Computer Intensive Methods using R

Part 3: bootstrap confidence intervals

Prof. Dr. Ziv Shkedy

Master of Statistics
Hasselt University

General Information

Overview of the course

- Bootstrap confidence intervals.
  - Bootstrap t intervals.
  - The BCa intervals.
  - The percentile intervals

Overview of the course (part 1)

The Bootstrap algorithm

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

Estimation:
- Accuracy of statistics.
- Confidence intervals.

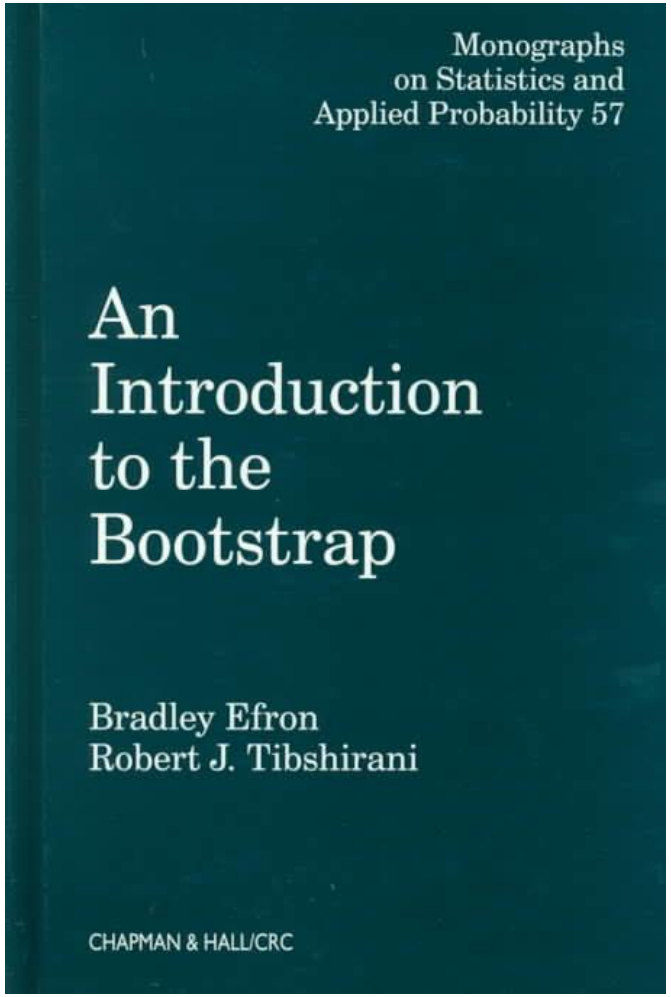
Inference:
- One sample tests.
- Two-samples tests.
- Bootstrap and permutation tests.

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

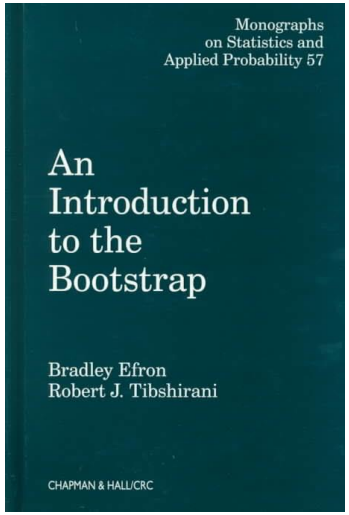
- Slides.
- R program.
- R datasets & External datasets.
- YouTube tutorials.
- Videos for the classes (highlights of each class in the course).

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

Bootstrap confidence intervals



Chapter 12,13 & 14

Topics

- Bootstrap confidence intervals:
  - Bootstrap t intervals.
  - Bootstrap standard normal interval.
  - The percentile interval.
  - The BCa method.

  the mouse data
- Examples:
  - Bootstrap interval for the standard error of the mean (the mouse data).
  - Bootstrap interval for correlation (the low school data).

The setting

Consider a population with distribution function

$$X \sim F(\theta)$$

Parameter of primary interest

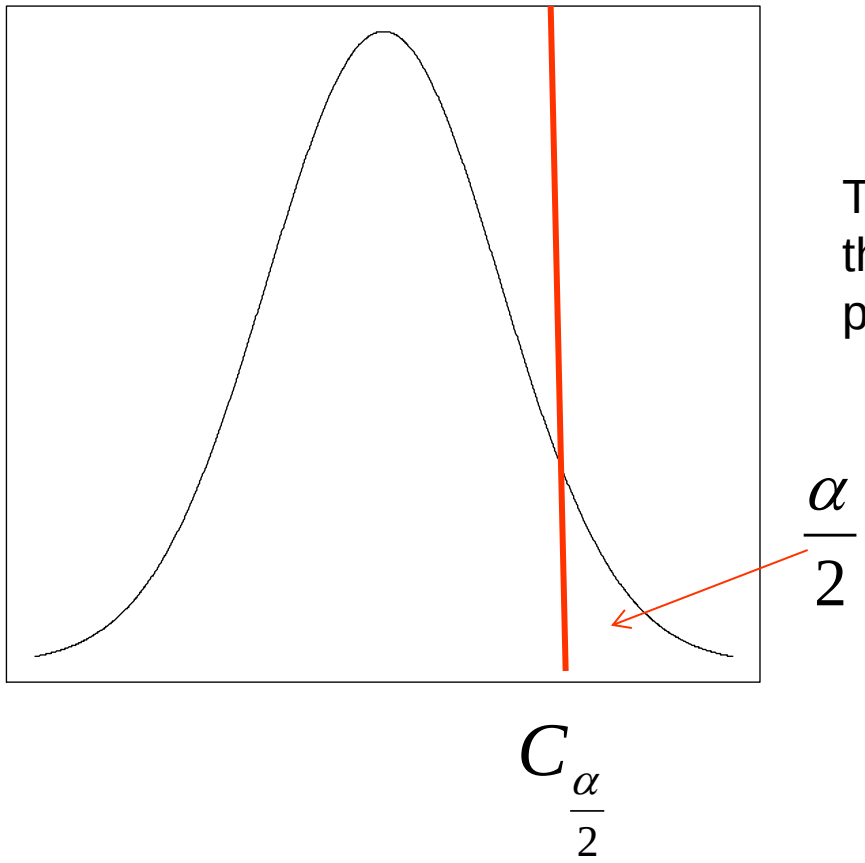
$$\theta = t(F)$$

Parameter estimate (point estimate)

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

$$s.e(\hat{\theta})$$

Confidence interval



$$\left[\hat{\theta} - C_\alpha s.e(\hat{\theta}); \hat{\theta} + C_\alpha s.e(\hat{\theta})\right]$$

The confidence interval is based on the asymptotic distribution of the parameter estimate

$$\frac{\hat{\theta} - \theta}{s.e(\hat{\theta})} \sim G$$

The classical confidence intervals

$$\frac{\hat{\theta}-\theta}{s.e(\hat{\theta})} \sim N(0,1)$$

$$\frac{\hat{\theta}-\theta}{\hat{s}.e(\hat{\theta})} \sim t_{n-1}$$

$$\left[\underbrace{\hat{\theta}-Z_{\frac{\alpha}{2}}s.e(\hat{\theta})}_{L};\underbrace{\hat{\theta}+Z_{\frac{\alpha}{2}}s.e(\hat{\theta})}_{U}\right]$$

$$\left[\hat{\theta}-t_{\left(n-1,\frac{\alpha}{2}\right)}\hat{s}.e(\hat{\theta});\hat{\theta}+t_{\left(n-1,\frac{\alpha}{2}\right)}\hat{s}.e(\hat{\theta})\right]$$

