Computer Intensive Methods using R

Part 2: the basic bootstrap

Prof. Dr. Ziv Shkedy

Master of Statistics
Hasselt University

General Information

Overview of the course

- The basic bootstrap algorithm
  - The bootstrap estimate of the standard error for the mean.
  - The correlation coefficient.

Overview of the course (part 1)

The Bootstrap algorithm

Introduction:
- Sampling from a population.
- The empirical distribution.
- Plug in principle.

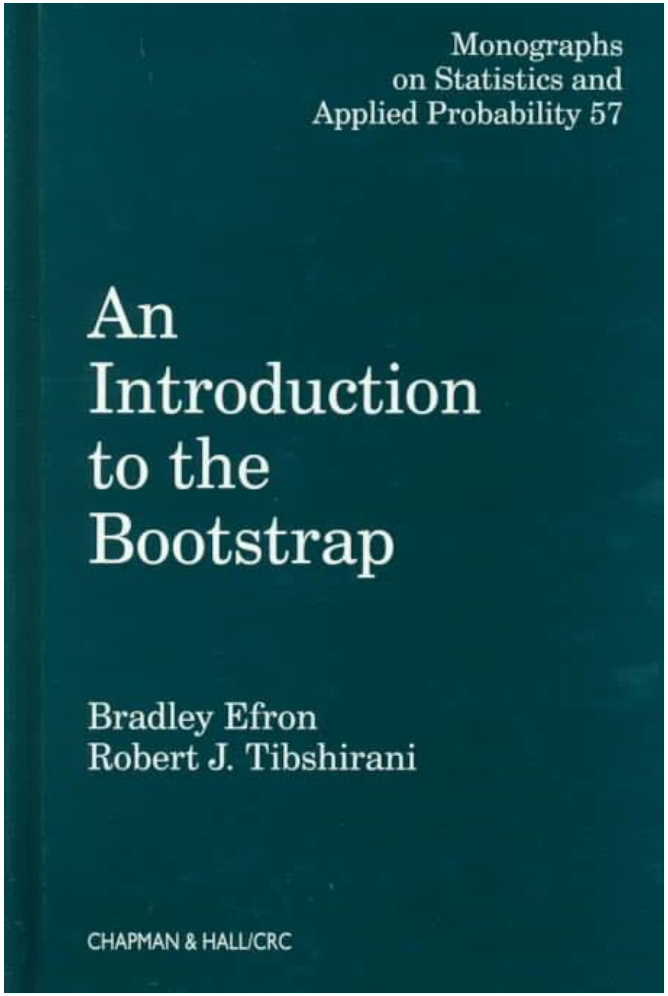
Estimation:
- Accuracy of statistics.

Inference:
- One sample tests.
- Two-samples tests.
- Bootstrap and permutation tests.

Modeling:
- Linear regression models.
- Non parametric regression.
- GLMs.

Reference

- Bradley Efron and Robert J. Tibshirani (1994): An introduction to bootstrap.
- Davison A.C. and Hinkley D.V: Bootstrap Methods and Their Application.



Course materials

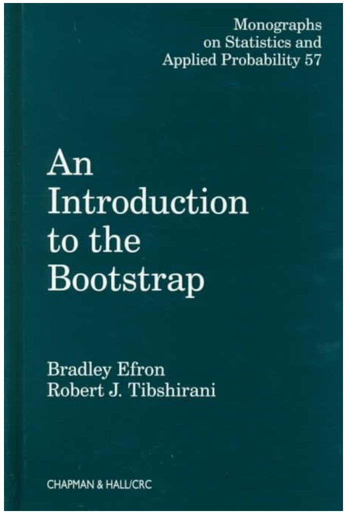
- Slides.
- R program.
- R datasets & External datasets.
- YouTube tutorials.
- Videos for the classes (highlights of each class in the course).

YouTube tutorials

- YouTube tutorials about bootstrap using R:
    1. One-sample bootstrap CI for the mean (host: LawrenceStats): https://www.youtube.com/watch?v=ZkCDYAC2iFg.
    2. Using the non-parametric bootstrap for regression models in R (host:Ian Dworkin):https://www.youtube.com/watch?v=ydtOTctg5So.
    3. Performing the Non-parametric Bootstrap for statistical inference using R (host: Ian Dworkin): https://www.youtube.com/watch?v=TP6r5CTd9yM
    4. Using the sample function in R for resampling of data - absolute basics (host: Ian Dworkin):https://www.youtube.com/watch?v=xE3KGVT6VLE
    5. Permutation tests in R - the basics (host:Ian Dworkin):https://www.youtube.com/watch?v=ZiQdzwB12Pk.
    6. Bootstrap Sample Technique in R software (host: Sarveshwar Inani):https://www.youtube.com/watch?v=tb6wb9ZdPH0
    7. Bootstrap confidence intervals for a single proportion (host: LawrenceStats):https://www.youtube.com/watch?v=ubX4QEPqx5o
    8. Bootstrapped prediction intervals (host:James Scott):https://www.youtube.com/watch?v=c3gD_PwsCGM.

- https://www.youtube.com/watch?v=gcPIyeqymOU

The bootstrap estimate of the standard error



Chapter 6

Topics

- Bootstrap:
  - Parametric
  - Non parametric
- Examples:
  - Standard error of the mean.
  - Correlation: distribution and standard error.
  - Quantile estimation and standard error.

Example 1:
the bootstrap estimate of the standard error for the mean

The observed data

A sample of 10 observations:

```
> x <- c(11.201, 10.035, 11.118, 9.055, 9.434, 9.663, 10.403, 11.662, 9.285,8.84)
> mean(x)
[1] 10.0696
```

$$x = \left(x_1, x_2, \ldots, x_{10}\right)$$

We wish to estimate the standard error of the sample mean

$$S.E(\bar{x}) = \frac{\sigma_F}{\sqrt{n}}$$

The observed data

An estimate of the standard error of the sample mean

$$\frac{\hat{\sigma}_F}{\sqrt{n}}$$

```
> x <- c(11.201, 10.035, 11.118, 9.055, 9.434, 9.663, 10.403, 11.662,
9.285,8.84)
> var(x)
[1] 0.9726152
> var(x)/10
[1] 0.09726152
```

