Computer Intensive Methods using R

Part 4: bootstrap tests
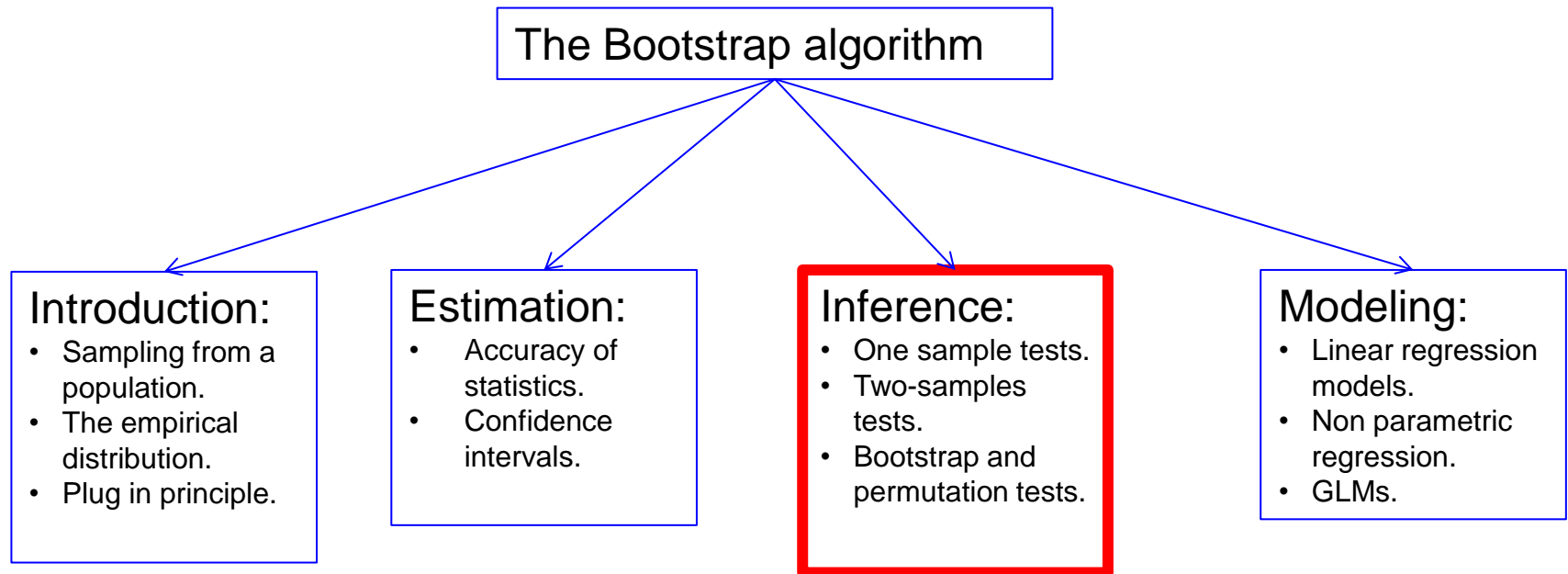
Prof. Dr. Ziv Shkedy

Master of Statistics

Hasselt University

General Information

Overview of the course

- Bootstrap tests.
  - One sample problem.
  - Two samples problem.
- Permutation tests:
  - Two samples problem.
- Correlation test.
- Binary data.

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.

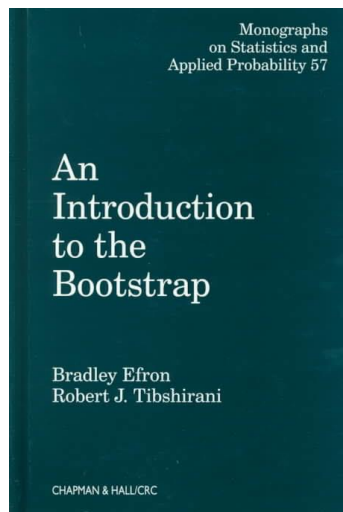- Davison A.C. and Hinkley D.V: Bootstrap Methods and Their Application.
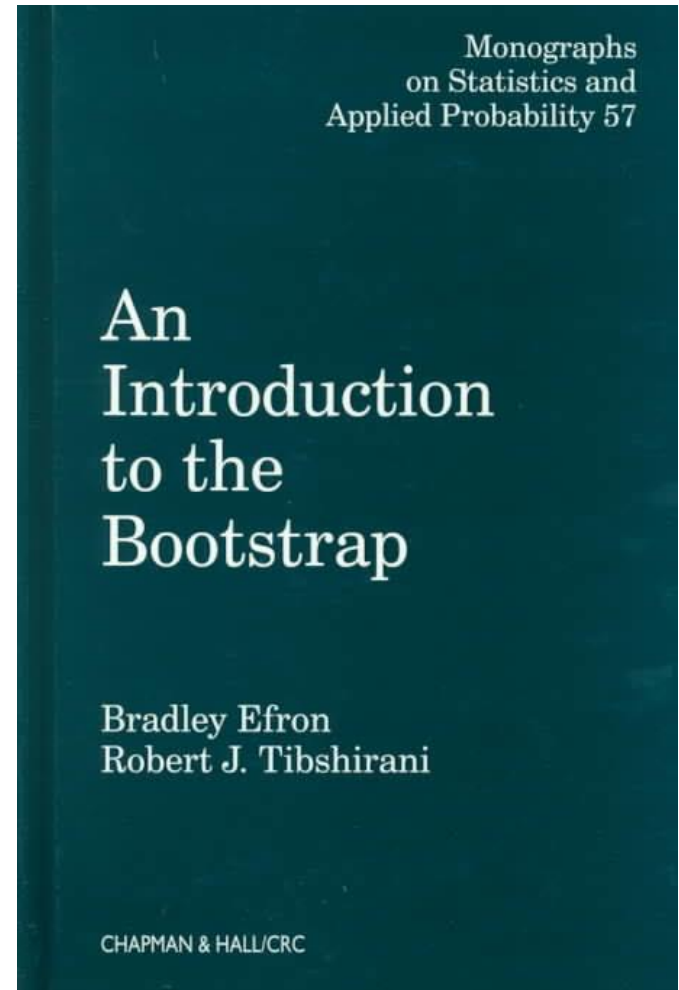
Course materials

- 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 tests

Monographs
on Statistics and
Applied Probability 57

An
Introduction
to the
Bootstrap

Bradley Efron
Robert J. Tibshirani

CHAPMAN & HALL/CRC

Chapter 16

Topics

- Testing of hypothesis in one sample.
- Testing equality of the means in two populations: algorithm 16.1.
- Testing equality of the means in two populations: algorithm 16.2.

The mouse data.

- Correlation test (the cars data).
- Ratio test and pair t-test (the city population data).

External dataset.

- Hypothesis testing for binary data (two populations).

Bootstrap tests

As before the bootstrap mechanism should reflect the original data generation mechanism and, in this setting of testing hypotheses, there is a second important rule

The bootstrap simulation should satisfy the null hypothesis.
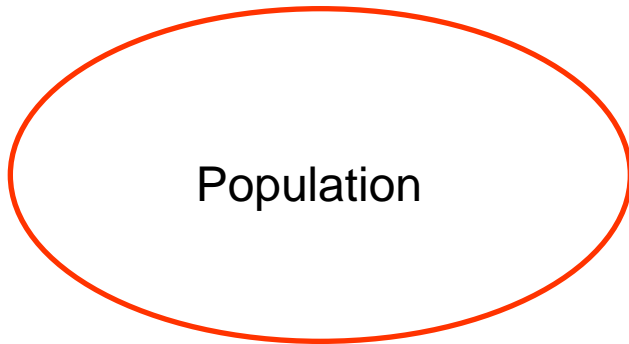
Bootstrap tests

The idea is to generate (repeatedly) new bootstrap data, reflecting the null hypothesis, recalculate the test statistic and in this way to simulate the null distribution of the test statistic.

These bootstrap test values can then be used to compute a p-value.

Example 1:
One-sample problem

The one sample problem

Population

sample

$$X_1, X_2, ..., X_n$$

Distribution of X in the population

$$X_i \sim G(\theta)$$

$$E(X_i) = \mu$$

We wish to test the hypotheses

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

One Sample problem

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Observed sample

$$x_1, x_2, ..., x_n$$

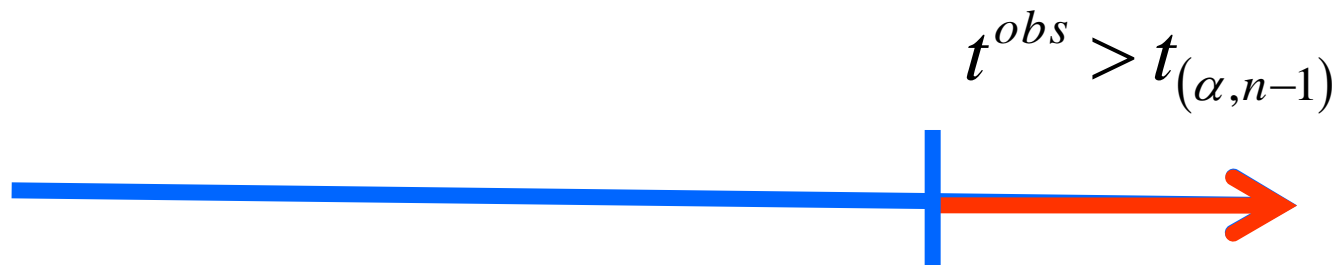Test statistic

$$t = \frac{\bar{x} - \mu_0}{se(\bar{x})}$$

Distribution of the test statistic under the null hypothesis

$$t \sim t_{(n-1)}$$
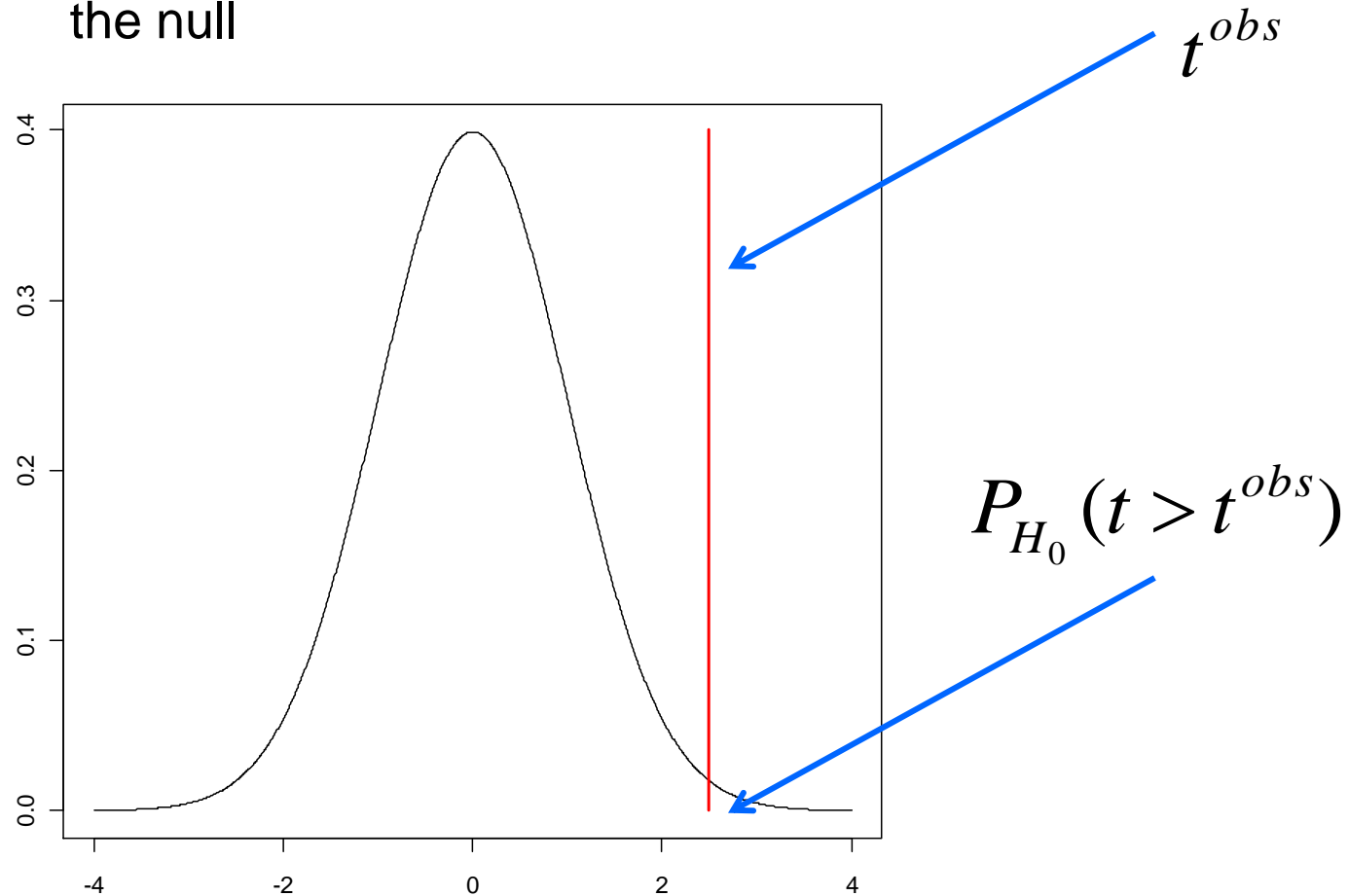
One sample problem: inference