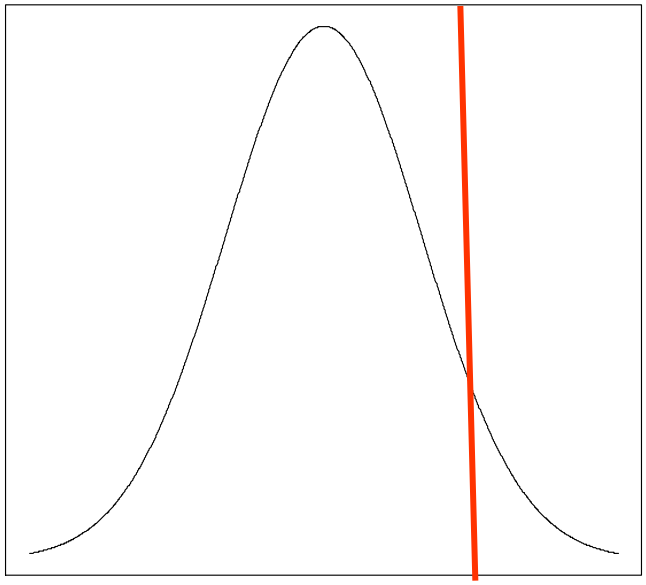
$$\hat{\theta}\in[L,U]$$

The classical confidence intervals

Interpretation

$$\left[\underbrace{\hat{\theta} - C_{\frac{\alpha}{2}} s.e(\hat{\theta})}_{L}; \underbrace{\hat{\theta} + C_{\frac{\alpha}{2}} s.e(\hat{\theta})}_{U}\right]$$

$$(L,U) = \left(\hat{\theta}_L, \hat{\theta}_U\right)$$



$C_{\frac{\alpha}{2}}$

$$P\left(\theta \in [L,U]\right) = 1 - \alpha$$

Bootstrap t interval

Draw B bootstrap samples and calculate the statistic

$$Z^*(b) = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{s}.e^*(\hat{\theta})}$$

Find the quantile of the bootstrap distribution of the replicates of Z

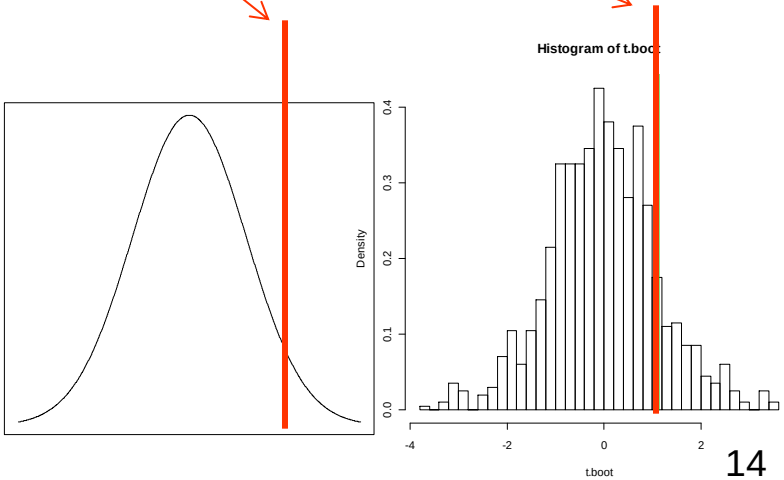
$$\frac{\#\left\{Z^*(b) > \hat{t}_\alpha\right\}}{B} = \alpha$$

1-2α confidence interval

$$\left[\hat{\theta} - \hat{t}_{(1-\alpha)}\hat{s}.e(\hat{\theta}); \hat{\theta} + \hat{t}_{(\alpha)}s.e(\hat{\theta})\right]$$

Intuition:
We replace the asymptotic distribution with the distribution of the bootstrap replicates for the statistic Z

$$C_\alpha \Rightarrow \hat{t}_\alpha$$



Bootstrap standard normal interval

Draw B bootstrap replicates from

$$\hat{\theta}^* \sim N(\hat{\theta}, \hat{s}.e(\hat{\theta}))$$

Find the quantile of the bootstrap distribution of the replicates

$$\left[\hat{\theta}_{lo}; \hat{\theta}_{up}\right] = \left[\hat{\theta}^*_{\alpha}; \hat{\theta}^*_{1-\alpha}\right]$$

Intuition:

Which type of bootstrap is applied ?

Which assumptions we need to make which we did not made for the previous method ?

What are the disadvantages of the two methods ?

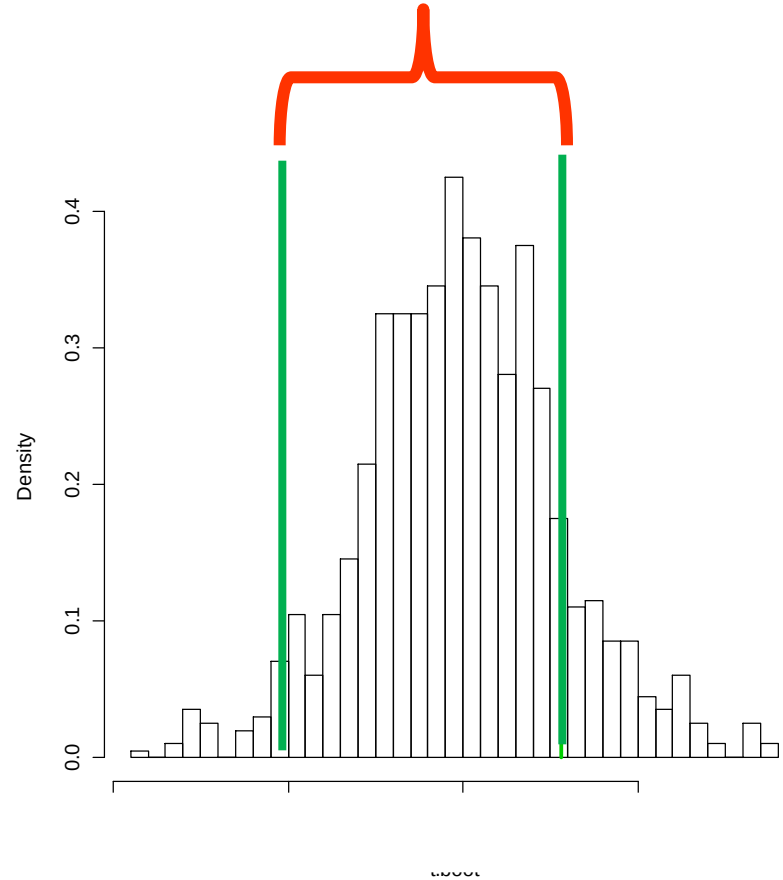
The percentile interval

Draw B bootstrap samples and calculate the bootstrap replicates

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

Find the quantile of the bootstrap distribution of the replicates

$$\left[\hat{\theta}_{lo}; \hat{\theta}_{up}\right] = \left[\hat{\theta}_B^{*(\alpha)}; \hat{\theta}_B^{*(1-\alpha)}\right]$$



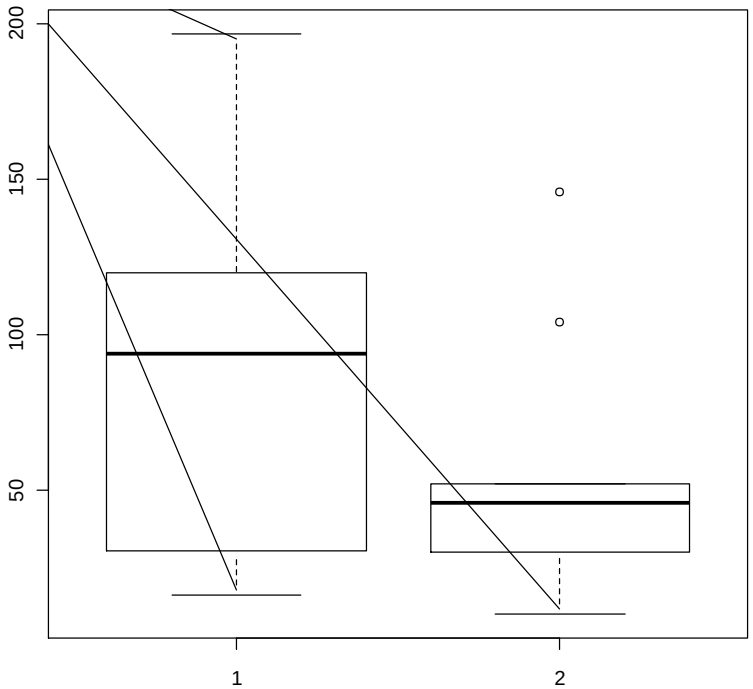
Distribution of the bootstrap replicates

Example 1
The mouse data

Bootstrap t interval

Example: the mouse data (first group: treatment)

```
> z<-c(94,197,16,38,99,141,23)
> y<-c(52,104,146,10,51,30,40,27,46)
> z
[1]  94 197  16  38  99 141  23
> y
[1]  52 104 146  10  51  30  40  27
   46
> boxplot(z,y)
```



- 16 mice randomly assigned to treatment and control group.
- Survival time following a test surgery.