Parametric and nonparametric bootstrap

$$F \rightarrow (x_1, x_2, \ldots, x_n) \Rightarrow \hat{\theta}$$

nonparametric bootstrap

$$\hat{F} \rightarrow (x_1, x_2, \ldots, x_n)$$

We resample from the empirical distribution

parametric bootstrap

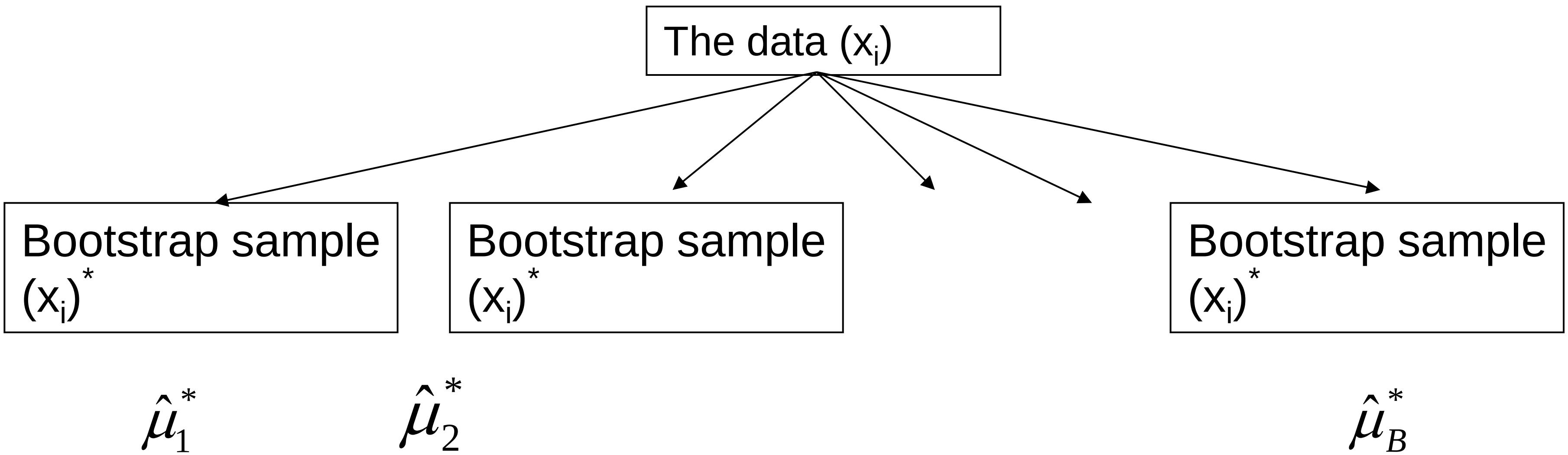
We assume a parametric model for F

$$F(\theta)$$

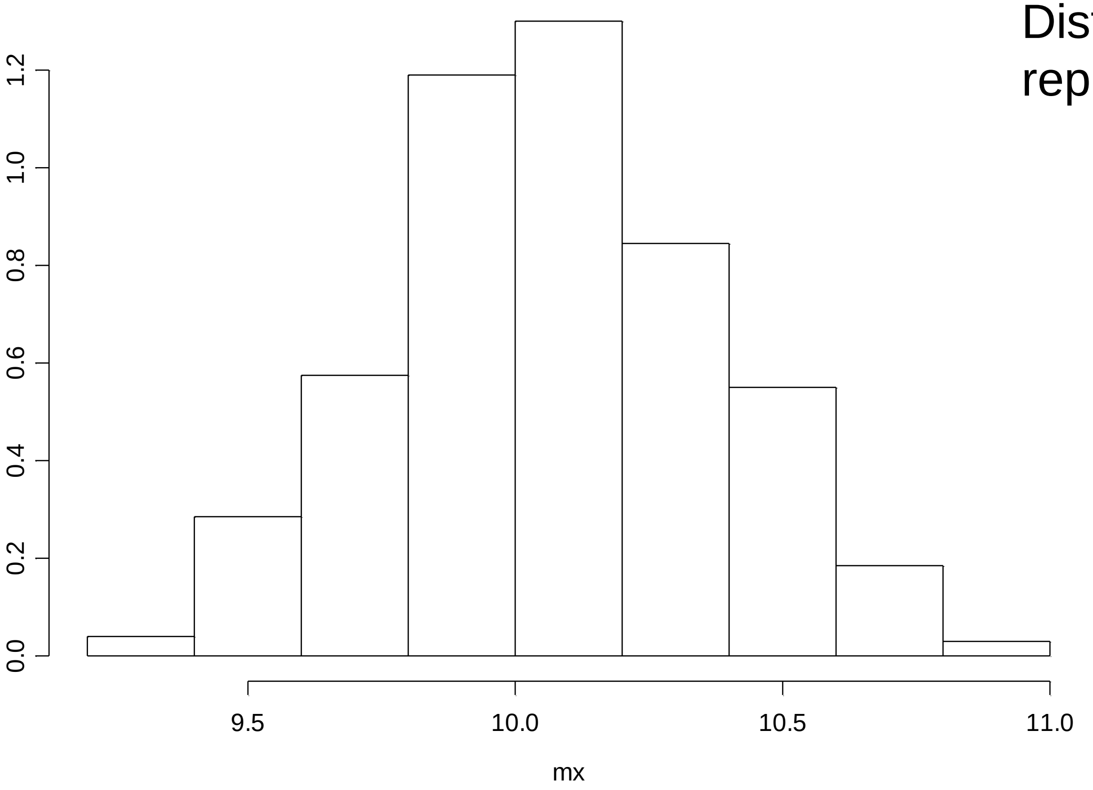
We resample from

$$F(\hat{\theta})$$

Nonparametric bootstrap



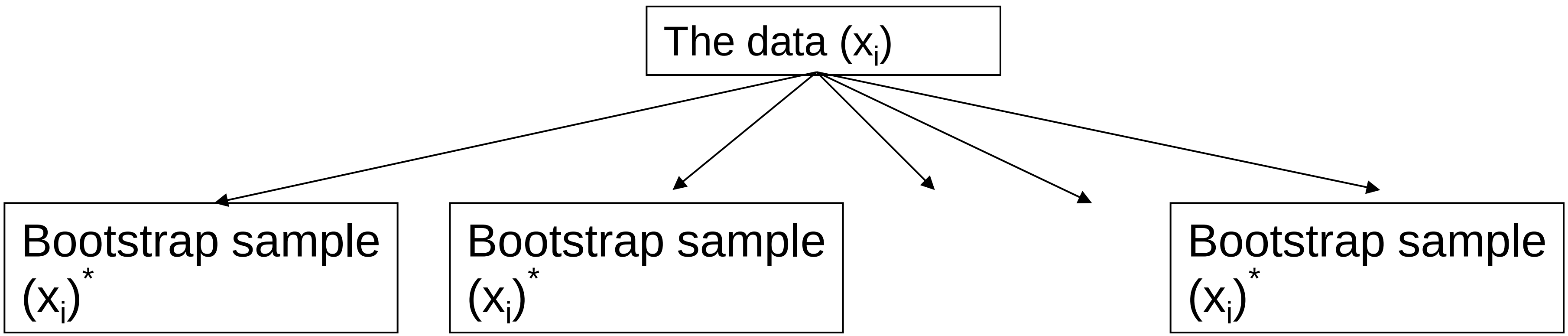
Nonparametric bootstrap



Distribution of the bootstrap replicates

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

Nonparametric bootstrap



$\hat{\mu}_1^*$ $\hat{\mu}_2^*$ $\hat{\mu}_B^*$

$$S.E.(\hat{\mu}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\mu}_b^* - \hat{\mu}^* \right)^2 \right\}^{0.5}$$

Bootstrap estimate for the standard error of the sample mean

The bootstrap algorithm (non parametric)

1) Draw *B bootstrap samples*

$$\boxed{x_1^*, x_2^*, \ldots\ldots, x_B^*}$$

with replacement from the observed data $x_1, x_2, \ldots\ldots, x_n$

2) Evaluate the bootstrap replications

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

3) Estimate se $F(\hat{\theta})$ or better approximate se $F(\hat{\theta}^*)$ by the sample deviation of the B replications

$$\boxed{S.E.(\hat{\theta}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}^* \right)^2 \right\}^{0.5}}$$

R code: non parametric bootstrap

```
> var(mx)
[1] 0.09357364
```

The estimated standard error

$$S.E(\hat{\mu}) = \sqrt{0.0935..}$$

```
n<-length(x)
B<-1000
mx<-c(1:B)
for(i in 1:B){
cat(i)
boot.i<-sample(x,n,replace=T)
mx[i]<-mean(boot.i)
}
```

$$\hat{F} \rightarrow \left(x_1^*, x_2^*, \ldots, x_n^*\right) \Rightarrow \hat{\theta}_b^*$$

Parametric bootstrap

We assume a parametric model for F

$$F = N(\mu, \sigma^2)$$

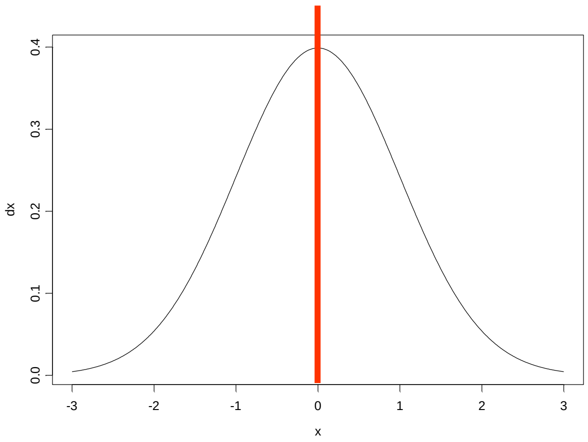
We estimate F by

$$\hat{F} = N(\hat{\mu}, \hat{\sigma}^2)$$

We replace the unknown parameters in F with their plug-in estimates

Parametric bootstrap

Population



$$\theta = (\mu, \sigma)$$

$$F = N(\mu, \sigma^2)$$

Sample

$$F \rightarrow \left(x_1, x_2, \ldots, x_n\right)$$

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma})$$

$$\hat{F} = N(\hat{\mu}, \hat{\sigma}^2)$$

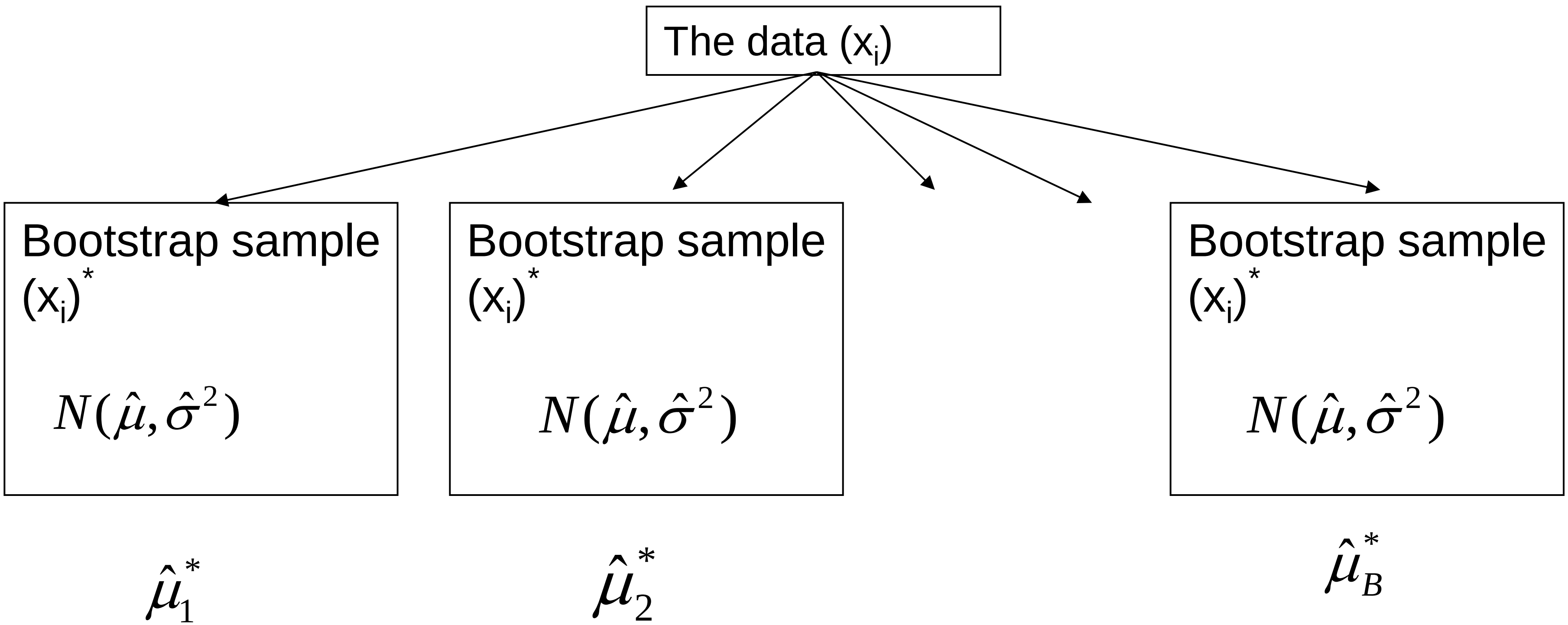
Parametric bootstrap

$$\hat{F} = N(\hat{\mu}, \hat{\sigma}^2)$$

$$x_1^* \quad x_2^* \quad x_B^*$$

B bootstrap samples from the empirical distribution

Parametric bootstrap: standard error of the mean



$$S.E.(\hat{\mu}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\mu}_b^* - \hat{\mu}^* \right)^2 \right\}^{0.5}$$

The bootstrap algorithm (parametric)