Rejection region

$$t^{obs} > t_{(\alpha, n-1)}$$

P value
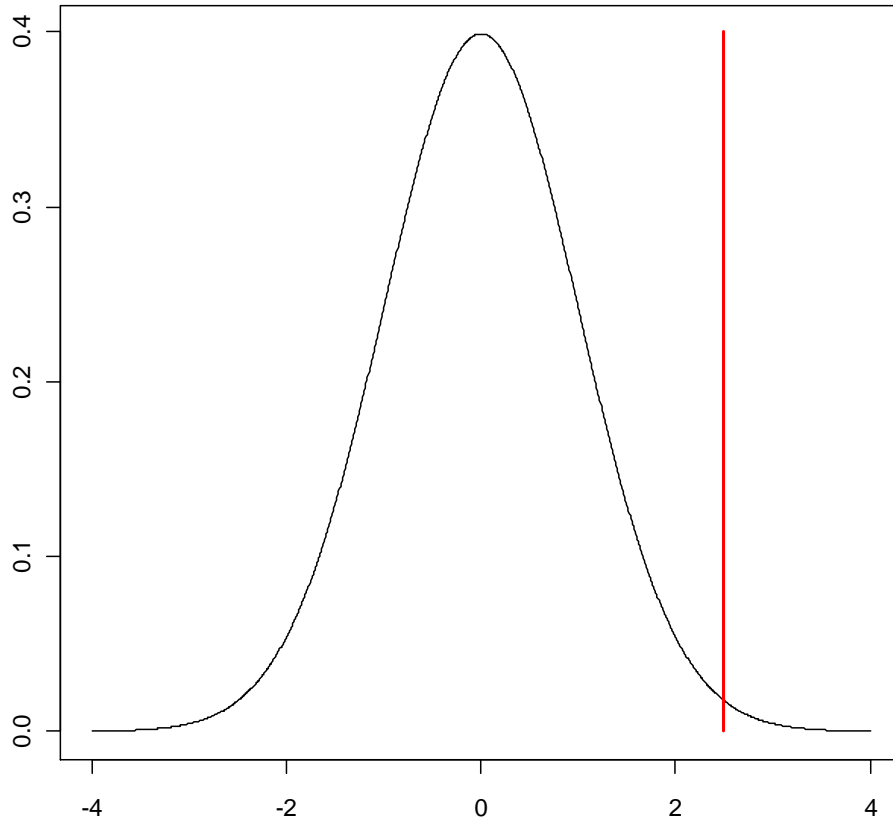
$$P_{H_0}(t > t^{obs}) < \alpha$$

The p value

Distribution of the test statistic under the null

t^{obs}

$P_{H_0}(t > t^{obs})$

16

Bootstrap test

Distribution of the test statistic under the null

We would like to approximate the distribution of the test statistic under the null hypothesis using bootstrap.

The bootstrap simulation should satisfy the null hypothesis.

17

Example: survival time of mice

- Z: survival time of 7 mice, in days, after surgery are:

```
> z
[1]  94 197  16  38  99 141  23

> mean(z)
[1] 86.85714

> sd(z)
[1] 66.76683
```

One Sample problem (two sided test)

We wish to test the null hypothesis:

$$H_0 : \mu = 129$$

$$H_1 : \mu \neq 129$$

Observed data:

$$z_1, z_2, \ldots, z_7$$

Test statistics and the distribution under the null:

$$t = \frac{\bar{z} - \mu_0}{se(\bar{z})} \qquad\qquad t \sim t_{(6)}$$

One sample t teat

```
> z <- c(94, 197, 16, 38, 99, 141, 23)
> t.test(z,mu=129)

        One Sample t-test

data:  z
t = -1.67, df = 6, p-value = 0.1460
alternative hypothesis: true mean is not
    equal to 129
95 percent confidence interval:
  25.10812 148.60616
sample estimates:
mean of x
 86.85714
```
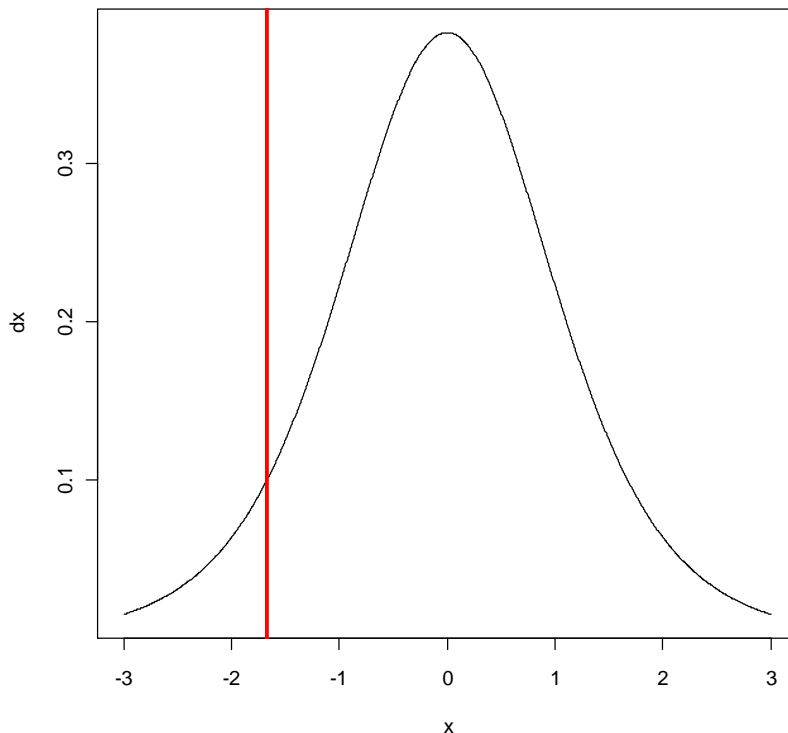
Two sided test

$$t = \frac{\bar{z} - \mu_0}{se(\bar{z})}$$

Under H0, the distribution of the test statistics:

$$t \sim t_{(6)}$$

Distribution of the test statistic under the null


```
>
t.test(z,mu=129,alternative="less")

        One Sample t-test

data:  z
t = -1.67, df = 6, p-value = 0.07298
alternative hypothesis: true mean is
less than 129
95 percent confidence interval:
    -Inf 135.8942
sample estimates:
mean of x
 86.85714
```

One sided test

```
 x<-seq(from=-3,to=3,length=1000)
> dx<-dt(x,6)
> plot(x,dx,type="l")
> lines(c(-1.67,-1.67),c(0,1),lwd=3,col=2)
> pt(-1.67,6)
[1] 0.07297866
```

Bootstrap test (1)

How can we generate the bootstrap sample under the null hypothesis ?
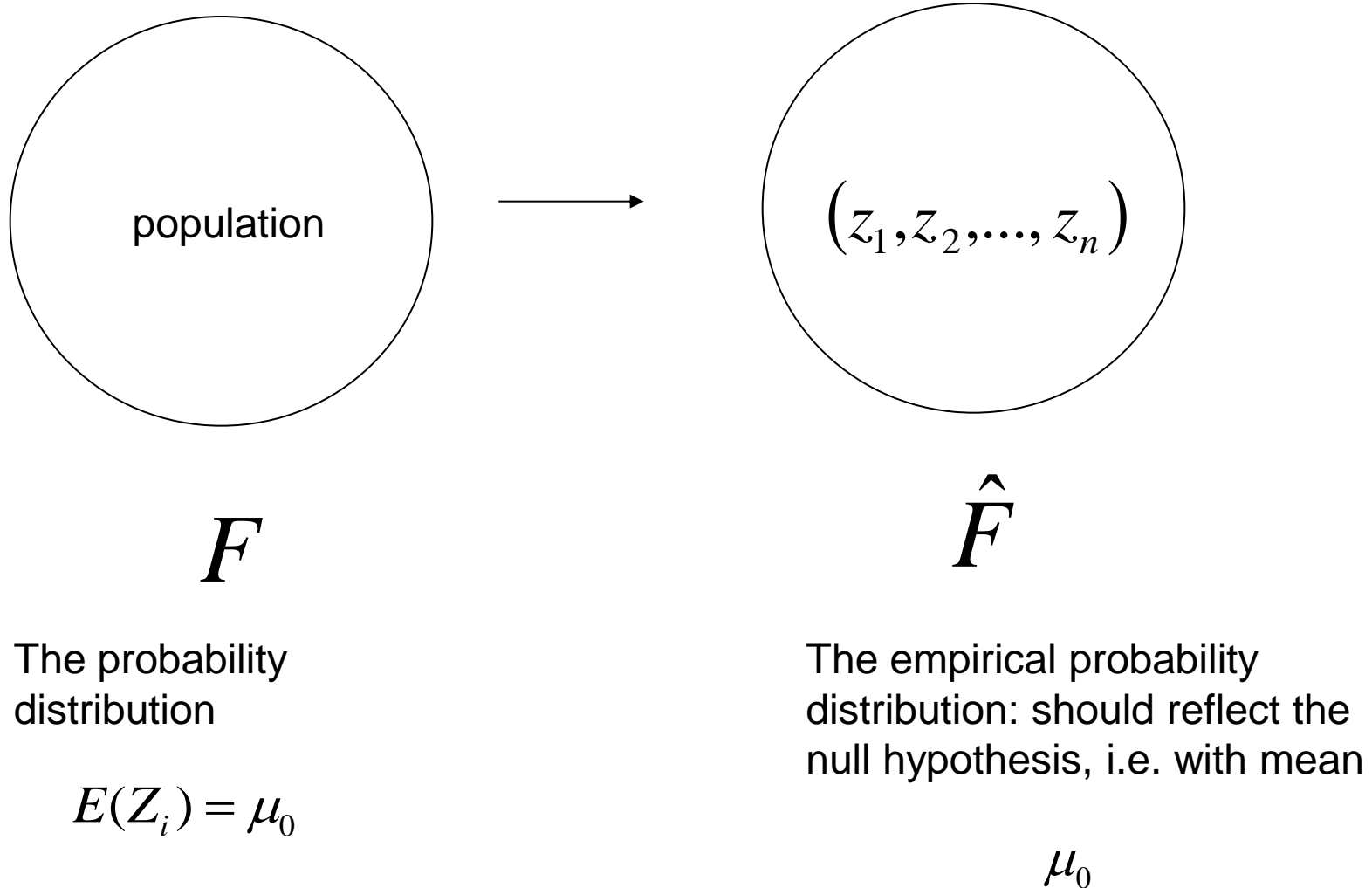
$$H_0 : \mu = 129$$

$$H_1 : \mu \neq 129$$

Assuming a normal population, the distribution of the the survival time and sample mean under the null hypothesis is given by

$$Z_i \sim N(129, \sigma^2)$$

$$\bar{Z}_i \sim N(129, \frac{\sigma^2}{7})$$

The empirical distribution under the null hypothesis

population

$\longrightarrow$

$(z_1, z_2, ..., z_n)$

F

$\hat{F}$

The probability distribution

The empirical probability distribution: should reflect the null hypothesis, i.e. with mean

$$E(Z_i) = \mu_0$$

$$\mu_0$$

Bootstrap test (1)

How can we generate the bootstrap sample under the null hypothesis ?

$$H_0 : \mu = 129$$

$$H_1 : \mu \neq 129$$

We define a new variable for which the mean is equal to 129

$$\tilde{Z}_i = Z_i - \bar{Z} + \mu_0$$

$$E(\tilde{Z}_i) = \mu_0$$

```
> mz <- mean(z)
> mz
[1] 86.85714
> n <- length(z)
> z.tilde <- z - mz + 129
> mean(z.tilde)
[1] 129
```

Two sided non parametric bootstrap t-test

Algorithm for the one sample problem

Simulate B samples of size n
with replacement from $\tilde{z}$

Calculate the test statistic values

$$t(z^*)_1, t(z^*)_2, ..., t(z^*)_B$$

Compute the Monte Carlo P-value

$$P = \frac{1 + \#\left\{\left|\hat{t}^*\right| > \left|t^{obs}\right|\right\}}{B + 1}$$

For two sided test

```
nz<-length(z)
t.obs<-t.test(z,mu=129)$statistic
B<-5000
t.boot<-c(1:B)
for(b in 1:B)
{
z.b<-sample(z.tilde,size=nz,replace=T)
t.boot[b]<-
    t.test(z.b,mu=129)$statistic
}
```

Bootstrap p value (two sided)