Classical C.I.

```
> z <- c(94, 197, 16, 38, 99, 141, 23)
> z
[1]  94 197  16  38  99 141  23
> t.hat<-mean(z)
> se.t.hat<-sqrt(var(z)/7)
> t.hat
[1] 86.85714
> se.t.hat
[1] 25.23549
> mean(z)+1.645*(sd(z)/sqrt(7))
[1] 128.3695
> mean(z)-1.645*(sd(z)/sqrt(7))
[1] 45.34476
```

```
> t.test(z,conf.level=0.9)

        One Sample t-test

data:  z
t = 3.4419, df = 6, p-value = 0.01377
alternative hypothesis: true mean is not
    equal to 0
90 percent confidence interval:
  37.82004 135.89425
sample estimates:
mean of x
 86.85714
```

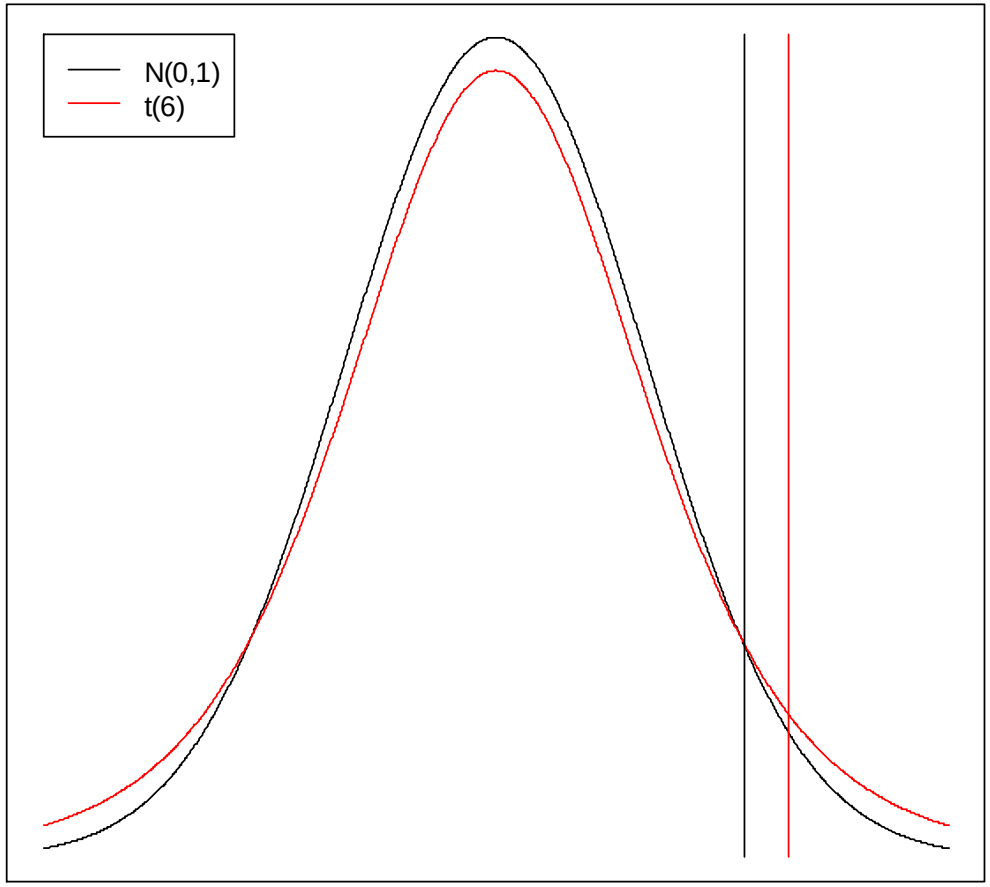
$$\left[\hat{\theta} - t_{\left(n-1,\frac{\alpha}{2}\right)} \hat{s}.e(\hat{\theta}); \hat{\theta} + t_{\left(n-1,\frac{\alpha}{2}\right)} \hat{s}.e(\hat{\theta})\right]$$

Classical C.I.
(the critical values)

$$\alpha = 0.05$$

$$C_\alpha = 1.645 \qquad N(0,1)$$

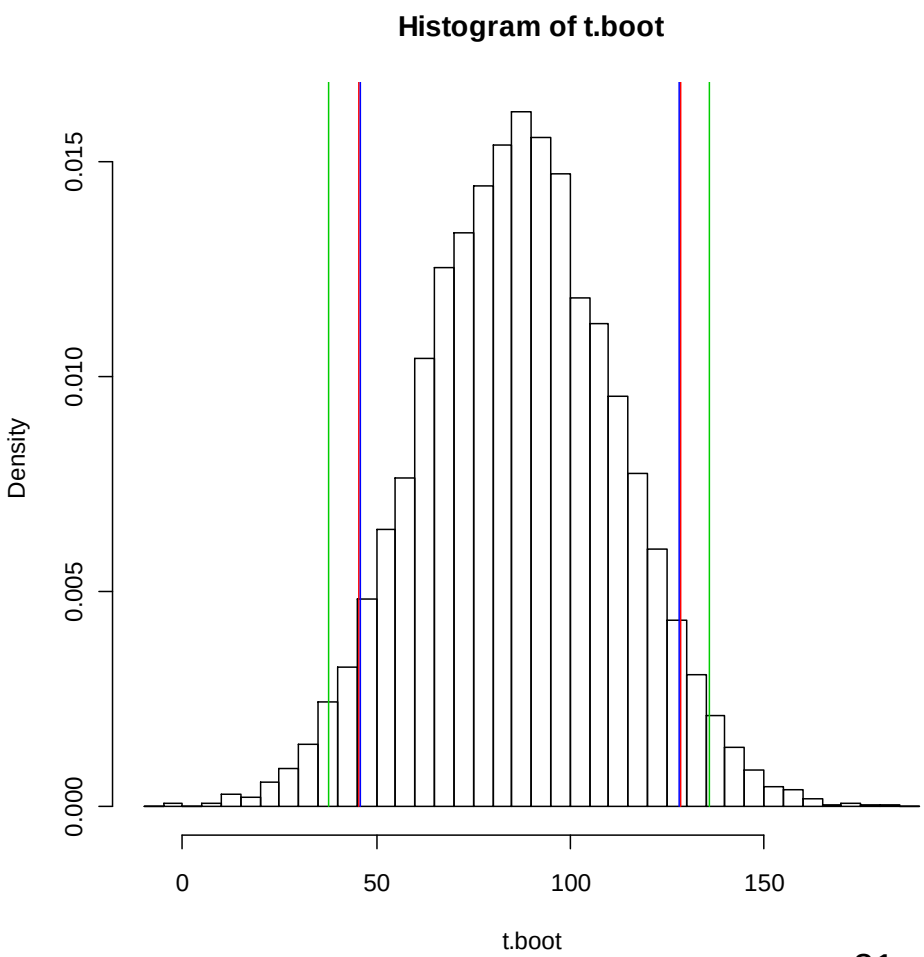
$$C_\alpha = 1.9431 \qquad t_{(6)}$$



Bootstrap standard normal interval

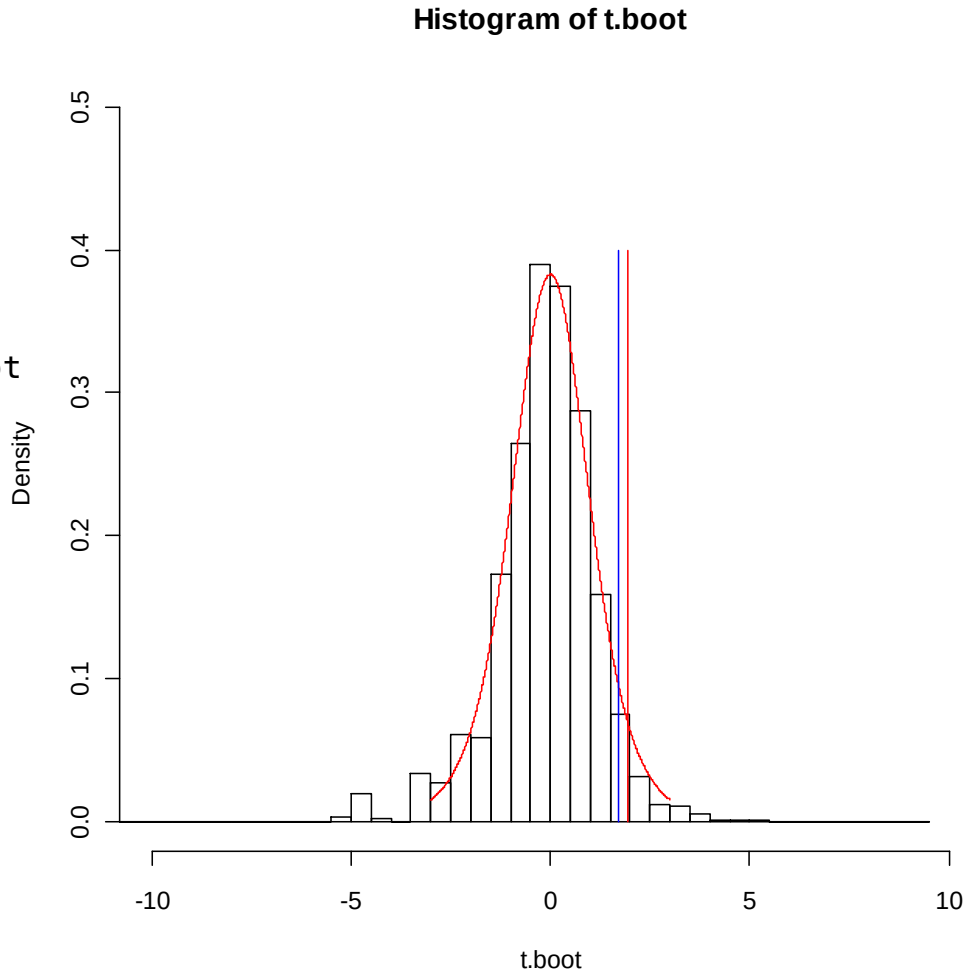
```
> B=10000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ t.boot[b]<-rnorm(1,t.hat,se.t.hat)
+ }
>
> quantile(t.boot,probs=c(0.05,0.95))
      5%      95%
 45.9621 128.2926
```