1) Draw *B bootstrap samples of size n*

$$\boxed{x_1^*, x_2^*, \ldots\ldots, x_B^*}$$

from the distribution $F(\hat{\theta})$

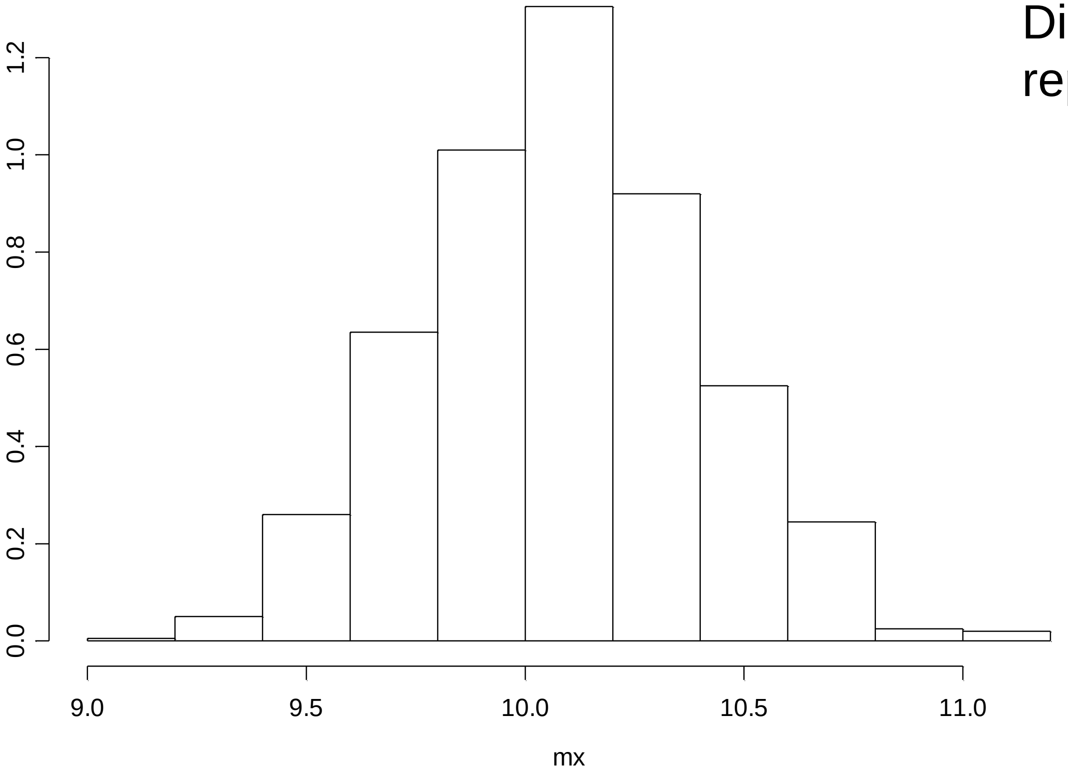
2) Evaluate the bootstrap replications

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

3) Estimate se $F(\hat{\theta})$ *or better approximate se* $F(\hat{\theta}^*)$ *by* the sample deviation of the *B replications*

$$\boxed{S.E.(\hat{\theta}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}^* \right)^2 \right\}^{0.5}}$$

Parametric bootstrap



Distribution of the bootstrap replicates

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

R code: parametric bootstrap

> var(mx)
[1] 0.1007613

Bootstrap estimate for the standard error for the mean

```
B<-1000
MLx<-mean(x)
Varx<-var(x)
mx<-c(1:B)
for(i in 1:B){
cat(i)
boot.i<-rnorm(n,MLx,sqrt(Varx))
mx[i]<-mean(boot.i)
}
```

$\hat{\theta}_b^*$

$$F = N(\hat{\mu}, \hat{\sigma}^2) = N(\bar{x}, s^2)$$

Example 2:
the correlation coefficient

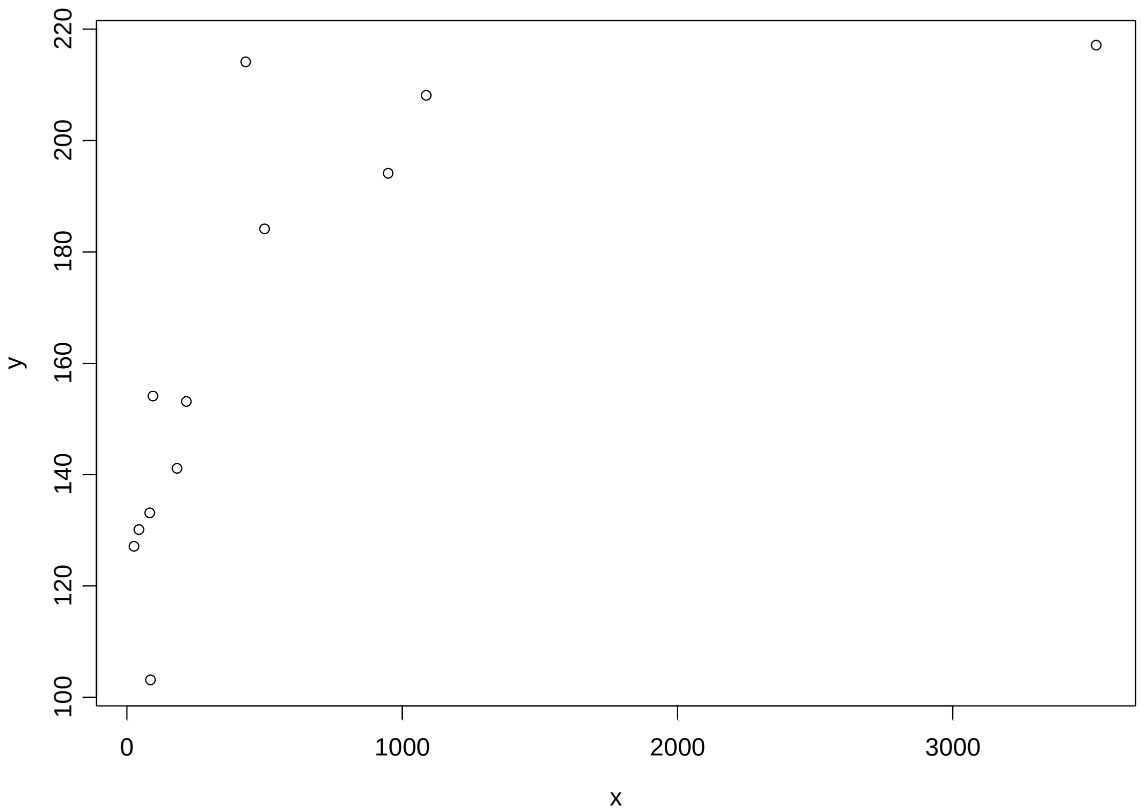
The sample

```
        x   y
 [1,]   29 127
 [2,]  435 214
 [3,]   86 133
 [4,] 1090 208
 [5,]  219 153
 [6,]  503 184
 [7,]   47 130
 [8,] 3524 217
 [9,]  185 141
[10,]   98 154
[11,]  952 194
[12,]   89 103
```

Observed correlation

```
> cor(x, y)
[1] 0.6738982
```

The sample




```
> cor(x, y)
[1] 0.6738982
```

The bootstrap algorithm: non parametric bootstrap

The observed sample

$$
\begin{array}{cc}
x_1 & y_1 \\
x_2 & y_2 \\
. & . \\
. & . \\
. & . \\
x_{12} & y_{12}
\end{array}
$$

We resample the **pair** (x_i, y_i) with replacement

n=12

$$
\begin{array}{cc}
x^*_1 & y^*_1 \\
x^*_2 & y^*_2 \\
. & . \\
. & . \\
. & . \\
x^*_{12} & y^*_{12}
\end{array}
$$

The bootstrap sample

The bootstrap algorithm

The bootstrap sample

$$
\begin{array}{cc}
x^*_1 & y^*_1 \\
x^*_2 & y^*_2 \\
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot \\
x^*_{12} & y^*_{12}
\end{array}
\longrightarrow \hat{\rho}^*_b(x^*, y^*)
$$