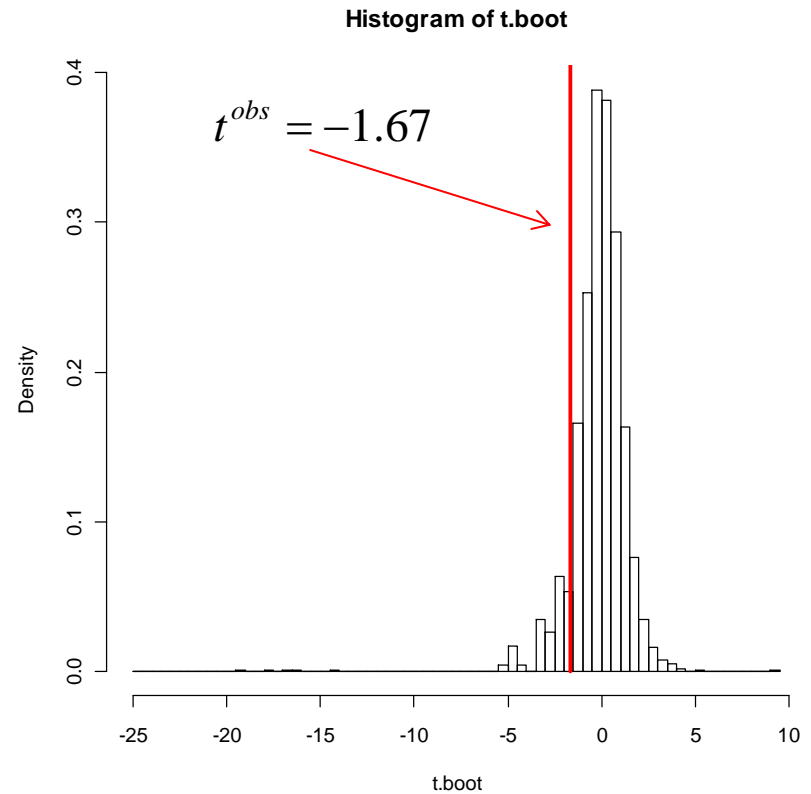$$P = \frac{1+\#\left\{\left|\hat{t}^*\right| > \left|t^{obs}\right|\right\}}{B+1}$$

> Pmc<-(1+sum(abs(t.boot) > abs(t.obs)))/(B+1)
> Pmc
[1] 0.1519696

```
> t.test(z,mu=129)

        One Sample t-test

data:  z
t=-1.67, df=6, p-value=0.1460
```

Histogram of t.boot

$t^{obs} = -1.67$

26

Two sided parametric bootstrap t-test

Algorithm for the one sample problem

Simulate B samples of size n from

$$F(\theta_0)$$

Calculate the test statistic values

$$t(z^*)_1, t(z^*)_2, ..., t(z^*)_B$$

Compute the Monte Carlo P-value

```
Theta.0<-129
Sig.0<-ssrt(var(z))
nz<-length(z)
t.obs<-t.test(z,mu=129)$statistic
B<-5000
t.boot<-c(1:B)
for(b in 1:B)
{
z.b<-rnorm(nz,theta.0,sig.0)

t.boot[b]<-t.test(z.b,mu=129)$statistic
}
```

$$N(\mu_0, \hat{\sigma}^2)$$

$$P = \frac{1 + \#\left\{ \left|\hat{t}^*\right| > \left|t^{obs}\right| \right\}}{B+1} \quad \Bigg\} \quad \text{For two sided test}$$

Example 2:
# two-sample problem

The two samples problem

- We would like to compare the distribution of a random variable in two populations (for example: EEG for two treatments in a clinical trial).

- Our main interest is to test if the distribution of the variable of primary interest is the same in the two populations.

Two populations – two distribution functions

$$H_0 : F = G$$

$$H_1 : F > G$$

We assume that F is a shift of the distribution G
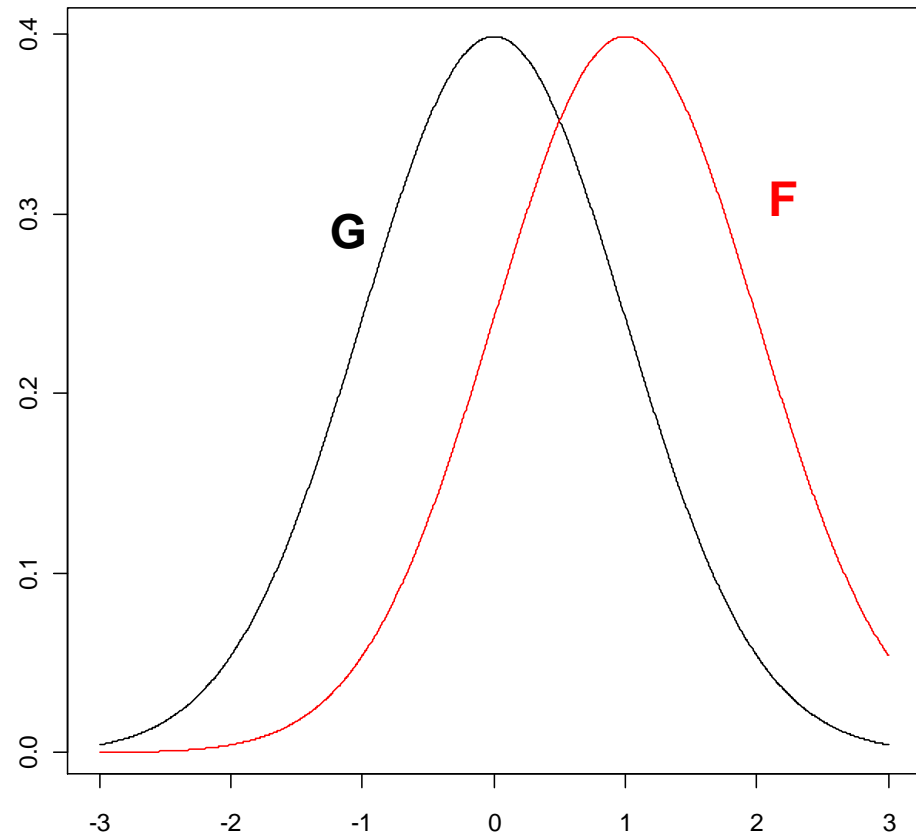
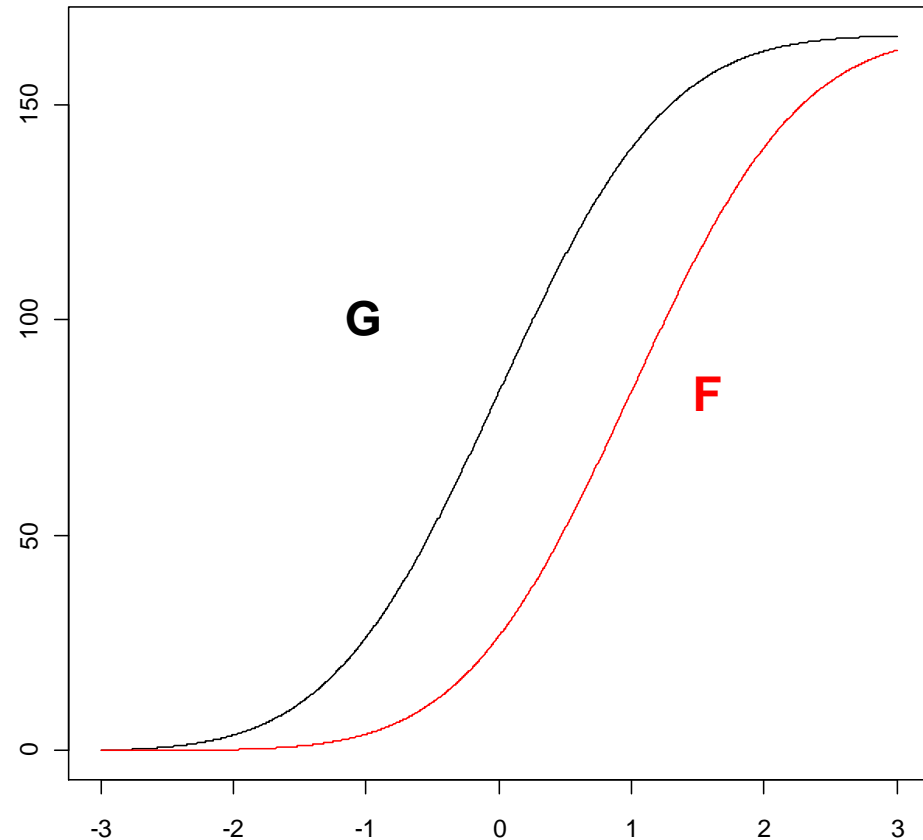# Two populations – two distribution functions
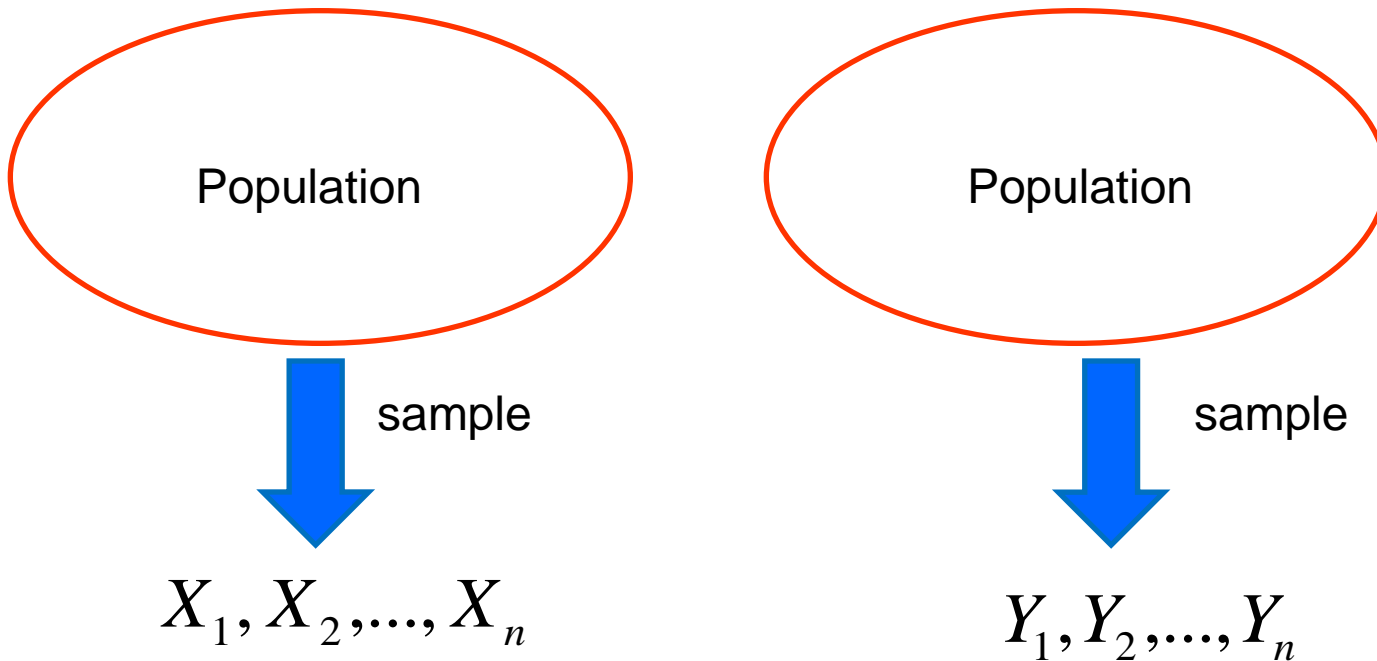
$H_0 : F = G$

$H_1 : F > G$

$H_0 : \mu_1 = \mu_2$
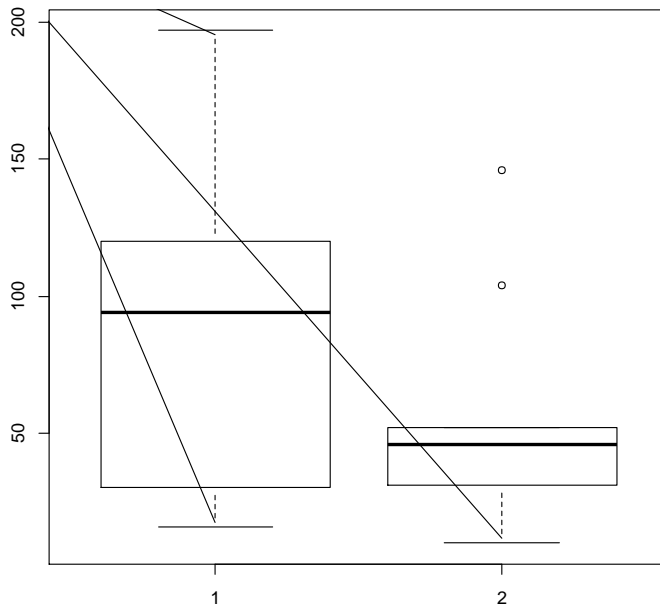
$H_1 : \mu_1 > \mu_2$

Populations and samples

Population

sample

$$X_1, X_2, ..., X_n$$

Population

sample

$$Y_1, Y_2, ..., Y_n$$

$$(Z_1, Z_2, ..., Z_n, Z_{n+1}, ..., Z_{n+m}) = (X_1, X_2, ..., X_n, Y_1, ..., Y_m)$$

The mouse data

- A randomized experiment with 16 mouse:
    - 7 to treatment group
    - 9 to control group.
- Treatment was intended to prolong survival after a test surgery.


```
> y <- c(10, 27, 31, 40, 46, 50, 52, 104, 146)
> x <- c(16, 23, 38, 94, 99, 141, 197)
> mean(y)
[1] 56.22222
> mean(x)
[1] 86.85714
```

Two samples problem

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

$$t = \frac{\bar{x} - \bar{y}}{\hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

$$t \sim t_{(n+m-2)}$$

```
> y
[1]    10    27    31    40    46
50   52  104  146
> x
[1]    16    23    38    94    99
141  197

> mean(y)
[1] 56.22222
> mean(x)
[1] 86.85714
```

Two-sample t test

```
> t.test(x,y,alternative="greater",var.equal=TRUE)

        Two Sample t-test

data:  x and y
t = 1.1214, df = 14, p-value = 0.1405
alternative hypothesis: true difference in means is greater than
   0
95 percent confidence interval:
 -17.48178       Inf
sample estimates:
mean of x mean of y
 86.85714  56.22222
```
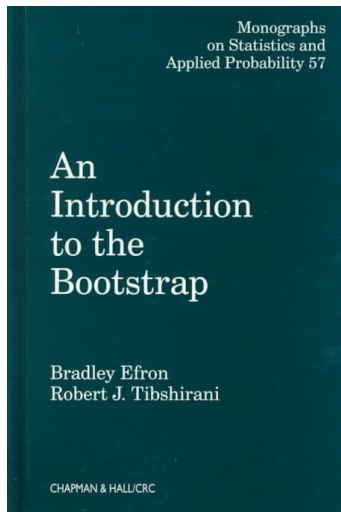
We do not reject the null hypothesis

Bootstrap algorithm 1

Monographs
on Statistics and
Applied Probability 57

An
Introduction
to the
Bootstrap

Bradley Efron
Robert J. Tibshirani

CHAPMAN & HALL/CRC

Algorithm 16.1

Algorithm 16.1

$$z = (y, x)$$

Simulate B samples of size m+ n with replacement from z.