$$\hat{\theta}_i^* \sim N\left(\hat{\theta}, \frac{s^2}{n}\right)$$



Bootstrap t interval

```
> B=10000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ x.boot<-
    sample(z,size=length(z),replace=T)
+ se.boot<-sqrt(var(x.boot)/length(z))
+ t.boot[b]<-(mean(x.boot)-t.hat)/se.boot
+ }
> quantile(t.boot,probs=c(0.05,0.95))
       5%       95%
-2.248523  1.712321
> qt(0.95,6)
[1] 1.943180
> qt(0.05,6)
[1] -1.943180
```



$$Z^*(b) = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{s}.e(\hat{\theta})}$$

$$\hat{s}.e(\hat{\theta}) = \frac{s^2}{n}$$

The usual estimate for the standard error at each bootstraap sample

Bootstrap t interval

$$Z^*(b) = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{s}.e^*(\hat{\theta})}$$

$\hat{s}.e^*(\hat{\theta})$: Bootstrap estimate for the standard error at each bootstrap sample.

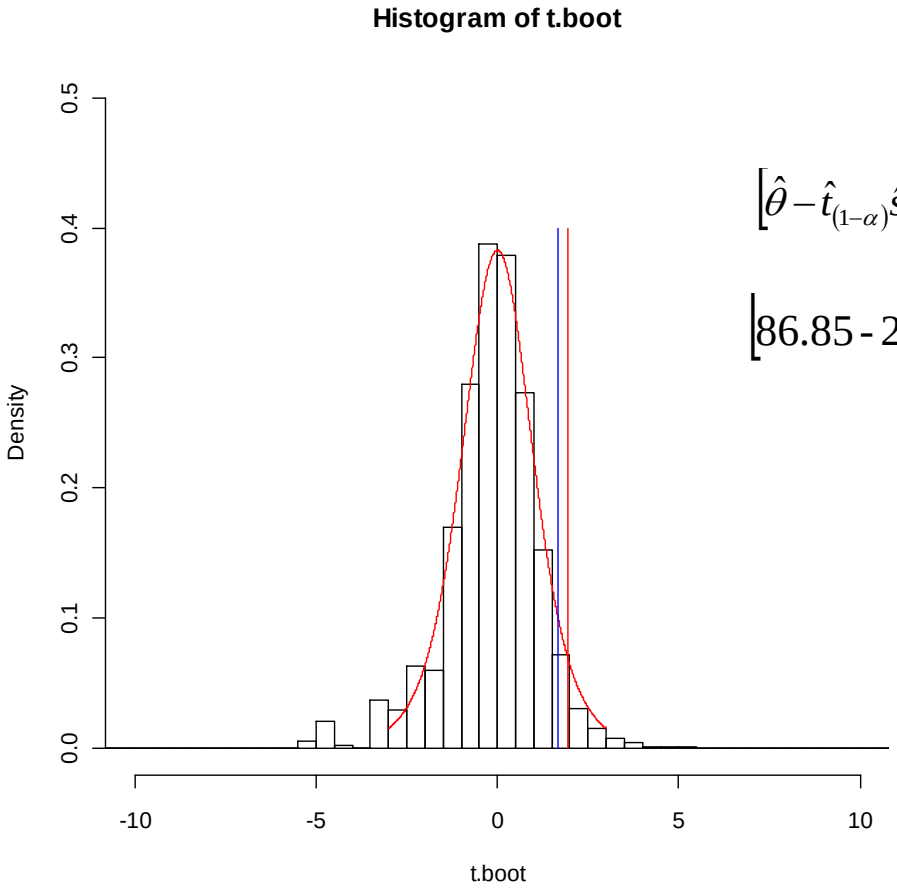
A nested bootstrap algorithm:

For each bootstrap sample: an inner of K bootstraps to obtain

The outer loop: B bootstrap samples to obtain

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

Bootstrap t interval



$$\left[\hat{\theta} - \hat{t}_{(1-\alpha)}\hat{s}.e(\hat{\theta}); \hat{\theta} + \hat{t}_{(\alpha)}s.e(\hat{\theta})\right]$$

$$\left[86.85 - 2.248523 \times \hat{s}.e(\hat{\theta}); 86.85 + 1.712321 \times s.e(\hat{\theta})\right]$$

Bootstrap t interval: transformation

A (1-2α)% C.I for θ:

$$\left[\underbrace{\hat{\theta}-\hat{t}_{(1-\alpha)}\hat{s}.e(\hat{\theta})}_{\hat{\theta}^{low}};\underbrace{\hat{\theta}+\hat{t}_{(\alpha)}s.e(\hat{\theta})}_{\hat{\theta}^{up}}\right]$$

Transformation for θ:

$$\eta = t(\theta)$$

The confidence interval:

$$\left[\eta\left(\hat{\theta}^{low}\right),\eta\left(\hat{\theta}^{up}\right)\right]$$

Does not have a coverage probability of a (1-2α)%.

Bootstrap t interval is not transformation respecting.

Example 2
The mouse data

Standard normal interval

Bootstrap standard normal interval

Draw B bootstrap replicates from

$$\hat{\theta}^* \sim N(\hat{\theta}, \hat{s}.e(\hat{\theta}))$$

Find the quantile of the bootstrap distribution of the replicates

$$\left[\hat{\theta}_{lo}; \hat{\theta}_{up}\right] = \left[\hat{\theta}^*_{\alpha}; \hat{\theta}^*_{1-\alpha}\right]$$

Intuition:

Which type of bootstrap is applied ?

Which assumptions we need to make which we did not made for the previous method ?

What are the disadvantages of the two methods ?

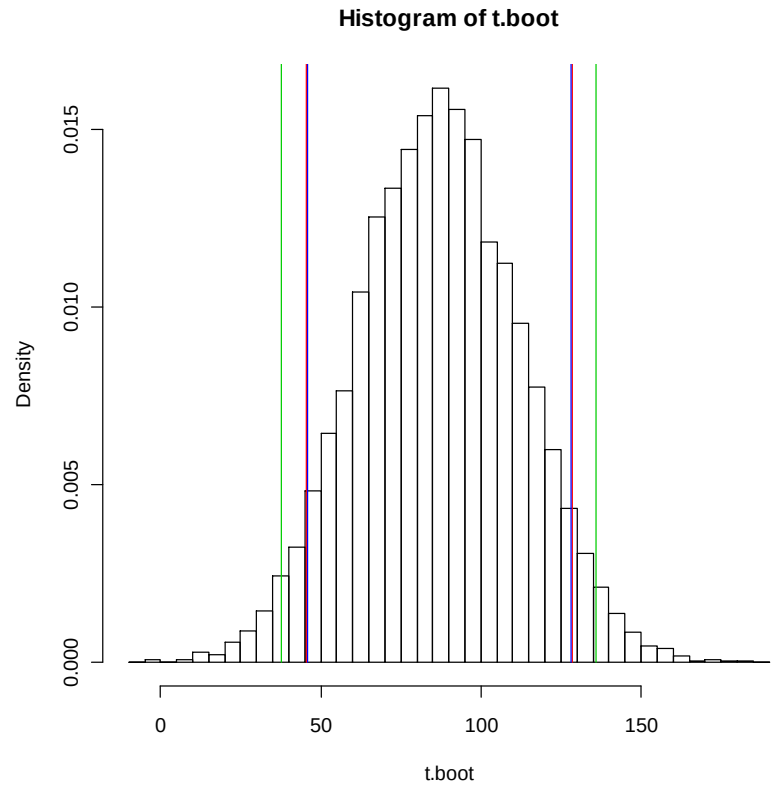
Bootstrap standard normal interval

$$\hat{\theta}_i^* \sim N\left(\hat{\theta}, \frac{s^2}{n}\right)$$

```
> B=10000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ t.boot[b]<-rnorm(1,t.hat,se.t.hat)
+ }
>
> quantile(t.boot,probs=c(0.05,0.95))
      5%      95%
 45.9621 128.2926
```