For each bootstrap sample we calculate the correlation

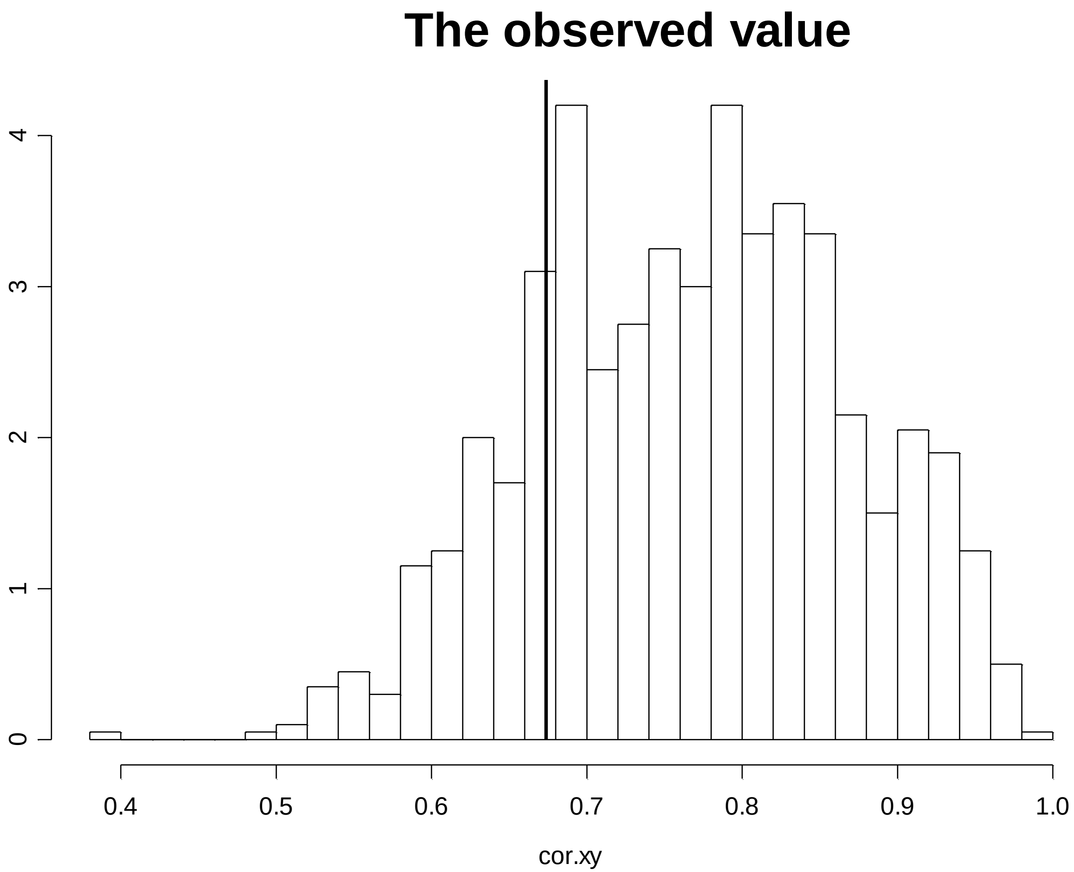
The bootstrap algorithm

B bootstrap replicates

$\hat{\rho}_1^*(x^*, y^*)$ $\hat{\rho}_2^*(x^*, y^*)$ $\hat{\rho}_B^*(x^*, y^*)$

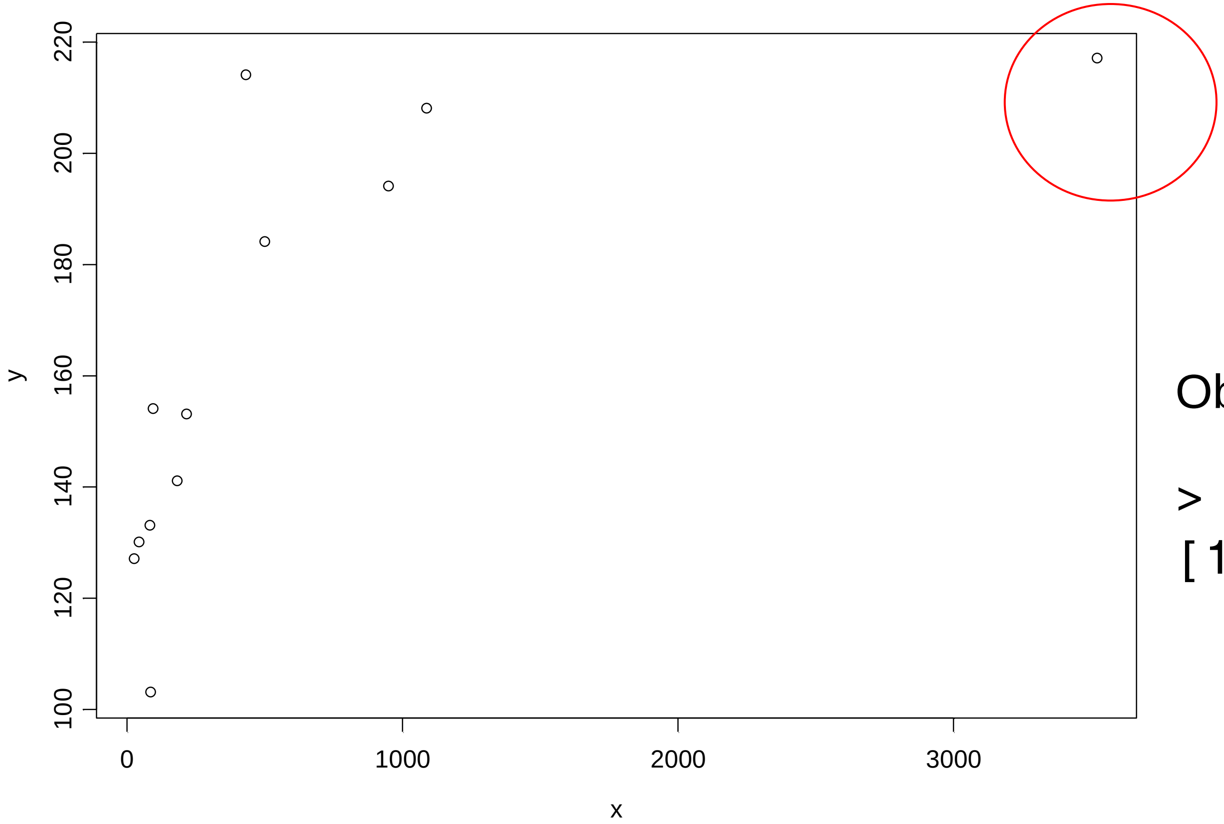
$$S.E(\hat{\rho}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\rho}_b^* - \hat{\rho}^* \right)^2 \right\}^{\frac{1}{2}}$$

1000 bootstrap replicates for the correlation



The observed sample

What is the influence of observation 8 on the estimate for the correlation ?



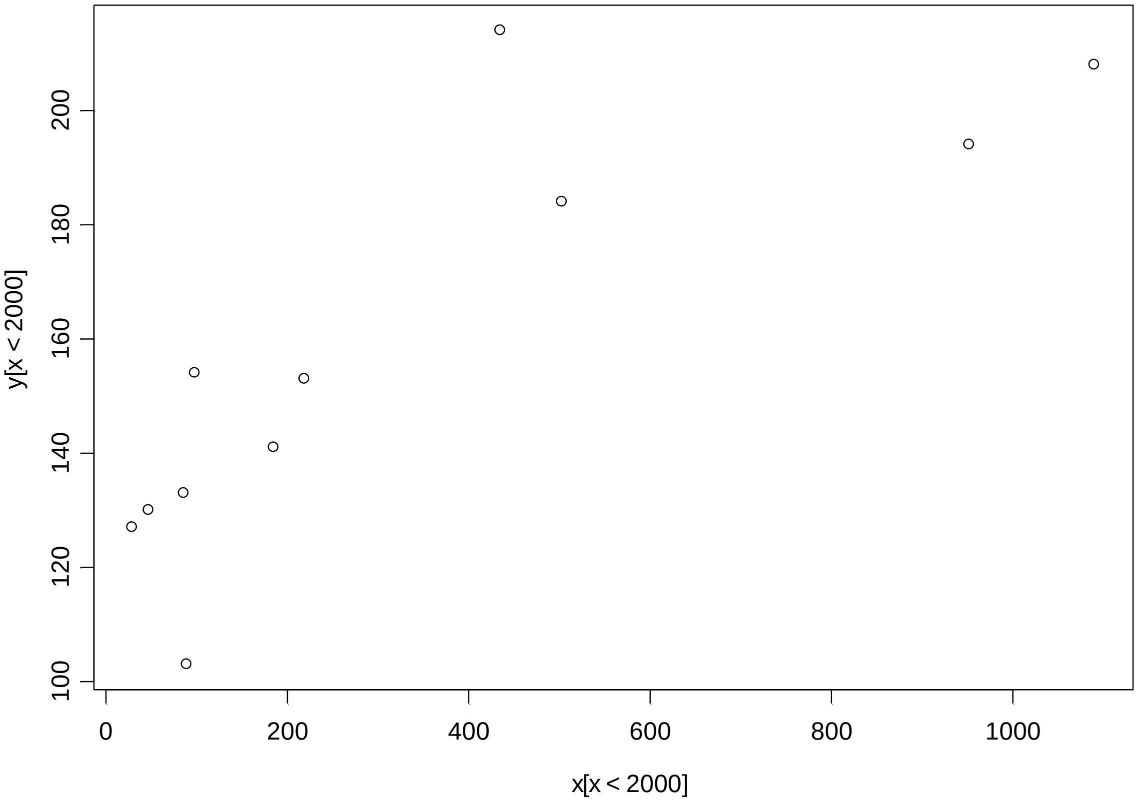
Observation 8

(3524,217)