The first n are considered as observations x* and the remaining m as observations y*.

Calculate the test statistic for each bootstrap sample

$$t(z^*)_1, t(z^*)_2, ..., t(z^*)_B$$

Two samples problem: test statistics I

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Test statistic

$$t = \bar{x} - \bar{y}$$

```
> t.obs <- mean(x)-mean(y)
> t.obs
[1] 30.63492
```
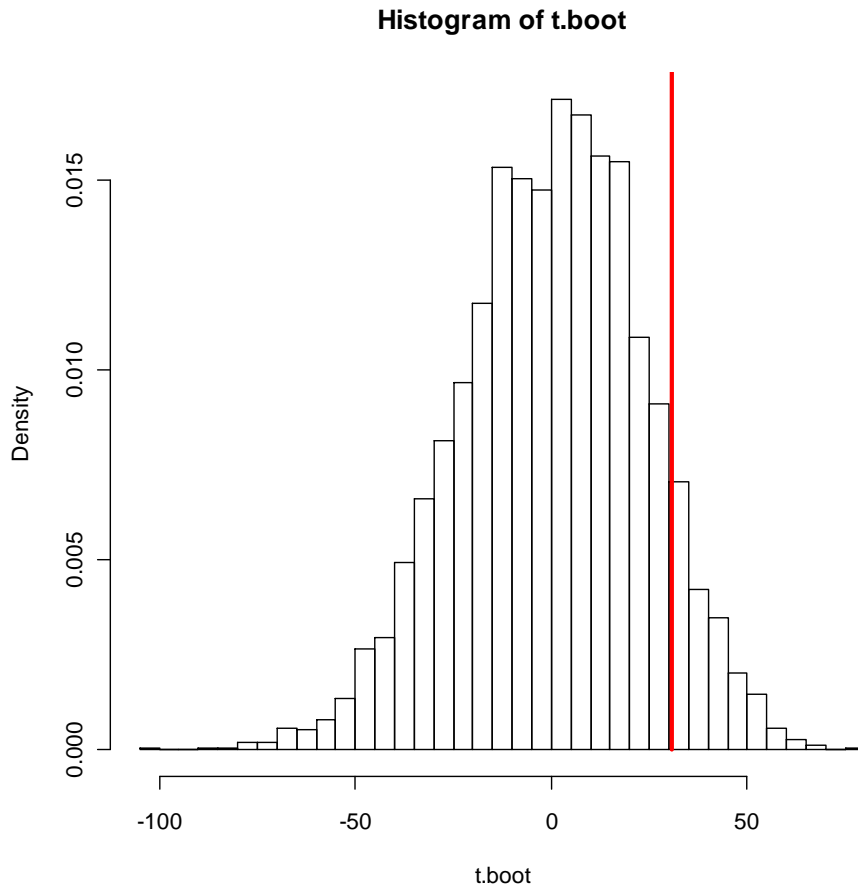
Simulate B samples of size m+ n with replacement from z.

The first n are considered as observations x* and the remaining m as observations y*.

Calculate the test statistic for each bootstrap sample

$$t(z^*)_1, t(z^*)_2, ..., t(z^*)_B$$

Two-sample bootstrap t test (one sided)

Histogram of t.boot


```
z <- c(x, y)
m <- length(x)
n <- length(y)
mn <- m + n
t.obs <- mean(x) - mean(y)
t.obs
B<-5000
t.boot<-c(1:B)
for(b in 1:B)
{
z.b<-sample(z,size=mn,replace=T)
x.b<-z.b[1:n]
y.b<-z.b[(n+1):mn]   ⬅
t.boot[b]<-mean(x.b)-mean(y.b)
}
hist(t.boot,nclass=50,probability=T)
lines(c(t.obs,t.obs),c(0,1),lwd=3,col=2)
```

```
> (sum(t.boot>t.obs)+1)/(B+1)
[1] 0.09318136
```

Two samples problem: test statistic II

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

What is the problem with the previous test statistic ?

$$t = \frac{\bar{x} - \bar{y}}{\hat{\sigma} \sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

$$t \sim t_{(n+m-2)}$$

Assuming
equal variance

```
> t.test(x,y,alternative="greater")

        Welch Two Sample t-test

data:  x and y
t = 1.0591, df = 9.645, p-value = 0.1577
alternative hypothesis: true difference
in means is greater than 0
95 percent confidence interval:
 -21.98945        Inf
sample estimates:
mean of x mean of y
 86.85714  56.22222
```
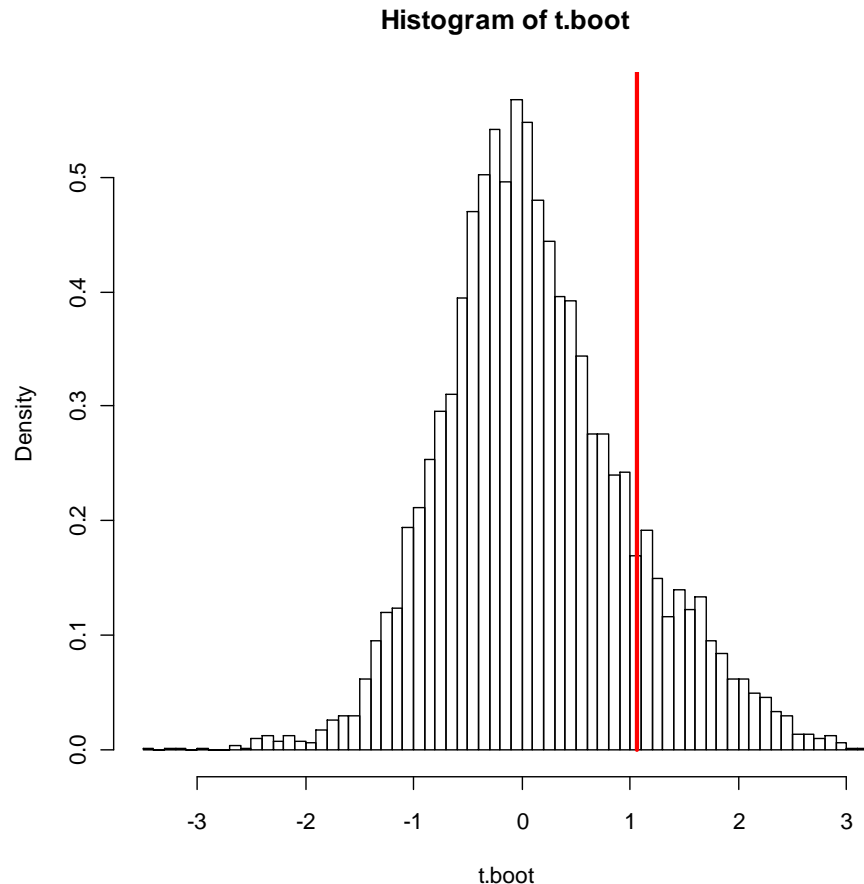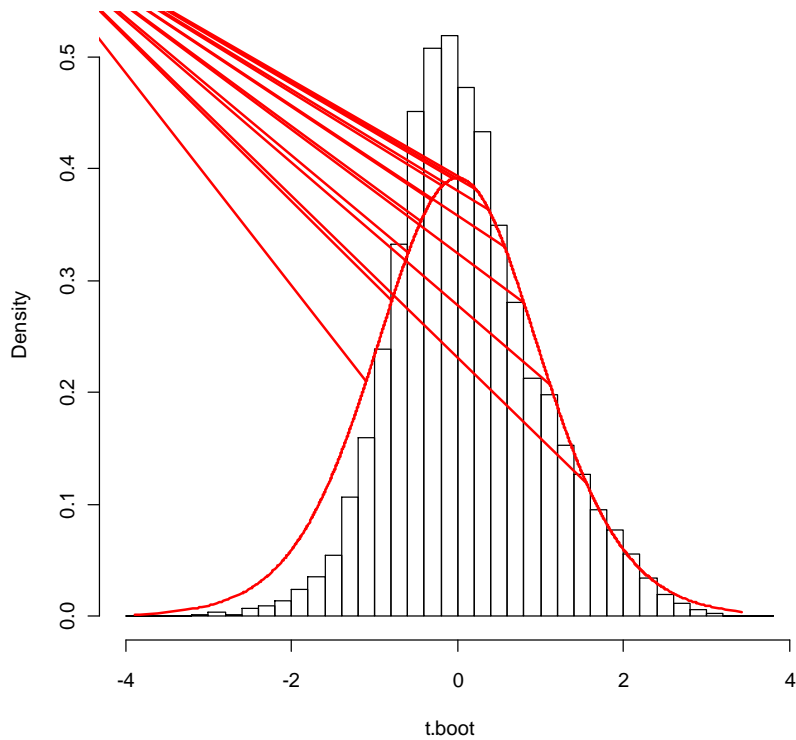
40

Two-samples bootstrap t test (one sided)

Histogram of t.boot


```
> B<-5000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ z.b<-sample(z,size=mn,replace=T)
+ x.b<-z.b[1:n]
+ y.b<-z.b[(n+1):mn]
+ t.boot[b]<-t.test(x.b,y.b)$statistic
+ }
> hist(t.boot,nclass=50,probability=T)
> lines(c(t.obs,t.obs),c(0,1),lwd=3,col=2)
```
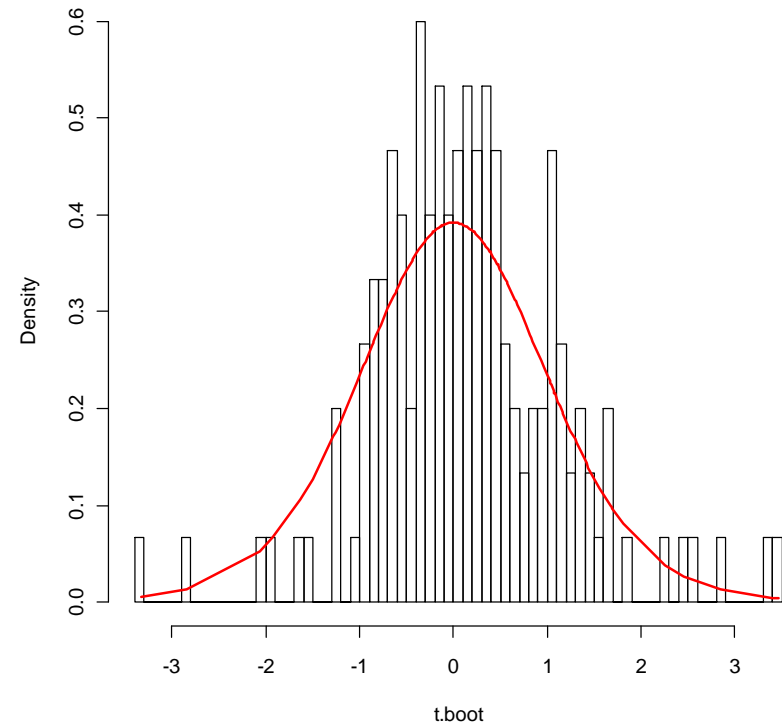
```
> Pmc<-(1+sum(t.boot > t.obs))/(B+1)
> Pmc
[1] 0.1445711
```