$$\left[\hat{\theta} - Z_{\frac{\alpha}{2}} s.e(\hat{\theta}); \hat{\theta} + Z_{\frac{\alpha}{2}} s.e(\hat{\theta})\right]$$

```
> t.hat
[1] 86.85714
> se.t.hat
[1] 25.23549
> mean(z)+1.645*(sd(z)/sqrt(7))
[1] 128.3695
> mean(z)-1.645*(sd(z)/sqrt(7))
[1] 45.34476
```



Do we really need this interval ?

Example 3
The mouse data

Bootstrap percentiles interval

The percentile interval

- Draw B bootstrap samples from $\hat{F}$.
- Calculate the bootstrap replicates:

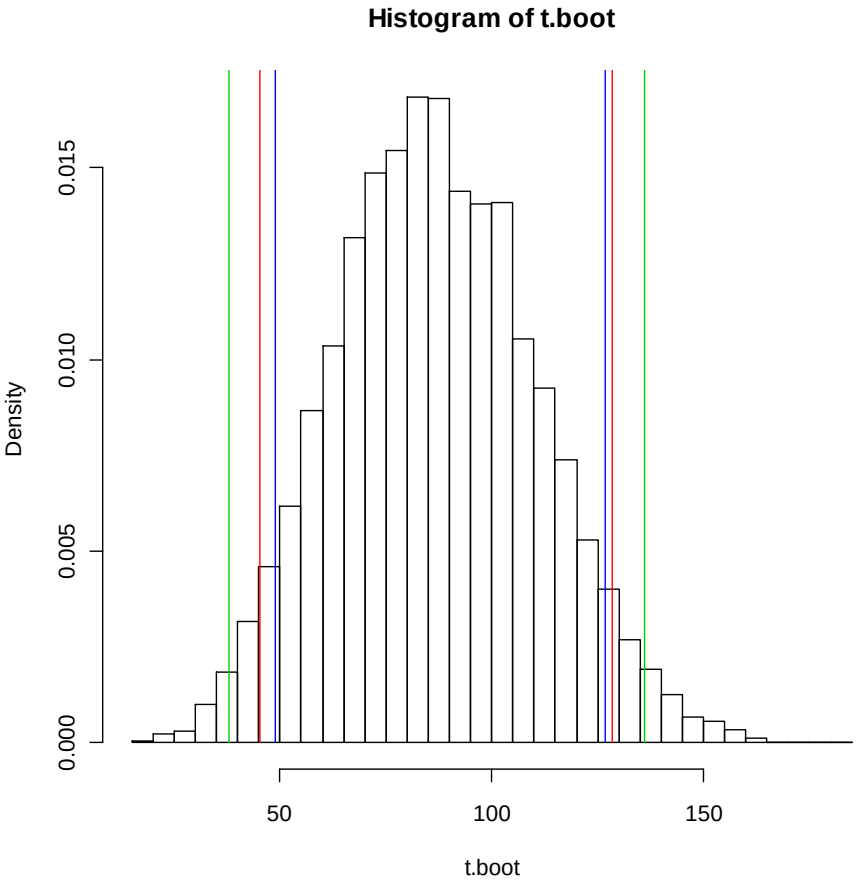
$$\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots\ldots\ldots\ldots \hat{\theta}_B^*$$

- The bootstrap percentile interval:

$$\left[\hat{\theta}_B^{*(1-\alpha)}; \hat{\theta}_B^{*(\alpha)}\right]$$

The percentile interval

```
> B=10000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ x.boot<-sample(z,size=length(z),replace=T)
+ t.boot[b]<-mean(x.boot)
+ }
>
> quantile(t.boot,probs=c(0.05,0.95))
       5%       95%
 49.42857 126.42857
```



Bootstrap confidence intervals summary (1)

method	**Bootstrap type**	**Assumption**	**Symmetry**
Classical	None		
Bootstrap t intervals			
Bootstrap Normal intervals			
Percentile confidence interval			

Transformation respecting property

Parameter of primary interest

$$\theta$$

Transformation

$$\lambda = m(\theta)$$

Percentile interval

$$\left[\hat{\theta}^*_{\alpha}; \hat{\theta}^*_{1-\alpha}\right]$$

Percentile interval

$$\left[\hat{\lambda}^*_{\alpha}; \hat{\lambda}^*_{1-\alpha}\right] = \left[m(\hat{\theta}^*_{\alpha}); m(\hat{\theta}^*_{1-\alpha})\right]$$

Both intervals have the same coverage probability.

Example
The mouse data

Bootstrap interval for the standard error of the mean

The standard error of the sample mean

population

$$X \sim (\mu_F, \sigma_F{}^2)$$

$$\sigma_F^2 = E_F[(x - \mu_F)^2]$$

sample

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

$$Var(\bar{x}) = \frac{1}{n^2} Var(\sum_{i=1}^{n} x_i) = \frac{\sigma_F{}^2}{n}$$

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

The percentile interval for the S.E of the sample mean

Population

$$X \sim (\mu_F, \sigma_F{}^2)$$

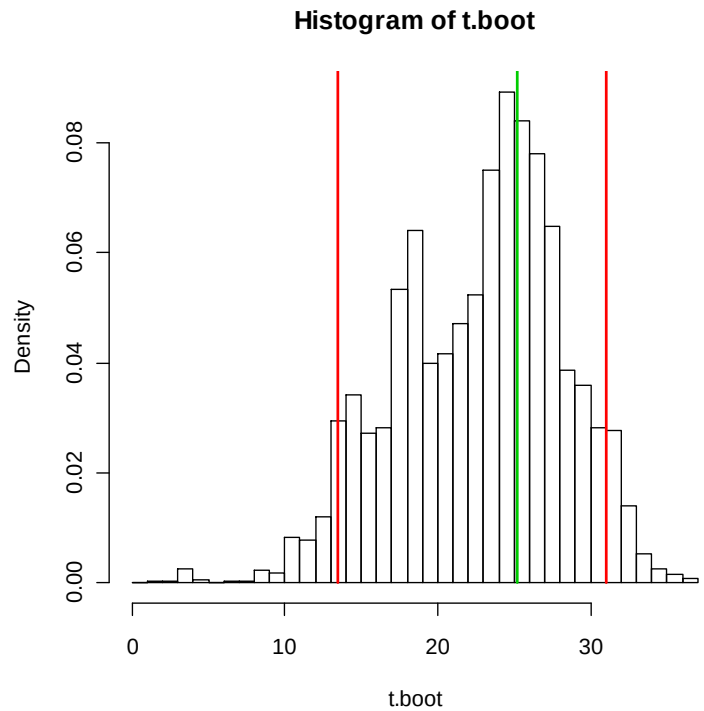
Sample

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

$$Var(\bar{x}) = \frac{1}{n^2} Var(\sum_{i=1}^{n} x_i) = \frac{\sigma_F{}^2}{n}$$

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

$$\frac{\hat{\sigma}_F}{\sqrt{n}} = 25.23$$




```
> quantile(t.boot,probs=c(0.05,0.95))
     5%      95%
13.51895 30.83807
```