Observed correlation

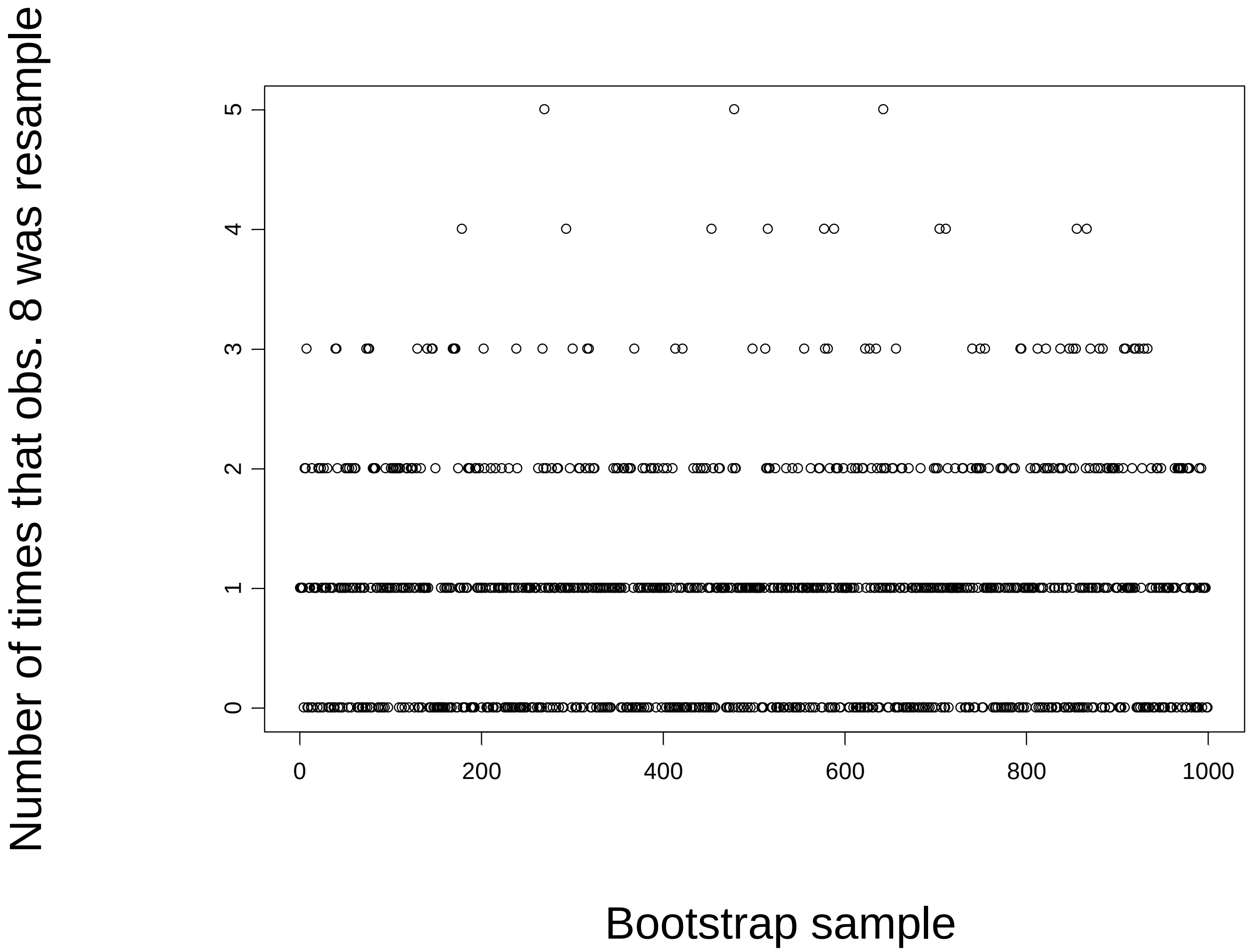
```
> cor(x, y)
[1] 0.6738982
```

Data without observation 8

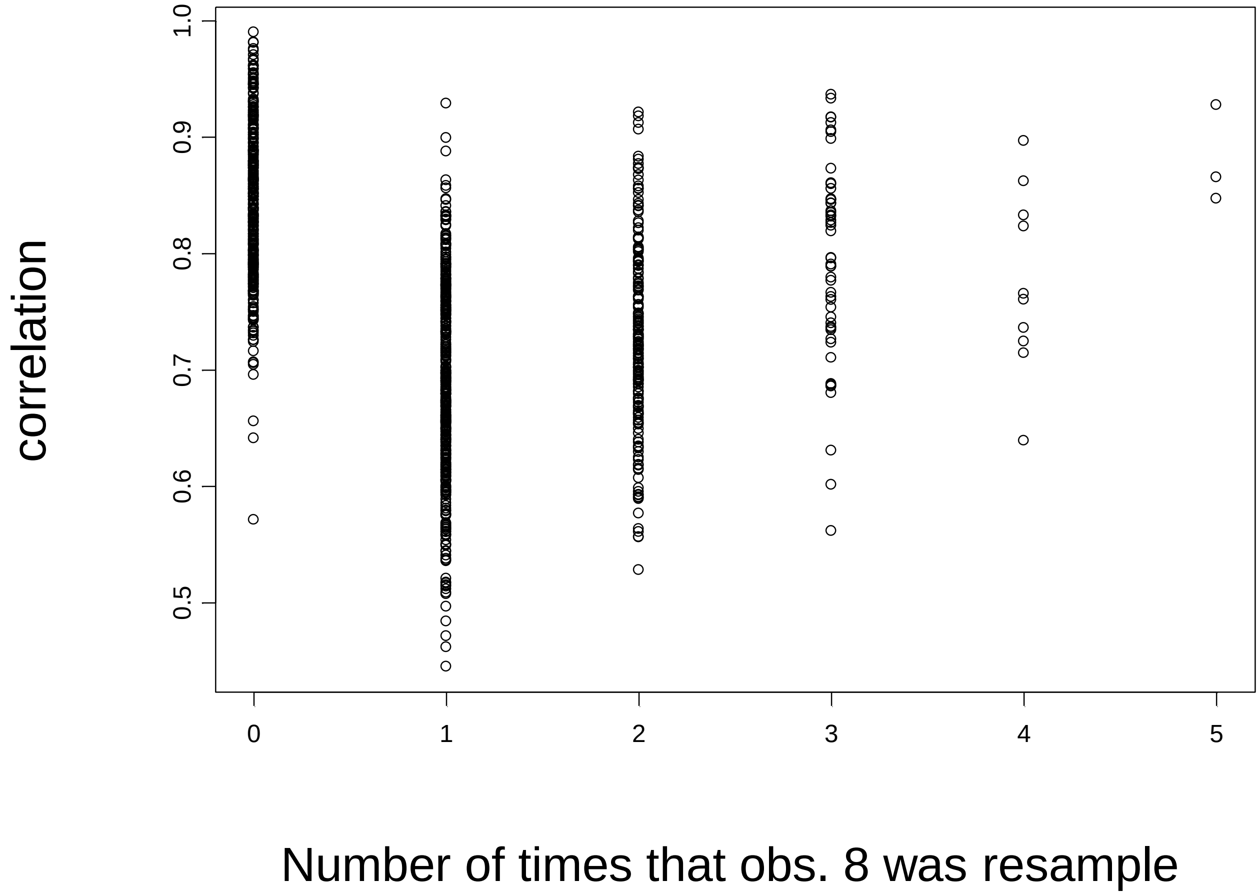



```
> cor(x[x < 2000], y[x < 2000])
[1] 0.820564
```

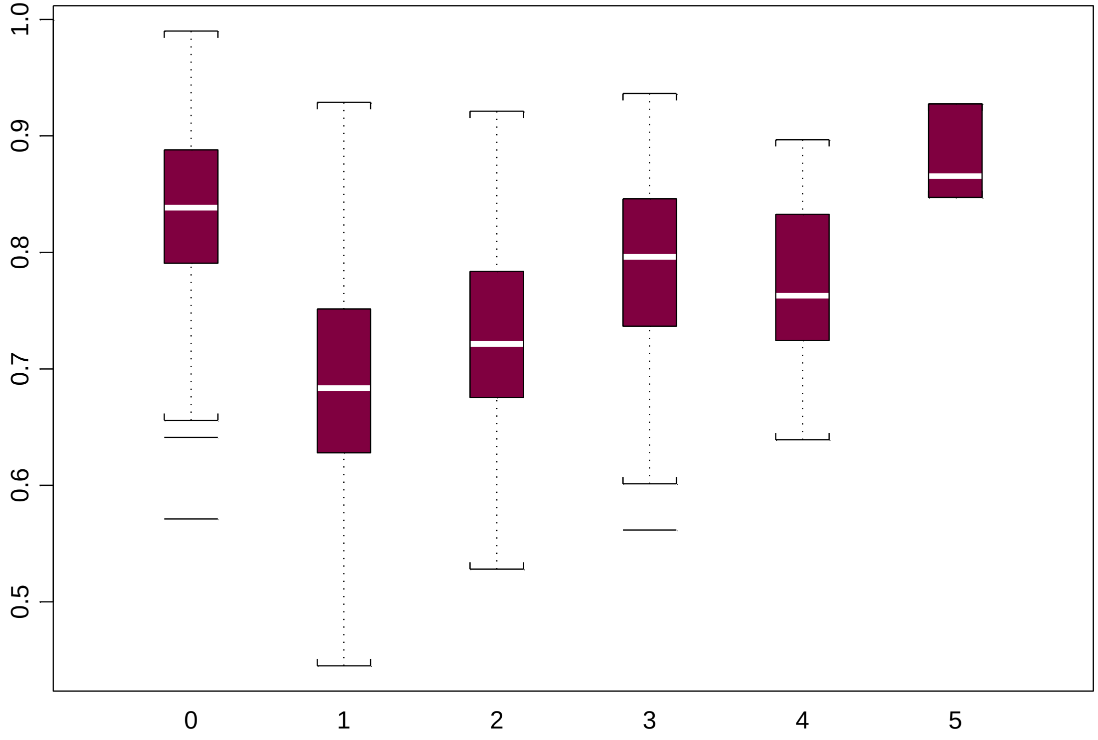
How many time observation 8 was resample in each bootstrap sample ?