How many bootstrap samples ?
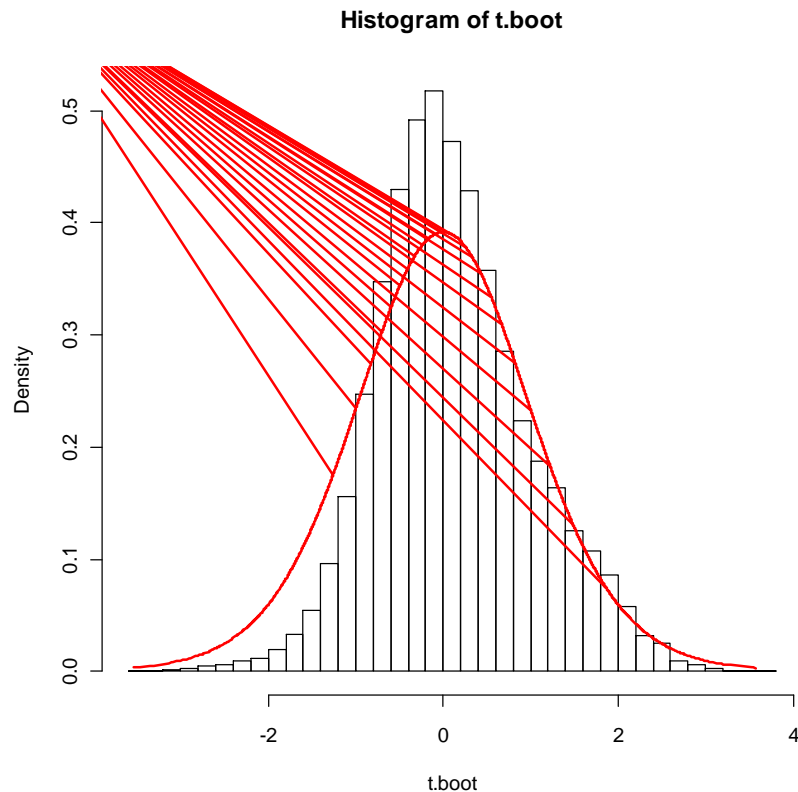

Histogram of t.boot

Histogram of t.boot

B=15000

B=150

B=25000

Histogram of t.boot

Two Sample t-test

data: x and y

t = 1.1214, df = 14, p-value = 0.1405

alternative hypothesis: true
   difference in means is greater than
   0

95 percent confidence interval:

 -17.48178   Inf
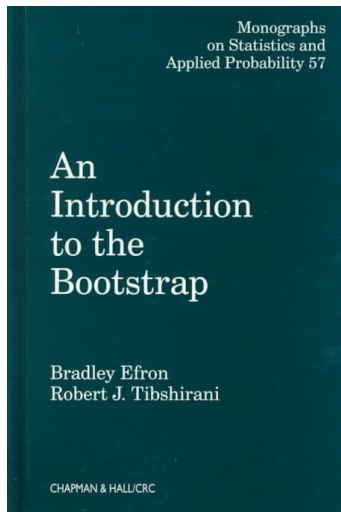
sample estimates:

mean of x mean of y

 86.85714 56.22222

> Pmc<-(1+sum(t.boot > t.obs))/(B+1)
> Pmc
[1] 0.1484341

43

Bootstrap algorithm 2

Monographs
on Statistics and
Applied Probability 57

An
Introduction
to the
Bootstrap

Bradley Efron
Robert J. Tibshirani

CHAPMAN & HALL/CRC

Algorithm 16.2

Two samples problem: test statistics I

$$z = (y, x)$$

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$$

$$\tilde{x}_i = x_i - \bar{x} + \bar{z} \quad \Leftarrow$$

$$\tilde{y}_i = y_i - \bar{y} + \bar{z} \quad \Leftarrow$$

$$t^* = \frac{\bar{x}^* - \bar{y}^*}{\hat{\sigma}^* \sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

Simulate B samples with replacement of size n form x-tilde and size m from y-tilde.

Calculate the test statistic for each bootstrap sample:

$$t(z^*)_1, t(z^*)_2, ..., t(z^*)_B$$

Non parametric bootstrap (algorithm 16.2)

```
> y <- c(10, 27, 31, 40, 46, 50, 52, 104, 146)
> x <- c(16, 23, 38, 94, 99, 141, 197)
> z<-c(x,y)
> m <- length(x)
> n <- length(y)
> my<-mean(y)
> mx<-mean(x)
> mz<-mean(z)
> x.tilde<-x-mx+mz
> y.tilde<-y-my+mz
> mean(x.tilde)
[1] 69.625
> mean(y.tilde)
[1] 69.625
```
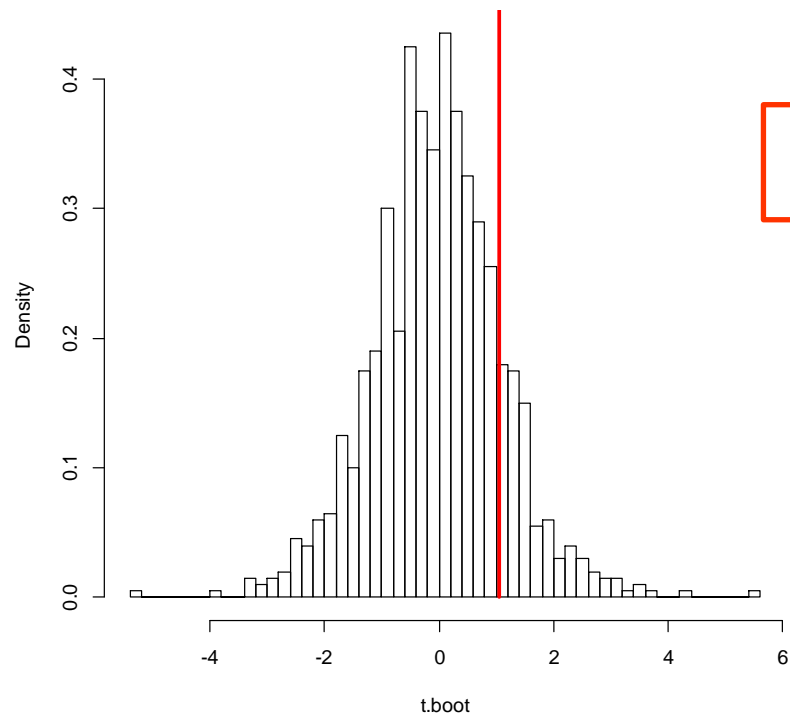
$$\tilde{x}_i = x_i - \bar{x} + \bar{z}$$

$$\tilde{y}_i = y_i - \bar{y} + \bar{z}$$

Two-samples bootstrap t test (one sided)

Histogram of t.boot

```
> B<-1000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+
+ x.b<-sample(x.tilde,m,replace=TRUE)
+ y.b<-sample(y.tilde,n,replace=TRUE)
+ t.boot[b]<-t.test(x.b,y.b)$statistic
+ }
> hist(t.boot,nclass=50,probability=T)
> lines(c(t.obs,t.obs),c(0,1),lwd=3,col=2)
> Pmc<-(1+sum(t.boot > t.obs))/(B+1)
> Pmc
[1] 0.1508492
```

Example 3: Two populations problem: an example of a ratio test

- Population of the most populated city in US in 1920 and 1930.

- Main question: do we see a growth in the population from 1920 to 1930 ?
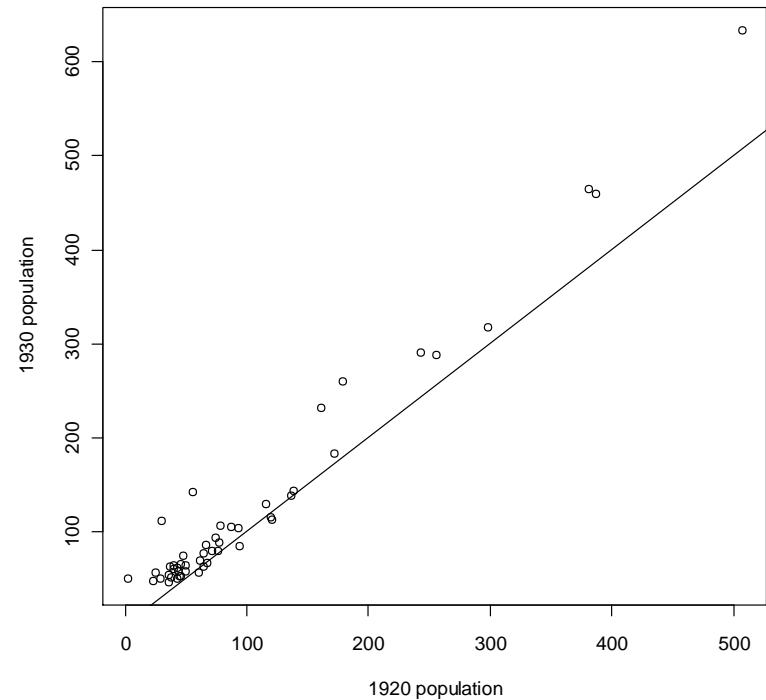
External dataset:

```
city<-read.table('C:/projects/cim/UpdatesSlides_2017/Data/city.txt',
header=FALSE,na.strings=".", dec=".",  strip.white=TRUE,
            col.names=c("pop1920","pop1930"))
```

Ratio test statistic

$U_i = pop.$ 1920

$X_i = pop.$ 1930

Parameter:

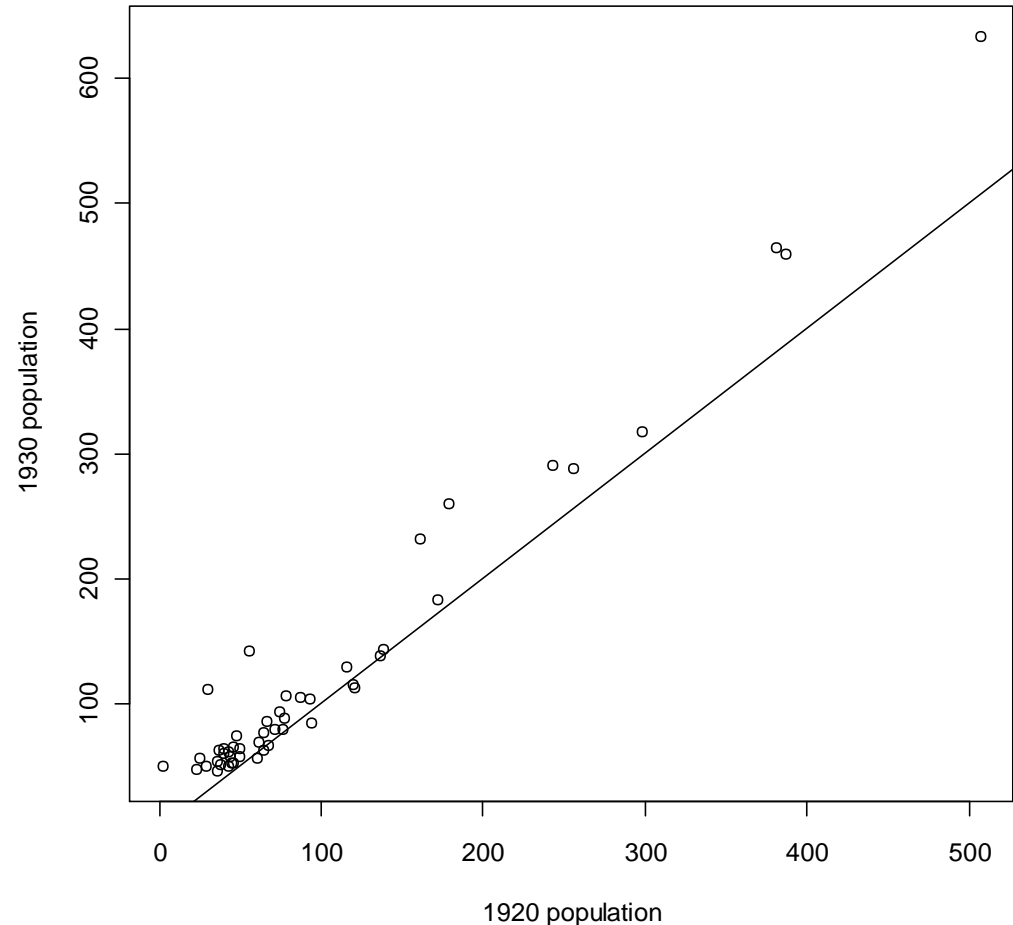$$\theta = \frac{E(X)}{E(U)} = \frac{\mu_{1930}}{\mu_{1920}}$$