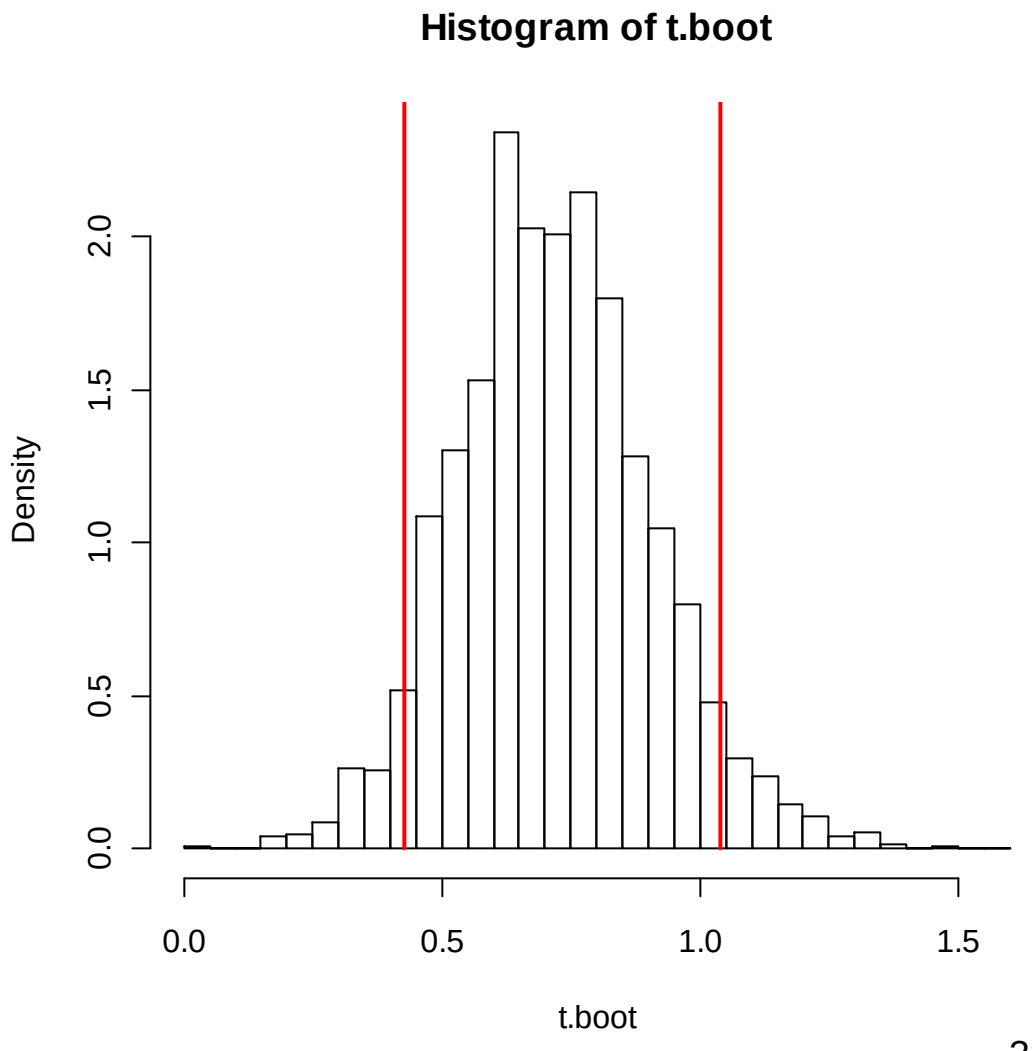
R code: the percentile interval for the S.E of the sample mean

```
> B=10000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ x.boot<-sample(z,size=length(z),replace=T)
+ t.boot[b]<-sqrt(var(x.boot)/7)
+ }
> quantile(t.boot,probs=c(0.05,0.95))
      5%      95%
13.51895 30.83807
```

Coefficient of variation

$$CV = \frac{\sigma_F}{\mu}$$

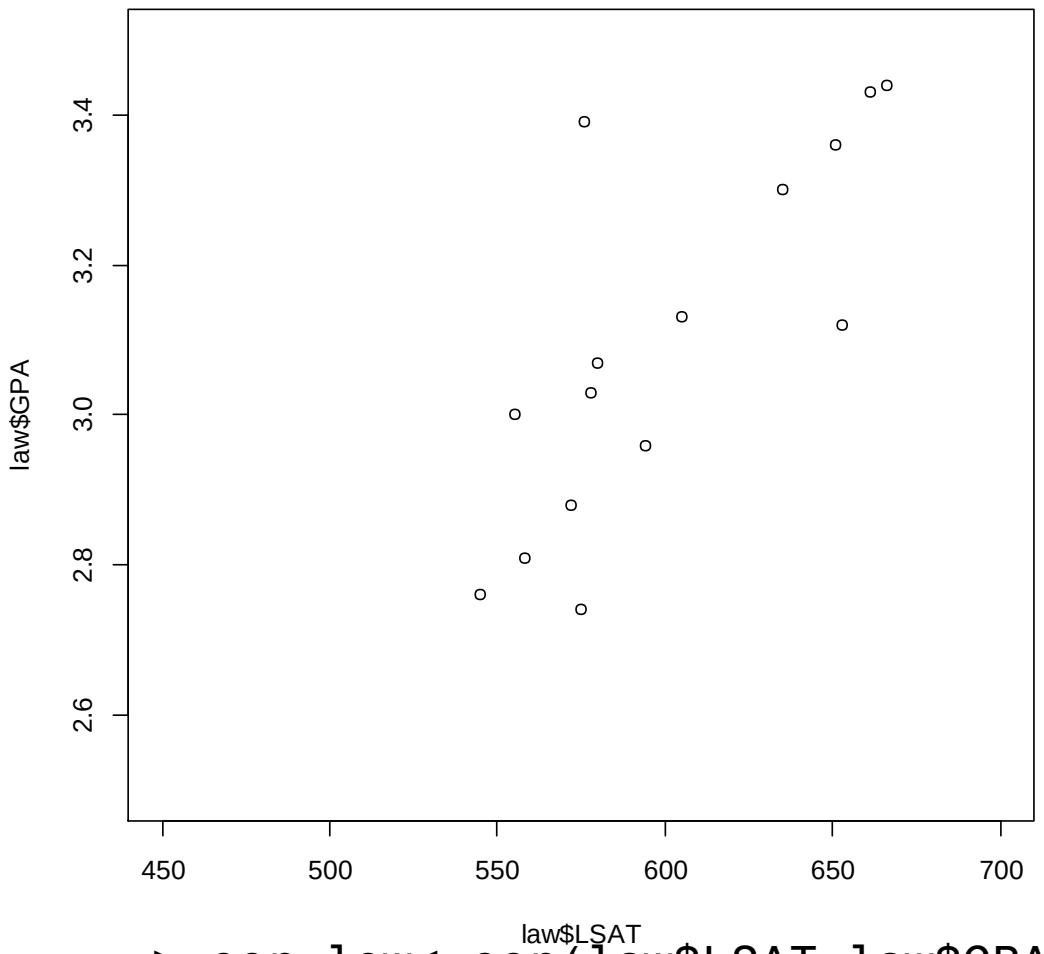
```
> sqrt(var(z))/mean(z)
[1] 0.768697
```



Example
The law school data

correlation

The law school data




```
> cor.low<-cor(law$LSAT,law$GPA)
> cor.low
[1] 0.7763745
```