The influence of observation 8



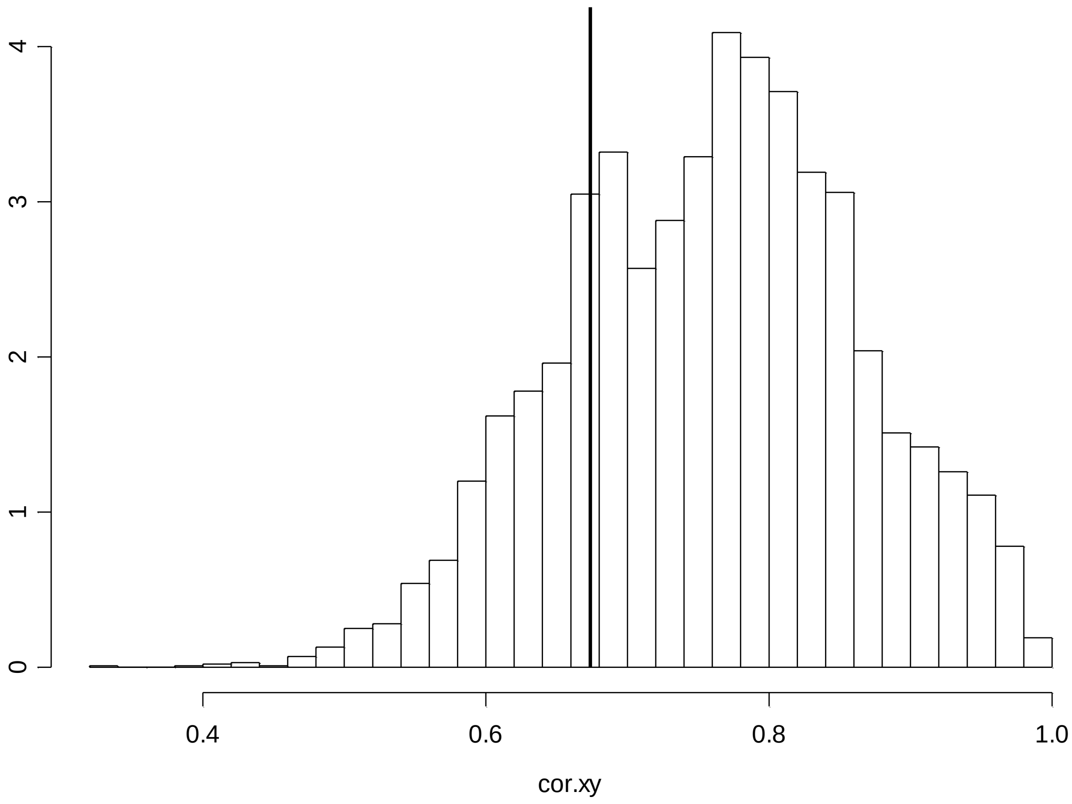
The influence of observation 8



Number of times that obs. 8 was resample

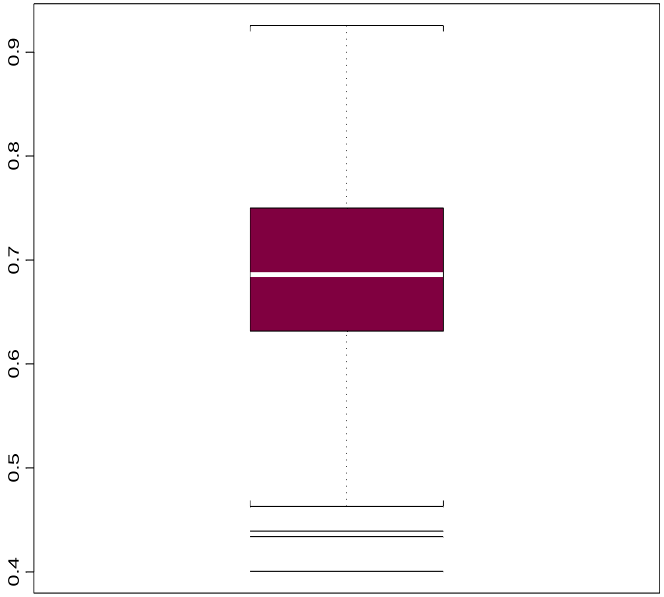
Why the correlation increases when the number of times that observation 8 included increases ?

B=5000



Number of bootstrap samples increases to 5000.

The influence of observation 8



Booxplot for the bootstrap replicates (of the correlation) for the samples in which observation 8 was sampled only once

```
> mean(cor.xy[obs.8 == 1])
[1] 0.6881557

> cor(x, y)
[1] 0.6737664
```

R code

```
x<-c(29,435,86,1090,219,503,47,3524,185,98,952,89)
y<-c(127,214,133,208,153,184,130,217,141,154,194,103)

cor.obs<-cor(x,y)
n<-length(x)
index<-c(1:n)
B<-1000
obs.8<-cor.xy<-c(1:B)

for(i in 1:B)
{

boot.i<-sample(index,n,replace=T)


x.b<-x[boot.i]
y.b<-y[boot.i]

cor.xy[i]<-cor(x.b,y.b)
}
```

We resample for the vector (1,2,3,4,5.,,,,n)

Bootstrapping pairs

Bootstrap replicates for the correlation.

The bootstrap algorithm: parametric bootstrap

The observed sample

$$\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{12} & y_{12} \end{array}$$

We resample the **pairs** (x_i, y_i) with replacement (n=12) BUT

What is the empirical distribution ?

$$F_{xy} = H(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho_{(x,y)})$$

The probability distribution function

Example 3
The air quality data

New York Air Quality Measurements

- Daily air quality measurements in New York, May to September 1973.
- Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island.
- In R:

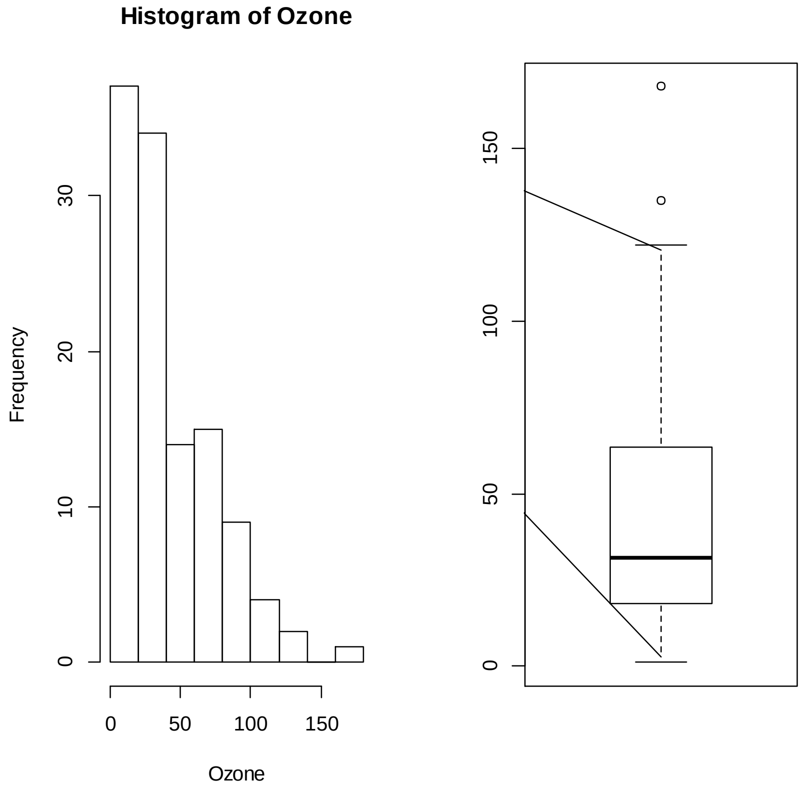
```
> help(airquality)
```

The Ozone levels

- n=116.

```
> length(Ozone)
[1] 116
> hist(Ozone)
> boxplot(Ozone)
```

```
> quantile(Ozone,probs=c(0.25,0.5,0.75))
  25%   50%   75%
18.00 31.50 63.25
```