Statistic:

$$t = \frac{\bar{x}}{\bar{u}}$$

```
> t.obs<-mean(x)/mean(u)
> t.obs
[1] 1.239019
```
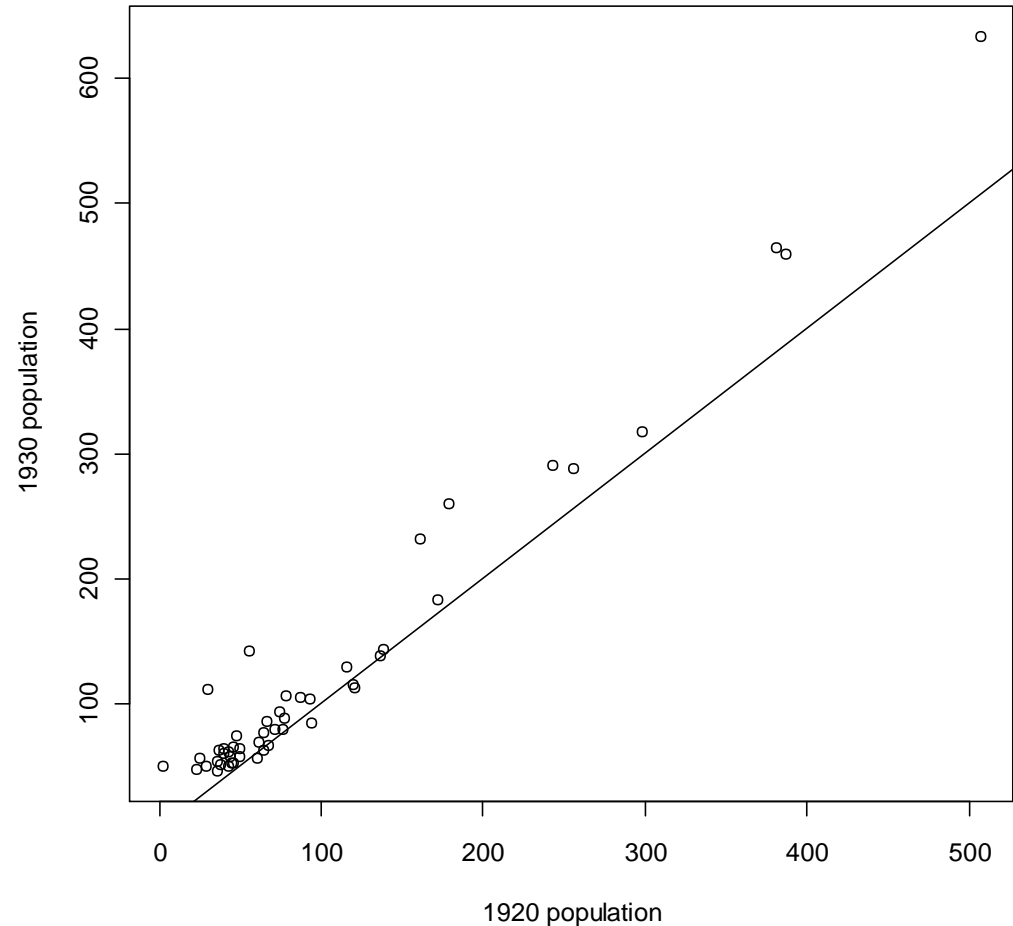
Ratio test statistic

The observed ratio is equal to 1.239019.

Formulation of the hypotheses:

$$H_0 : \frac{\mu_{1920}}{\mu_{1930}} = 1$$

$$H_1 : \frac{\mu_{1920}}{\mu_{1930}} < 1$$

$$t = \frac{\overline{u}}{\overline{x}}$$

50

Bootstrap algorithm

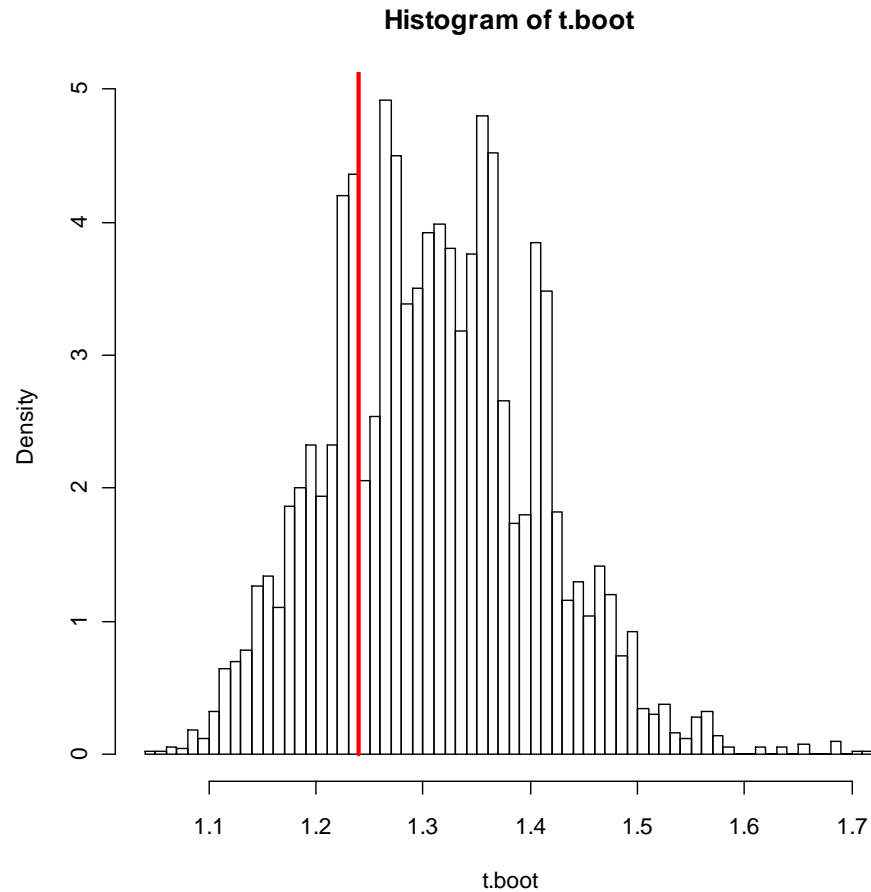- Bootstrap algorithm.
- Bootstrap pairs (why ?).

```
> B-5000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ index<-sample(c(1:n),size=n,replace=T)
+ u.boot<-u[index]
+ x.boot<-x[index]
+ t.boot[b]<-mean(x.boot)/mean(u.boot)
+ }
```

Bootstrap p-value

Histogram of t.boot

$$P = \frac{1 + \#\left\{t_b^* > t^{obs}\right\}}{B + 1}$$

```
> (1+sum(t.boot > t.obs))/(B+1)
[1] 0.7476505
```

What is wrong here?

City population: paired t-test

$$Z_i = U_i - X_i = pop._{1920} - pop._{1930}$$

Under the null hypothesis:

$$E(Z_i) = E(U_i - X_i) = 0$$

$$H_0 : \frac{\mu_{1920}}{\mu_{1930}} = 1$$

$$H_1 : \frac{\mu_{1920}}{\mu_{1930}} < 1$$

$$H_0 : \mu_z = 0$$

$$H_1 : \mu_z < 0$$

Each city was measured twice (1920 & 1930).

Under the null hypothesis (no growth) the mean of the difference is expected to be zero.

Resampling under the null

How can we generate the bootstrap sample under the null hypothesis ?

$$H_0 : \mu_Z = 0$$

$$H_0 : \mu_Z > 0$$

$$\tilde{Z}_i = Z_i - \overline{Z} + \mu_0$$

$$E(\tilde{Z}_i) = \mu_0$$

```
> z<-u-x
> mean(z)
[1] -24.65306
> z.tilde<-z-mean(z)
> mean(z.tilde)
[1] 5.794596e-16
```

Two samples paired t-test (two sided)

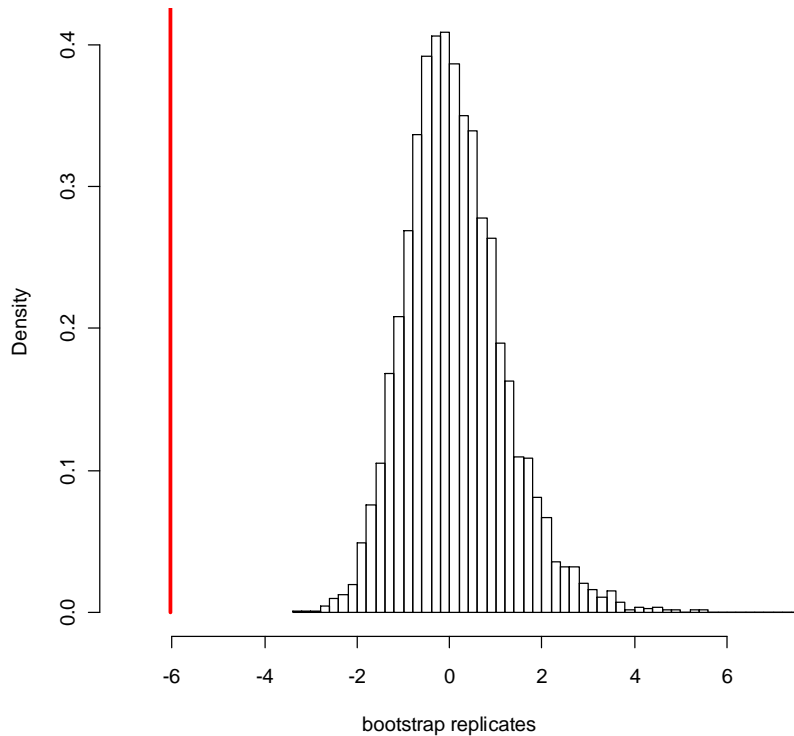```
> t.test(z,mu=0)

        One Sample t-test

data: z
t = -6.0276, df = 48, p-value = 2.275e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -32.87662 -16.42950
sample estimates:
mean of x
-24.65306
```

We reject the null hypothesis

# Bootstrap algorithm



Density

bootstrap replicates

> Pmc<-(1+sum(abs(t.boot) > abs(t.obs)))/(B+1)
> Pmc
[1] 0.00059988

What is the bootstrap algorithm of this problem ?

How can we make sure that we bootstrap under the null hypothesis ?

```
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ z.b<-
    sample(z.tilde,size=nz,replace=T)
+ t.boot[b]<-
    t.test(z.b,mu=0)$statistic
+ }
>
```

How can we test the original null hypothesis ?

Recall that the original hypotheses were formulated in terms of the ratios.

$$H_0 : \frac{\mu_{1920}}{\mu_{1930}} = 1$$

$$H_1 : \frac{\mu_{1920}}{\mu_{1930}} < 1$$

Test statistic:

$$t = \frac{\bar{x}}{\bar{u}}$$

Bootstrapping under the null ?
Homework 3

Example 3: correlation test

The cars data

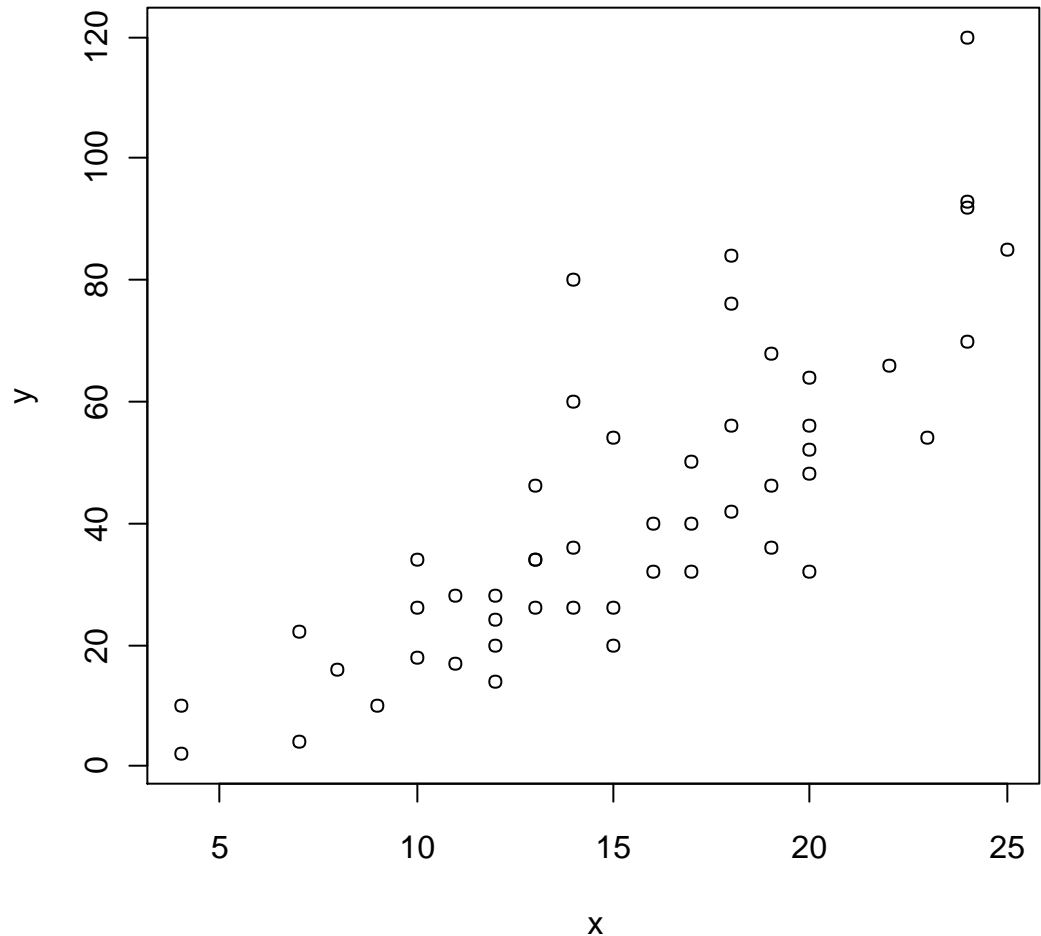The data give the speed of 50 cars and the distances taken to stop.

The data were recorded in the 1920s.

In R:

```
> help(cars)
```

Observed correlation

```
> x<-cars$speed
> y<-cars$dist
> cor(x,y)
[1] 0.8068949
```
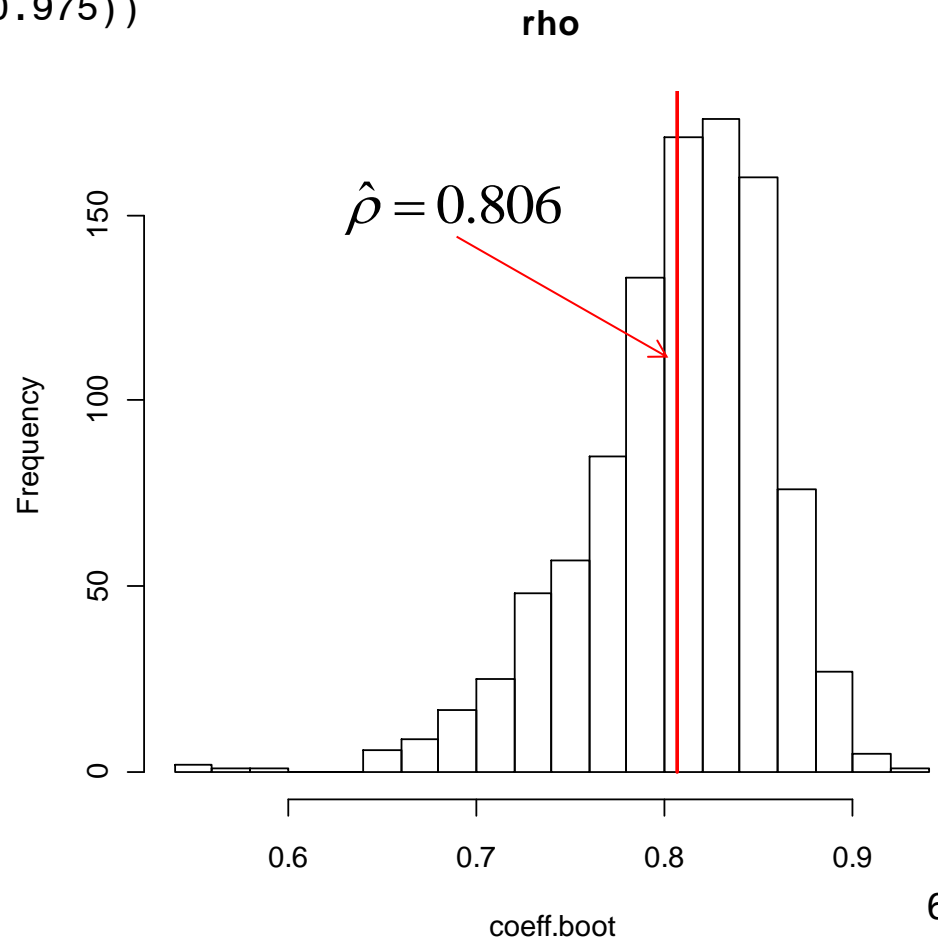$\to \hat{\rho}$

Bootstrap C.I. (non parametric)

```
> quantile(coeff.boot,probs=c(0.025,0.975))
     2.5%      97.5%
0.6906236 0.8831039
```

Non parametric bootstrap algorithm.

Resampling pairs.

rho

$$\hat{\rho} = 0.806$$

Frequency

coeff.boot

60

The hypotheses

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Under the null hypothesis the speed and stopping distance are not correlated.

This should be reflected in the bootstrap algorithm.

Non parametric bootstrap algorithm: generate a bootstrap sample under the null hypothesis

The observed sample

$x_1 \quad y_1$

$x_2 \quad y_2$

. .

. .

. .

$x_{50} \quad y_{50}$

We resample xi and yi with replacement

n=50

$x^*_1 \quad y^*_1$

$x^*_2 \quad y^*_2$

. .   The bootstrap sample

. .

. .

$x^*_{50} \quad y^*_{50}$

The bootstrap algorithm

The bootstrap sample

For each bootstrap sample we calculate the correlation

$$x^*_1 \qquad y^*_1$$

$$x^*_2 \qquad y^*_2$$

$$\cdot \qquad \cdot \longrightarrow \hat{\rho}^*_b(x^*, y^*)$$

$$\cdot \qquad \cdot$$

$$\cdot \qquad \cdot$$

$$x^*_{50} \qquad y^*_{50}$$

Bootstrap p value (two sided)

$$P = \frac{1 + \#\left\{\left|\hat{\rho}^*\right| > \left|\hat{\rho}^{obs}\right|\right\}}{B + 1}$$

The bootstrap distribution of the test statistics **rho**

$\hat{\rho} = 0.806$

64
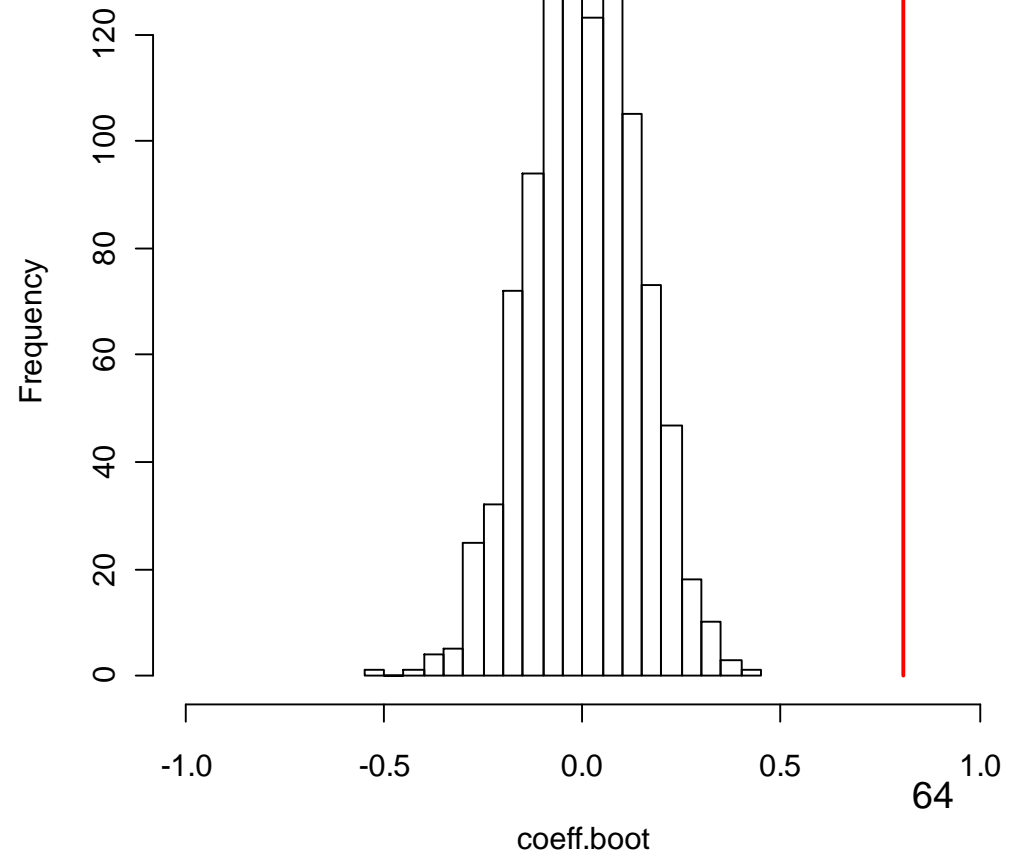
R code for the C.I. and bootstrap test: non parametric bootstrap

Bootstrap confidence intervals

```
n<-length(x)
B<-1000
coeff.boot<-c(1:B)
index<-c(1:n)
for(b in 1:B)
{
index.b<- sample(index,size=n,replace=TRUE)
x.boot<- x[index.b]
y.boot<- y[index.b]
coeff.boot[b]<-cor(x.boot,y.boot)
}
```

Bootstrap test

```
n<-length(x)
B<-1000
coeff.boot<-c(1:B)
index<-c(1:n)
for(b in 1:B)
{
x.boot<- sample(x,size=n,replace=TRUE)
y.boot<- sample(y,size=n,replace=TRUE)
coeff.boot[b]<-cor(x.boot,y.boot)
}
```

Permutation tests

Monographs
on Statistics and
Applied Probability 57

An
Introduction
to the
Bootstrap

Bradley Efron
Robert J. Tibshirani

CHAPMAN & HALL/CRC

Chapter 15
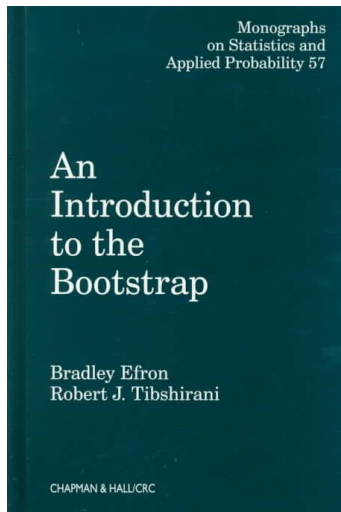
Topics

- Permutation tests for the two samples problems.

- Bootstrap without replacement and permutation tests.

Two populations problem

$F \rightarrow \quad z = (z_1, z_2, ..., z_n)$

$G \rightarrow \quad y = (y_1, y_2, ..., y_m)$

Shift model:

$$H_0 : F = G$$
$$H_1 : F > G$$
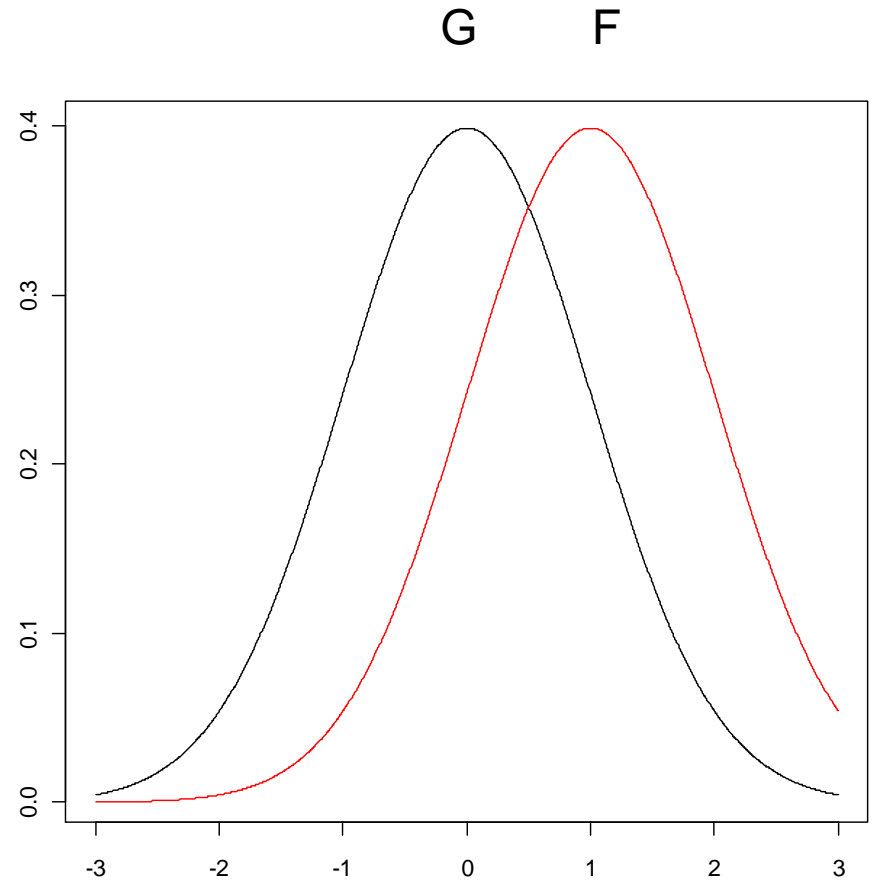
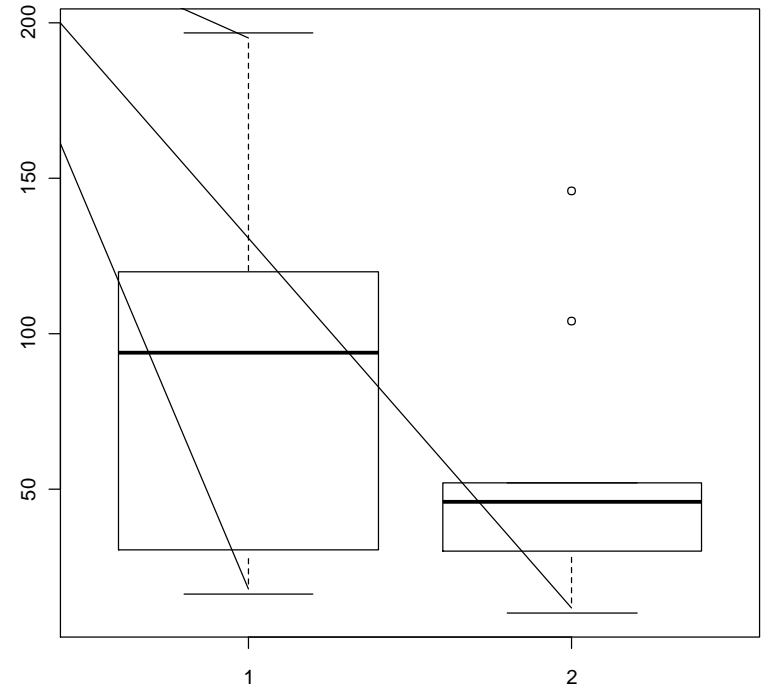# The mouse data

```
> 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.