The non parametric bootstrap algorithm

The observed sample

x_1	y_1
x_2	y_2
.	.
.	.
.	.
x_{15}	y_{15}

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

n=15

x^*_1	y^*_1
x^*_2	y^*_2
.	.
.	.
.	.
x^*_{15}	y^*_{15}

The bootstrap sample

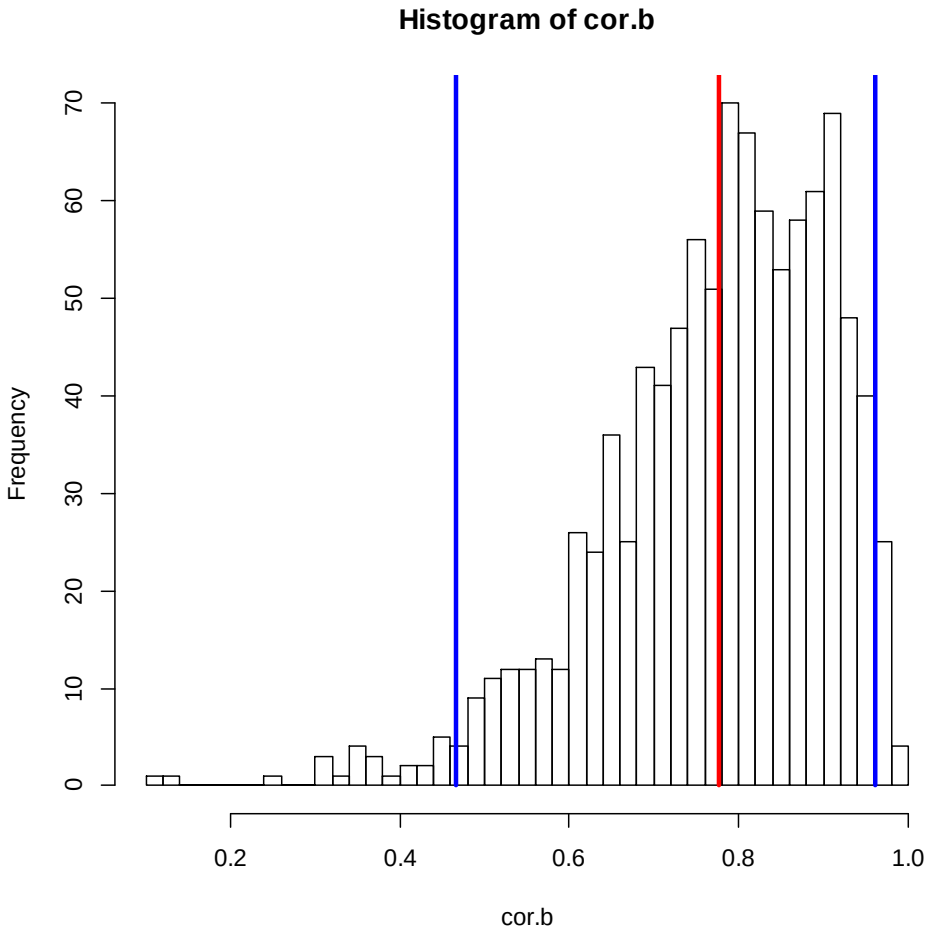
The non parametric bootstrap algorithm

The bootstrap sample

For each bootstrap sample we calculate the correlation

$$\begin{array}{cc} x^*_1 & y^*_1 \\ x^*_2 & y^*_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x^*_{15} & y^*_{15} \end{array} \longrightarrow \hat{\rho}^*_b(x^*, y^*)$$

Percentile interval for the correlation




```
> ci<-
quantile(cor.b,probs=c(0.025,0.975)
)
> ci
     2.5%     97.5%
0.4466008 0.9607133
```

R code

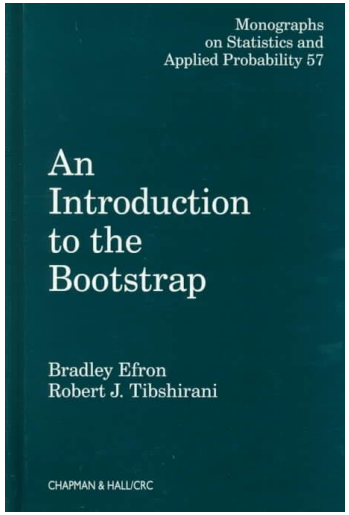
```
> B<-1000
> cor.b<-c(1:B)
> n<-length(law$LSAT)
> index<-c(1:n)
>
> for(i in 1:B)
+ {
+ index.b<-sample(index,n,replace=TRUE)

+ LAST.b<-law$LSAT[index.b]
+ GPA.b<-law$GPA[index.b]

+ cor.b[i]<-cor(LAST.b,GPA.b)
+ }
>
```

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

Example
The spatial test data



Chapter 14, Secion:14.1-14.3

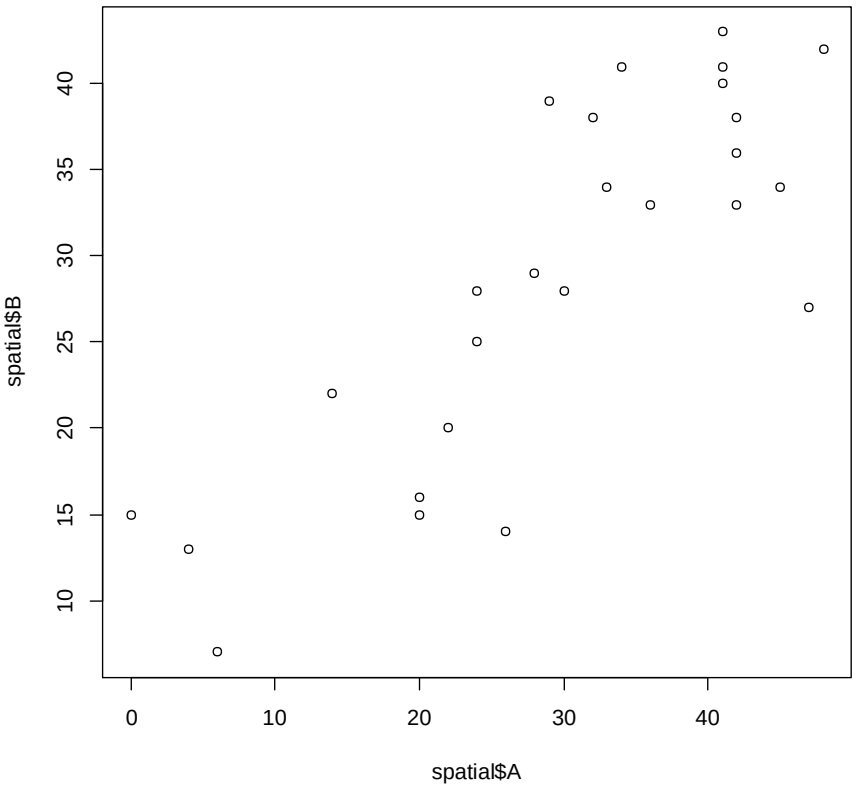
The BCa method

- Bias corrected intervals.
- For the example, two steps:
  - Step 1: produce bootstrap replicates for percentile interval (non parametric and parametric).
  - Step 2: modify the lower and upper limits for the BCa.

Example: the spatial test data

- Twenty-six neurologically impaired children have each taken two tests of spatial perception, called "A" and "B".
- In R:

```
> help(spatial)
```



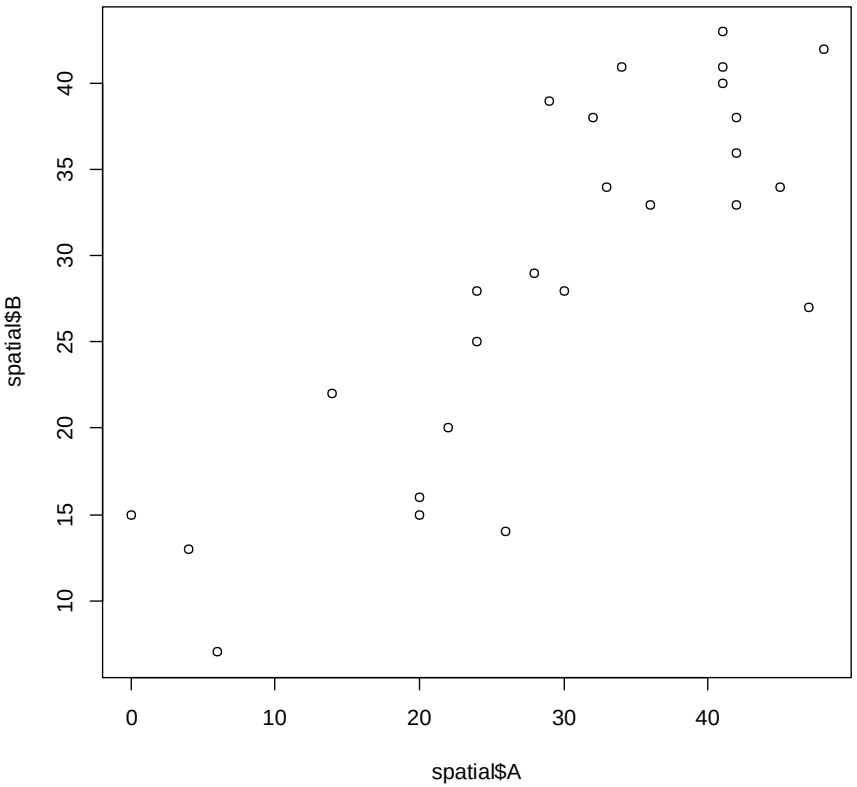
Parameter of interest

- The variance of A:

$$\sigma_A^2 = \mathrm{var}(A)$$

```
> Ai<-spatial$A
> Bi<-spatial$B
> mean(Ai)
[1] 29.65385
> mean(Bi)
[1] 28.88462
> cov(spatial)
          A        B
A 178.3954 116.9585
B 116.9585 113.7862
```

$\hat{\sigma}_A^2$



The covariance matrix of A and B.

Non parametric bootstrap percentile interval

Bootstrap B samples for the empirical distribution (with replacement).

