−6
−4
−2
0
2
4
t.valores