The aim of the analysis:
Estimate the standard error of the quantiles

The empirical distribution

Population:
ozone level

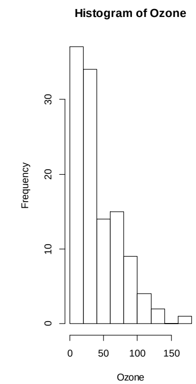
$$(x_1, x_2, \ldots, x_n)$$

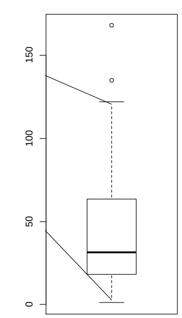
Sample: May to
September 1973

$$F$$

The probability
distribution

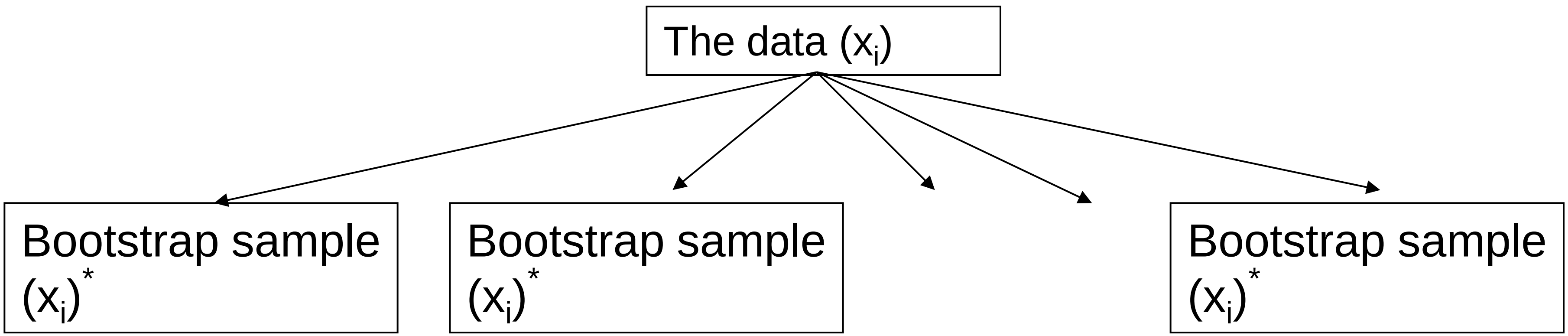
$$\hat{F}$$





The empirical probability
distribution

Nonparametric bootstrap



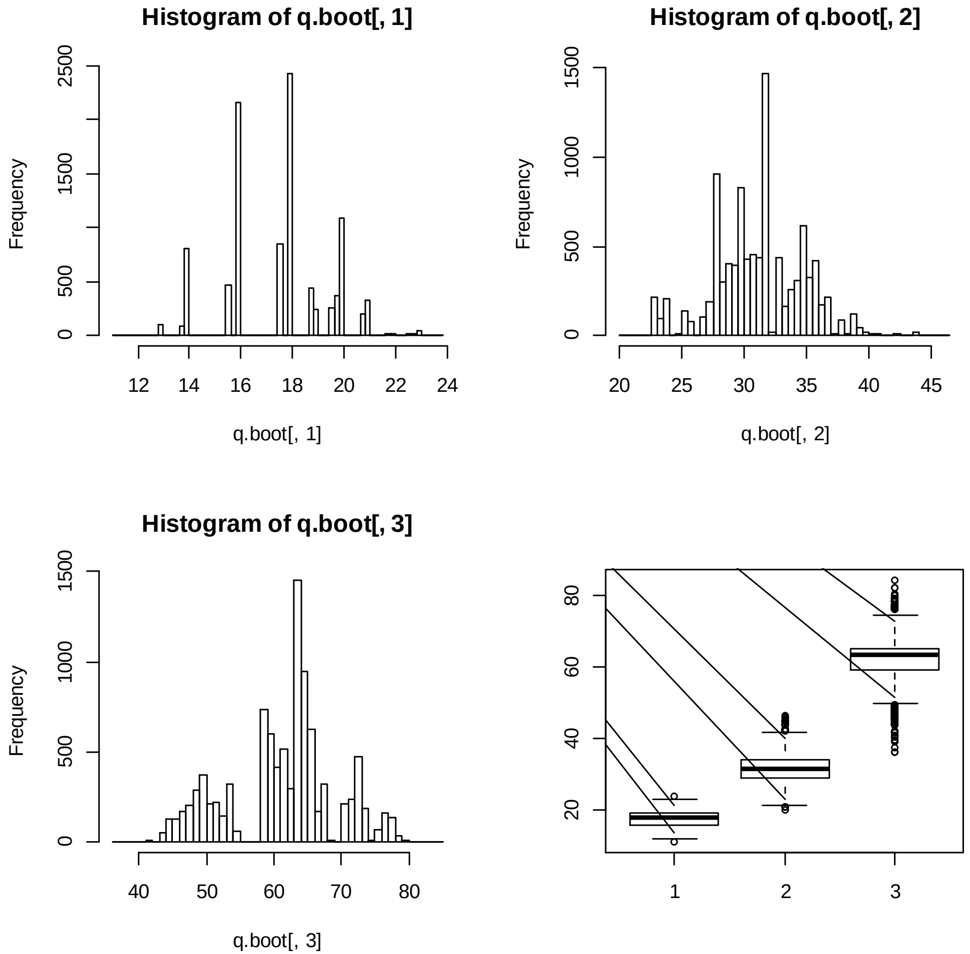
$\hat{q}^*_{0.25}, \hat{q}^*_{0.5}, \hat{q}^*_{0.75}$ $\hat{q}^*_{0.25}, \hat{q}^*_{0.5}, \hat{q}^*_{0.75}$ $\hat{q}^*_{0.25}, \hat{q}^*_{0.5}, \hat{q}^*_{0.75}$

B bootstrap sample

Distribution of the bootstrap replicates for q25, q50 and q75

- B=10000.
- Observed quantiles:

```
25%   50%   75%
18.00 31.50 63.25
```



Standard error for the quartiles

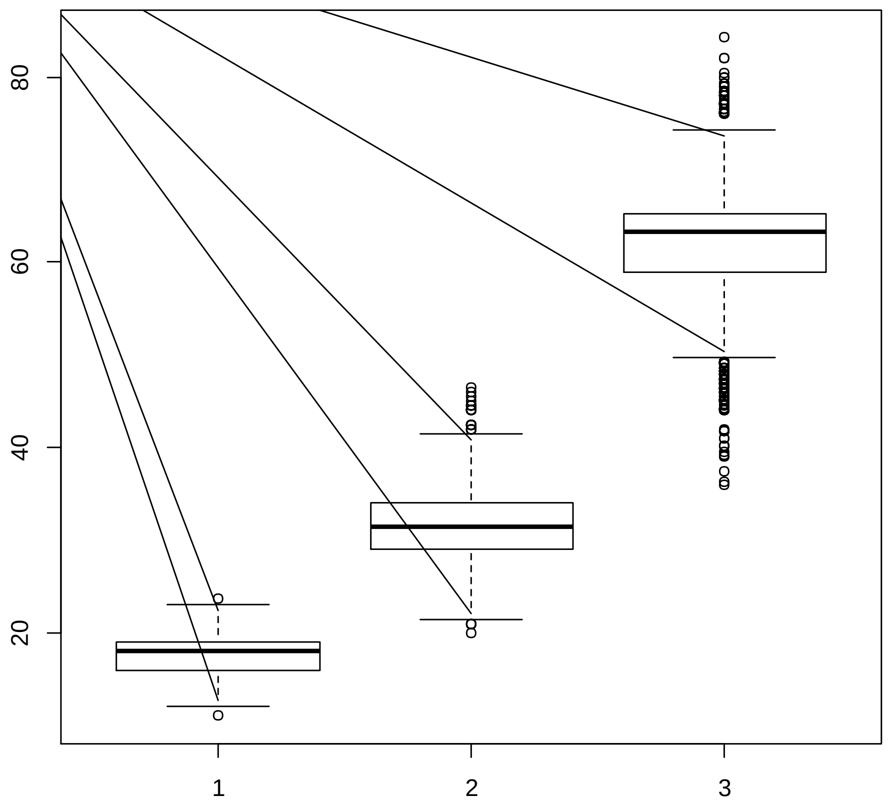
B bootstrap replicates (per quantile)

$$\hat{q}^*_{\ell,1}, \hat{q}^*_{\ell,2}, \ldots, \hat{q}^*_{\ell,B}$$

$$S.E(\hat{q}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(q^*_{\ell,b} - \bar{q}^* \right)^2 \right\}^{\frac{1}{2}}$$

Estimation of the standard error of q25, q50 and q75

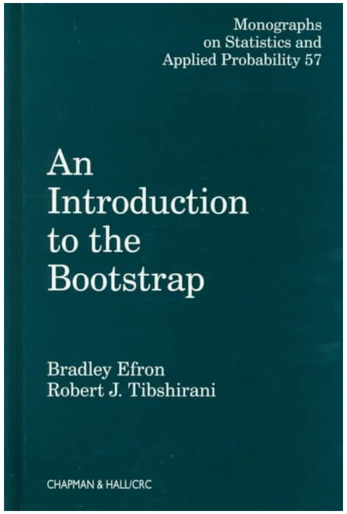
```
25%   50%   75%
18.00 31.50 63.25

> boxplot(q.boot[,1],q.boot[,2],q.boot[,3])
> var(q.boot[,1])
[1] 4.102476
> var(q.boot[,2])
[1] 13.22946
> var(q.boot[,3])
[1] 64.60844
```



R code

```
> B<-10000
> q.boot<-matrix(0,B,3)
> for(b in 1:B)
+ {
+ Ozone.boot<-sample(Ozone,size=116,replace=TRUE)
+ q.boot[b,]<-quantile(Ozone.boot,probs=c(0.25,0.5,0.75))
+ }
>
> par(mfrow=c(2,2))
> hist(q.boot[,1],nclass=50)
> hist(q.boot[,2],nclass=50)
> hist(q.boot[,3],nclass=50)
> boxplot(q.boot[,1],q.boot[,2],q.boot[,3])
```

Bootstrap standard error: some examples



Topics

- Examples:
  - The score data:
    - Distribution of the covariance matrix.
    - Ratio between variables.
  - The fuel data:
    - Non parametric regression: a loess model for the fuel data.

Example 1: the score data

The score data

- 88 students who took examinations in 5 subjects.
- Some where with open book and other with closed book.
- In R:

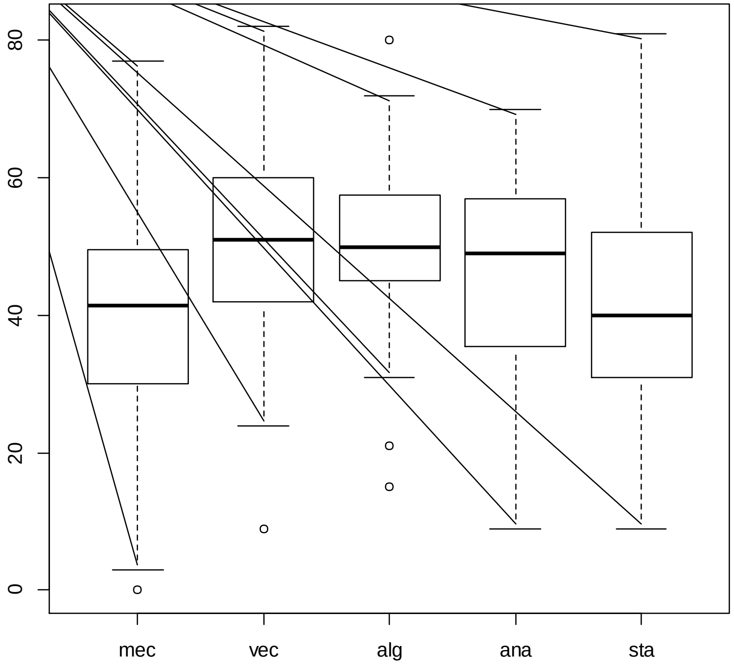
```
> help(scor)
```

- Variables in the data:

```
> head(scor)
  mec vec alg ana sta
1  77  82  67  67  81
2  63  78  80  70  81
3  75  73  71  66  81
4  55  72  63  70  68
5  63  63  65  70  63
6  53  61  72  64  73
```

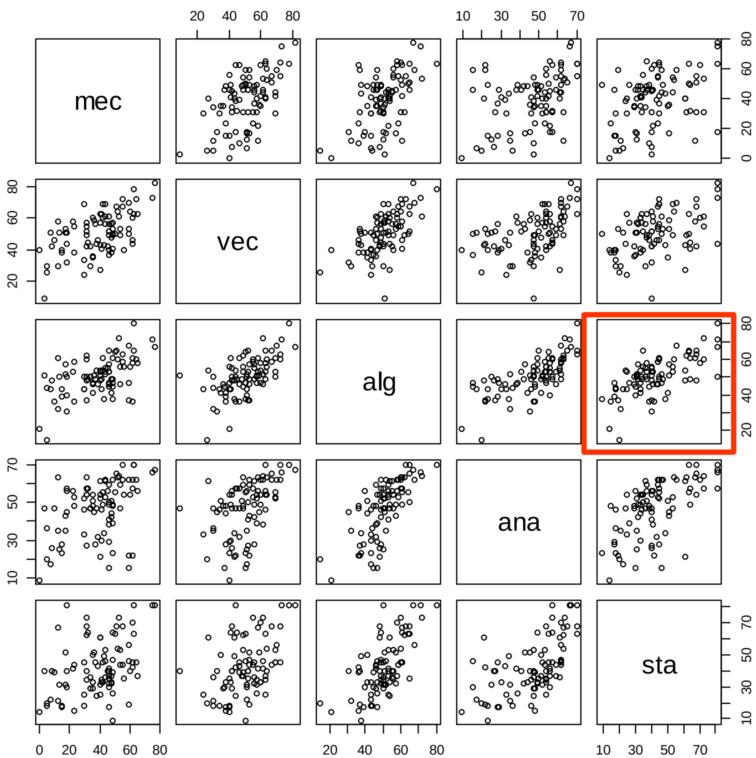
- Mec: mechanics, closed book note
- Vec: vectors, closed book note
- Alg: algebra, open book note
- Ana: analysis, open book note
- Sta: statistics, open book note

The score data



Boxplot for the score data.