Resampling pairs:

$$x_i^* = (A_i, B_i)^*$$

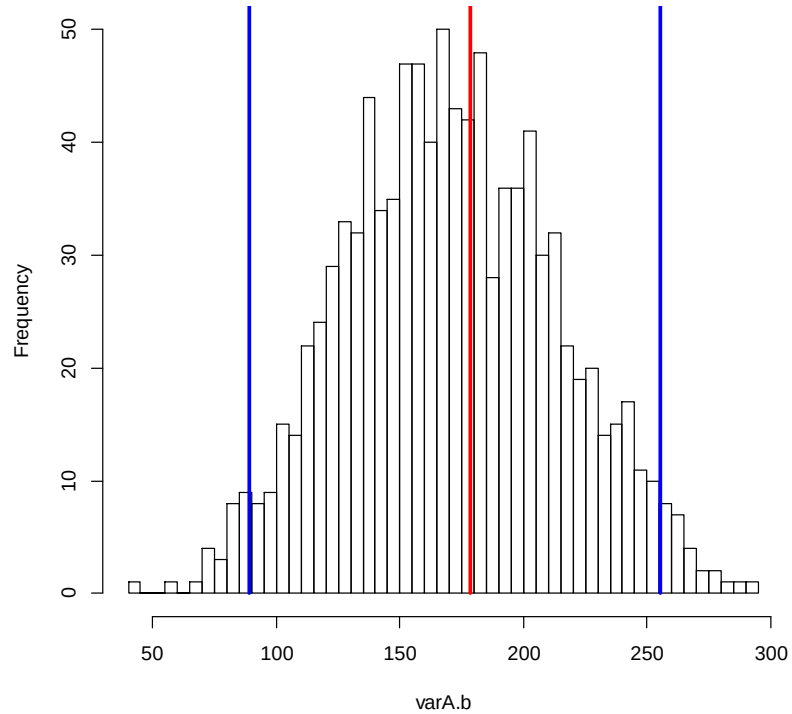
For each bootstrap sample:

$$\hat{F} \rightarrow (x_1^*, x_2^*, .., x_n^*)$$

Bootstrap replicates:

$$\theta_1^*, \theta_B^*, ..., \theta_B^*$$

Distribution of the bootstrap replicates and 95% percentile interval




```
> ci
     2.5%     97.5%
 89.10865 255.32465
```

Parametric bootstrap percentile interval

Bootstrap B samples for the bivariate normal distribution:

$$F_{norm} = \begin{bmatrix} A_i \\ B_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \Sigma \right)$$

Estimate the unknown parameters:

$$\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} = \begin{bmatrix} \overline{A} \\ \overline{B} \end{bmatrix}$$ The mean vector

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_A^2 & \hat{\sigma}_{AB} \\ \hat{\sigma}_{AB} & \hat{\sigma}_B^2 \end{bmatrix}$$ Covariance matrix

Bootstrap B samples for the bivariate normal distribution:

$$F_{norm} = \begin{bmatrix} A_i^* \\ B_i^* \end{bmatrix} \sim N\left(\begin{bmatrix} \overline{A} \\ \overline{B} \end{bmatrix}, \hat{\Sigma} \right)$$

Bootstrap replicates:

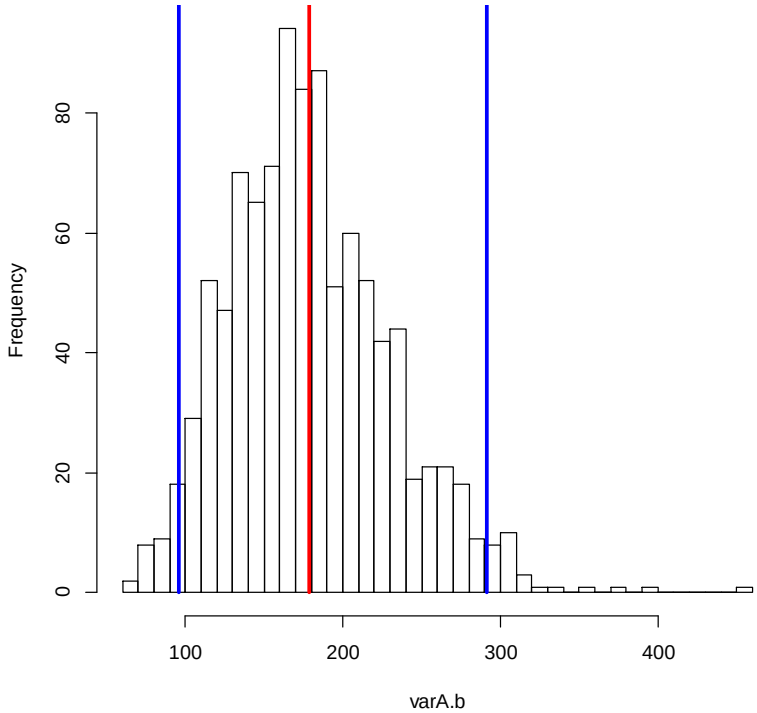
$$\theta_1^*, \theta_B^*, \ldots, \theta_B^*$$

Parametric bootstrap percentile interval

Parameter estimates

```
> mean.ab<-c(mean(Ai),mean(Bi))
> mean.ab
[1] 29.65385 28.88462
> cov.ab<-cov(spatial)
> cov.ab
         A        B
A 178.3954 116.9585
B 116.9585 113.7862
```

Distribution of the bootstrap replicates and 95% percentile interval




```
> ci
    2.5%    97.5%
95.7222 291.3740
```

The BCa method

Percentile interval

$$\left[\hat{\theta}_{lo};\hat{\theta}_{up}\right]=\left[\hat{\theta}^{*(\alpha)};\hat{\theta}^{*(1-\alpha)}\right]$$

BCa interval

$$\left[\hat{\theta}_{lo};\hat{\theta}_{up}\right]=\left[\hat{\theta}^{*(\alpha_1)};\hat{\theta}^{*(\alpha_2)}\right]$$

$$\alpha_1=\phi\left(\hat{z}_0+\frac{\hat{z}_0+z^{(\alpha)}}{1-\hat{a}\left(\hat{z}_0+z^{(\alpha)}\right)}\right)$$

$$\alpha_2=\phi\left(\hat{z}_0+\frac{\hat{z}_0+z^{1-(\alpha)}}{1-\hat{a}\left(\hat{z}_0+z^{(1-\alpha)}\right)}\right)$$

A percentile interval with modification for the upper and lower limits of the intervals.

See extra slides about BCa

The BCa method

$$\hat{z}_0 = \phi^{-1}\left(\frac{\#\left\{\hat{\theta}_b^* < \hat{\theta}\right\}}{B}\right)$$

The proportion of bootstrap replicates smaller than the observed statistic:

$$\frac{\#\left\{\hat{\theta}_b^* < \hat{\theta}\right\}}{B}$$

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}^{(.)} - \hat{\theta}^{(-i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}^{(.)} - \hat{\theta}^{(-i)}\right)^2\right\}^{\frac{3}{2}}}$$

Can be estimated using jackknife.

$$se(\theta^*) = se(\hat{\theta}) \times \left[1 + a\left(\theta^* - \hat{\theta}\right)\right] \Rightarrow se(\theta^*) = se(\hat{\theta}), a = 0$$

The BCa method in R

```
> x<-Ai
> theta <- function(x){var(x)}
>results <- bcanon(x,1000,theta,alpha=c(0.025,0.975))

> results
$confpoints
     alpha bca point
[1,] 0.025  114.8554
[2,] 0.975  298.5046

$z0
[1] 0.2585273

$acc
[1] 0.06124012

$u
 [1] 171.2433 184.0833 181.7900 185.8100 179.2233 179.2233 181.7900 179.2233
 [9] 183.2900 180.2500 175.6233 175.2100 161.5833 147.7233 185.3433 185.7100
[17] 185.0100 157.3100 185.5900 184.4433 172.7900 180.2500 184.4433 185.2500
[25] 185.8233 180.2500
```

The parameter of interest $\mathrm{var}(A)$

$$\left[\hat{\theta}_{lo};\hat{\theta}_{up}\right]=\left[\hat{\theta}^{*(\alpha_1)};\hat{\theta}^{*(\alpha_2)}\right]$$

$$\hat{z}_0=\phi^{-1}\left(\frac{\#\left\{\hat{\theta}_b^*<\hat{\theta}\right\}}{B}\right)$$

$$\hat{a}=\frac{\sum_{i=1}^{n}\left(\hat{\theta}^{(.)}-\hat{\theta}^{(-i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}^{(.)}-\hat{\theta}^{(-i)}\right)^2\right\}^{\frac{3}{2}}}$$