The score data




```
> pairs(scor)
> 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
```

Main focus: variance/covariance matrix.

What is the standard error of the covariance between algebra and statistics ?

The score data

- The joint distribution of the scores:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \sim H(\Sigma)$$

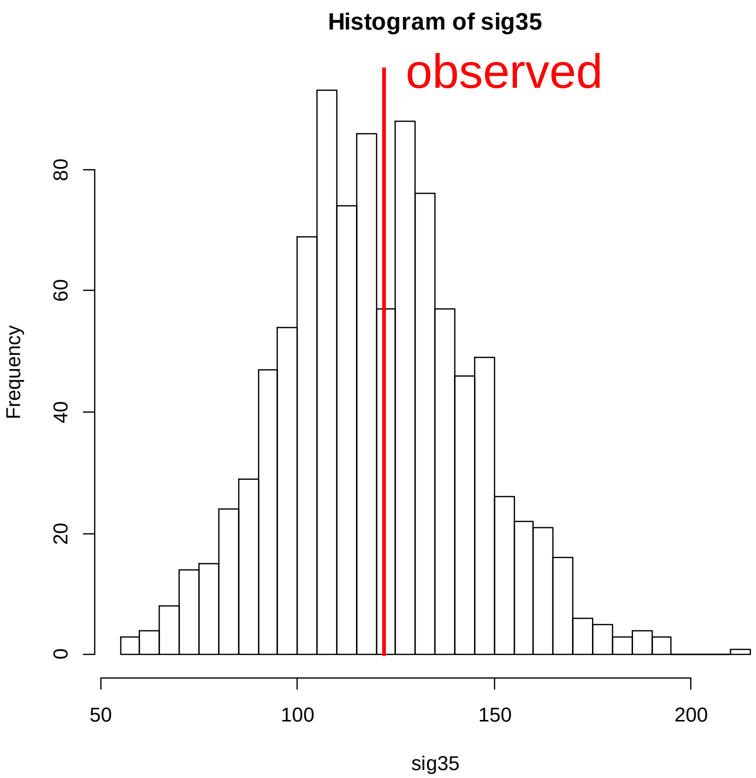
$$\Sigma_{3,5} = 121.87056$$

- Main interest: variability and distribution of the element of the covariance matrix.

- For example:

$$\Sigma_{3,5} \sim ??$$
$$\mathrm{var}(\Sigma_{3,5})$$

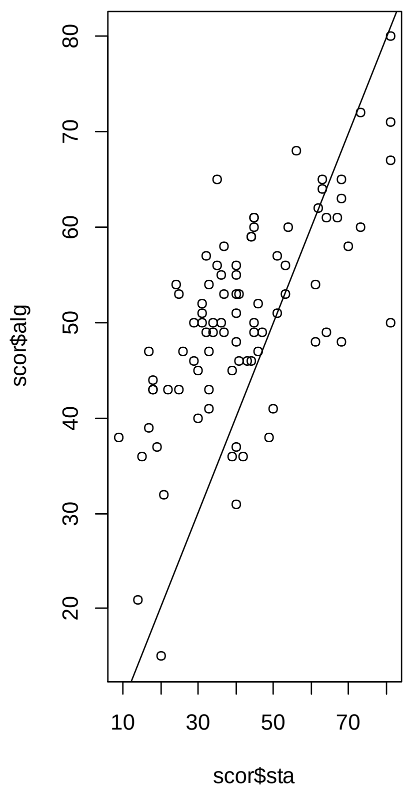
Non parametric bootstrap

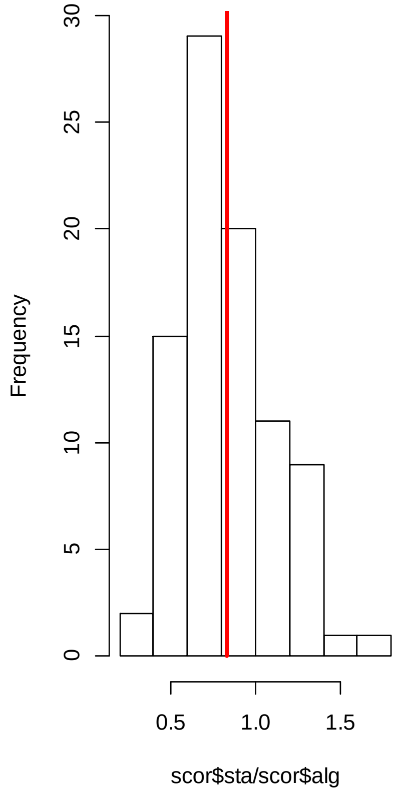



```
> cov(scor)
         mec       vec       alg       ana       sta
mec 305.7680 127.22257 101.57941 106.27273 117.40491
vec 127.2226 172.84222  85.15726  94.67294  99.01202
alg 101.5794  85.15726 112.88597 112.11338 121.87056
ana 106.2727  94.67294 112.11338 220.38036 155.53553
sta 117.4049  99.01202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

```
> var(sig35)
[1] 587.4874
> sqrt(var(sig35))
[1] 24.23814
```

The ratio between statistics and algebra scores





The ratio between the scores:

$$\theta_i = r_i = \frac{x_{5i}}{x_{3i}}$$

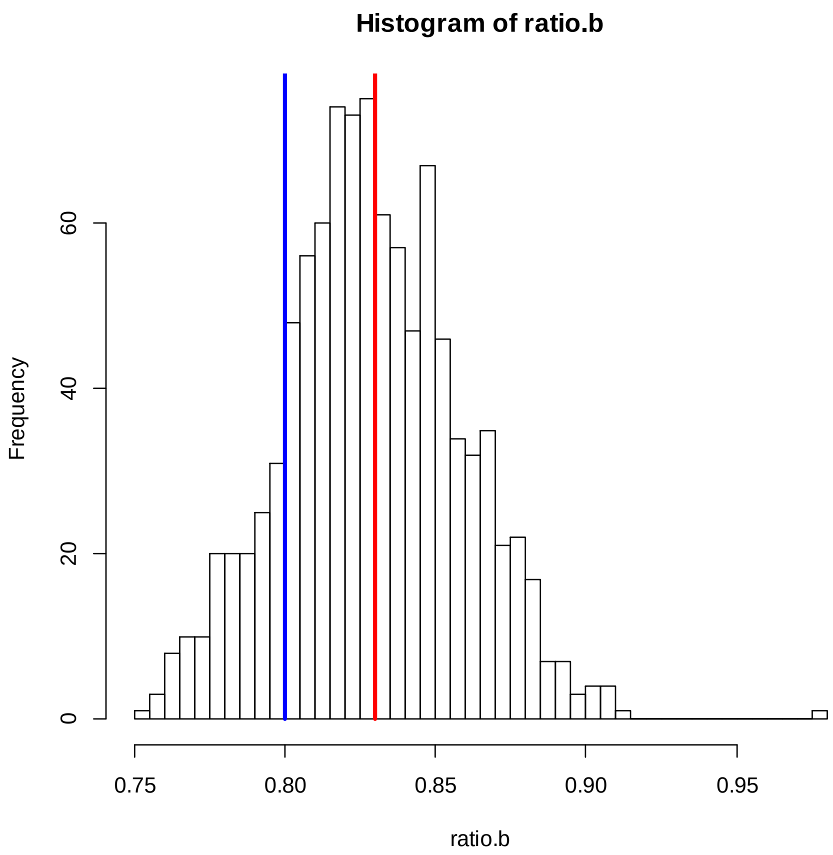
```
> m.r<-mean(scor$sta/scor$alg)
> m.r
[1] 0.8300346
```

$$\bar{r} = \hat{\theta}$$

$$\text{var}(\bar{r}) = ?$$

$$P(\bar{r} < 0.8) = ?$$

Non parametric bootstrap



- Distribution of the bootstrap replicates for the ratio.
- All values are smaller than 1, what does this mean ?

```
> var(ratio.b)
[1] 0.0008710094
> sum(ratio.b<0.8)
[1] 148
```

$$\hat{P}(\bar{r} < 0.8) = \frac{148}{1000}$$

R code

```
n<-length(scor$sta)
B<-1000
index<-c(1:n)
sig35<-ratio.b<-c(1:B)
for(i in 1:B)
{
index.b<-sample(index,n,replace=TRUE)
scor.b<-scor[index.b,]
cov.b<-cov(scor.b)
sig35[i]<-cov.b[5,3]
ratio.b[i]<-mean(scor.b$sta/scor.b$alg)
}

hist(sig35,nclass=50)
lines(c( 121.8706,121.8706),c(0,500),col=2,lwd=3)
var(sig35)
sqrt(var(sig35))


par(mfrow=c(1,2))
plot(scor$sta,scor$alg)
abline(0,1)
hist(scor$sta/scor$alg,main=" ")
m.r<-mean(scor$sta/scor$alg)
lines(c(m.r,m.r),c(0,100),col=2,lwd=3)

par(mfrow=c(1,1))
hist(ratio.b,nclass=50)
lines(c(m.r,m.r),c(0,500),col=2,lwd=3)
lines(c(0.8,0.8),c(0,500),col=4,lwd=3)


var(ratio.b)
sum(ratio.b<0.8)
```

Non parametric bootstrap

$$x^*_{i,b} = (x_1, x_2, x_3, x_4, x_5)^*_i$$

We re sample the lines (cases) in the data matrix:

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{n5} \end{pmatrix}$$