Hypotheses and test statistic

$$H_0 : F = G$$
$$H_1 : F > G$$

$$\Rightarrow$$

$$H_0 : F(\mu_F) = G(\mu_G)$$
$$H_1 : F(\mu_F) > G(\mu_G)$$

$$\Rightarrow$$

$$H_0 : \mu_F = \mu_G$$
$$H_1 : \mu_F > \mu_G$$

$$\hat{\theta} = \bar{z} - \bar{y}$$

$$\hat{\theta} \sim N\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

Standard error is unknown
We use the parameter estimate for the standard error

$$\frac{\hat{\theta}}{se(\hat{\theta})} = \frac{\hat{\theta}}{\sigma(\hat{\theta})} \sim N(0,1)$$

$$\Rightarrow$$

$$\frac{\hat{\theta}}{\hat{\sigma}(\hat{\theta})} \sim t_{n+m-2}$$

The p value (achieved significance level)

$$ASL = P_{H_0}(\theta \geq \hat{\theta})$$

$$ASL = P_{H_0}\left(t_{14} \geq \frac{\hat{\theta}}{\hat{\sigma}(\hat{\theta})} \right) = P_{H_0}\left(t_{14} \geq \frac{30.63}{54.21\sqrt{\frac{1}{9} + \frac{1}{7}}} \right) = 0.141$$

The p value (achieved significance level)

- The distribution of the test statistic under the null.

- The p value:

$$P_{H_0}\left(t \geq t^{obs}\right) = P\left(\theta \geq \hat{\theta}\right)$$
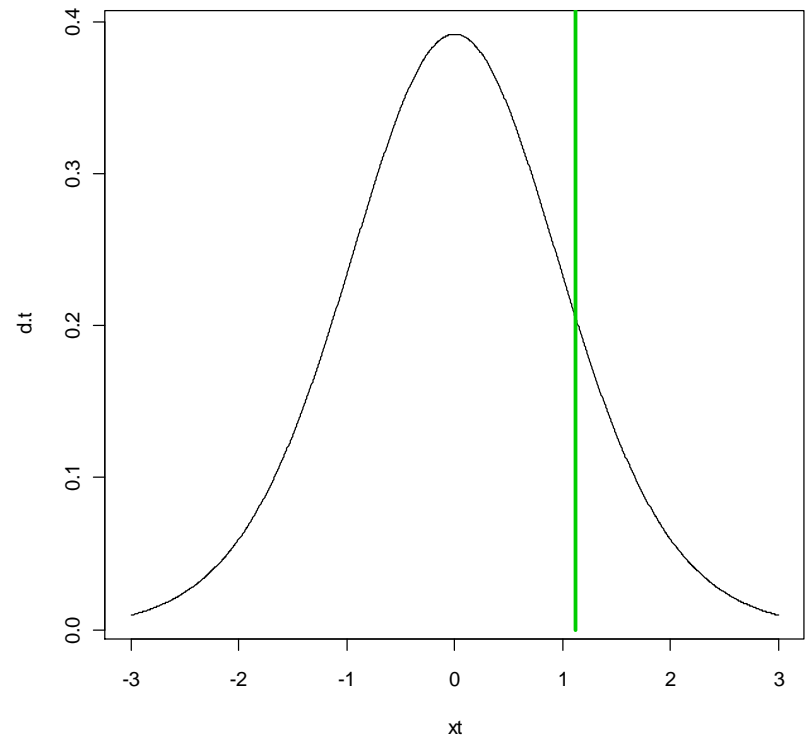
Two samples problem: one sided test

```
> t.test(z,y,alternative="greater")
```

Welch Two Sample t-test $\longrightarrow \sigma_1 \neq \sigma_2$

```
data:  z and y
t = 1.0587, df = 9.654, p-value = 0.1578
alternative hypothesis: true
difference in means is greater than 0
95 percent confidence interval: -22.00168
    Inf
sample estimates:
mean of x mean of y
 86.85714  56.22222
```
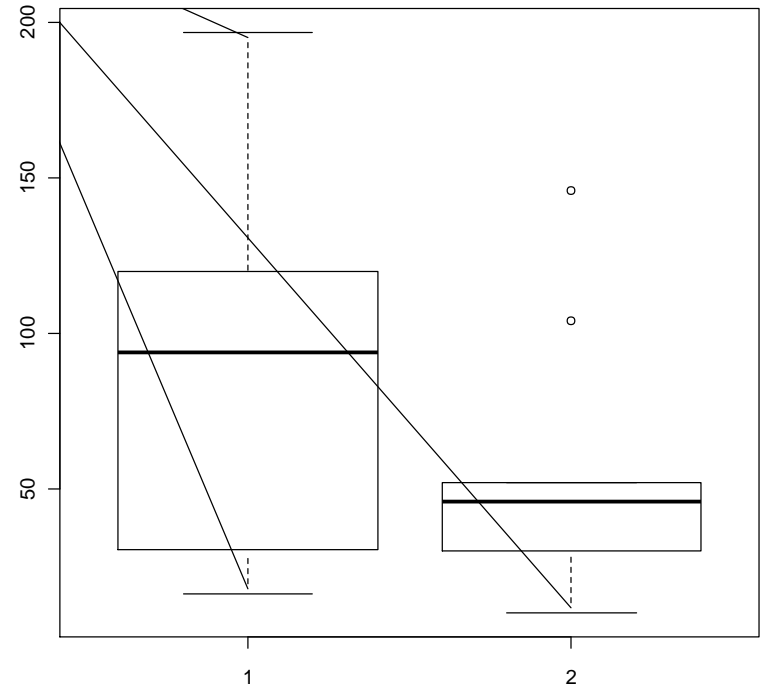
Permutations

- We combine the two samples.

- Under the null hypothesis the survival time of any mice could come equally from either two groups.

```
> data.frame(sort(x),r.x[order(x)])
   sort.x.  r.x.order.x..
```

	sort.x.	r.x.order.x..
1	10	y
2	16	z
3	23	z
4	27	y
5	30	y
6	38	z
7	40	y
8	46	y
9	51	y
10	52	y
11	94	z
12	99	z
13	104	y
14	141	z
15	146	y
16	197	z

Permutations

We define two vectors:

Response (survival time)

$$v = (v_1, v_2, ..., v_N)$$

$$v_i = \begin{cases} y_i & observation \in G \\ z_i & observation \in F \end{cases}$$

Membership

$$g = (g_1, g_2, ..., g_N)$$

$$g_i = \begin{cases} y & observation \in G \\ z & observation \in F \end{cases}$$

Permutations

Under $H_0\text{:}F{=}G$, the vector g has equal probability, π, of being equal to one of its possible values.

$$\pi \;=\; \frac{1}{\dbinom{N}{n}}$$

n: sample size from F

Under $H_0\text{:}F{=}G$, each permutation of the data has the same probability (all permutations of z's and y's are equally likely)

Test Statistic

$$\hat{\theta} = S(g, v)$$

$$\hat{\theta} = \frac{1}{n} \sum_{g_i = z} v_i - \frac{1}{m} \sum_{g_i = y} v_i$$

The test statistic is a function of the two vectors.

The mean difference.

Permutations

$$\hat{\theta}^* = S(g^*, v)$$

$$ALS = \frac{\#\left(\hat{\theta}^* \geq \hat{\theta}\right)}{\binom{N}{n}}$$

If we do all possible permutations than the p value is the number of time that the test statistic (in the permutation) is grater or equal to the observed test statistic divided by the number of all possible permutations.

What is the problem here ?

Permutations

$$\hat{\theta}^*(b) = S(g^*(b), v)$$

We draw B samples without replacement (why ?) from the vector v.

$$ALS = \frac{\#\left(\hat{\theta}^*(b) \geq \hat{\theta}\right)}{B}$$

The first n are assigned to the first group the last m are assigned to the second group.

Permutations

Histogram of t.boot
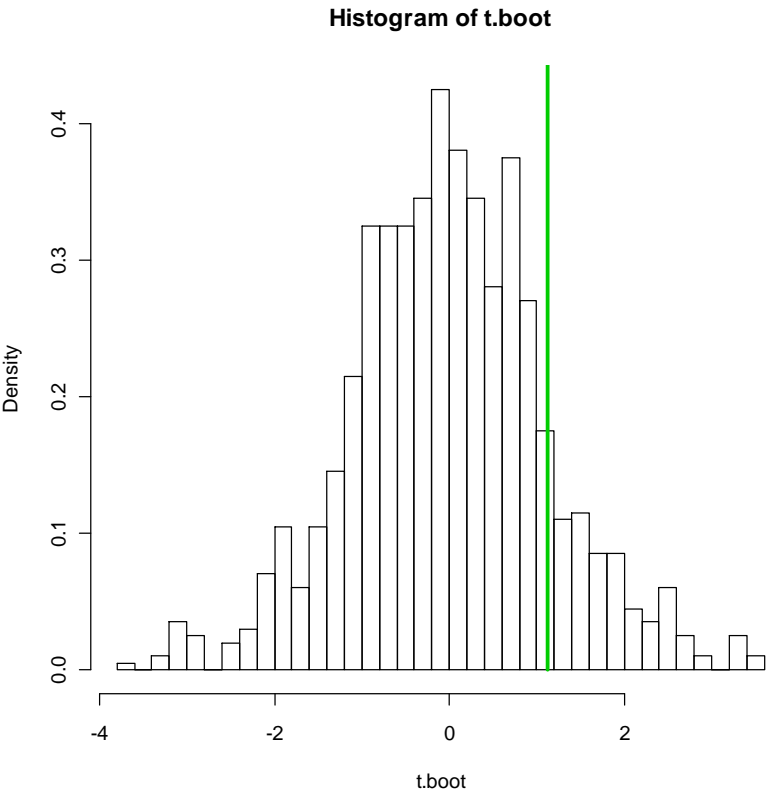

```
> x<-c(z,y)
> B<-1000
> t.boot<-c(1:B)
> for(b in 1:B)
+ {
+ x.boot<-sample(x,size=n+m,replace=FALSE)
+ z.boot<-x.boot[1:n]
+ y.boot<-x.boot[(n+1):(n+m)]
+ t.boot[b]<-
    t.test(z.boot,y.boot,alternative="greater",
        var.equal=TRUE)$statistic
+ }
> hist(t.boot,nclass=50,probability=TRUE)
> lines(c(t.obs,t.obs),c(0,10),col=3,lwd=3)
```

```
> (1+sum(t.boot>t.obs))/(B+1)
[1] 0.1348651
```

Extra example: binary data

Two populations and two proportions

Population 1

$$X_i = \begin{cases} 1 & \pi_1 \\ 0 & 1 - \pi_1 \end{cases}$$

Population 2

$$Y_i = \begin{cases} 1 & \pi_2 \\ 0 & 1 - \pi_2 \end{cases}$$

Sample 1

Sample 2

$$X_1, X_2, ..., X_{n_1}$$

$$Y_1, Y_2, ..., Y_{n_2}$$

The testing problem

The null hypothesis that we want to test

$$H_0 : \pi_2 - \pi_1 = 0$$

$$H_1 : \pi_2 - \pi_1 \neq 0$$

The sample proportions

- Suppose S_1 the number of successes in the first sample and S_2 the number of successes in the second sample.
- The proportions of the sample are then given by

$$\hat{\pi}_1 = \overline{P}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{the\ number\ of\ "1s"\ in\ sample_1}{n_1}$$

$$\hat{\pi}_1 = \overline{P}_1 = \frac{S_1}{n_1} \qquad\qquad \hat{\pi}_2 = \overline{P}_2 = \frac{S_2}{n_2}$$

The variance of the samples means

$X_1, X_2, ..., X_{n_1}$

a sample from the first population

$Y_1, Y_2, ..., Y_{n_2}$

a sample from the second population

$$E(\overline{P_1}) = \pi_1$$

$$E(\overline{P_2}) = \pi_2$$

$$Var(\overline{P_1}) = \frac{\pi_1(1-\pi_1)}{n_1}$$

$$Var(\overline{P_2}) = \frac{\pi_2(1-\pi_2)}{n_2}$$

The expected value and the variance of the difference

$$E(\overline{P}_2 - \overline{P}_1) = \pi_2 - \pi_1$$

$$Var(\overline{P}_1 - \overline{P}_2) = Var(\overline{P}_2) + Var(\overline{P}_1) = \frac{\pi_2(1-\pi_2)}{n_2} + \frac{\pi_1(1-\pi_1)}{n_1}$$

The proportion under H_0

The null hypothesis says that $\pi_1=\pi_2=\pi$ (for some unknown value π).

Under H_0 :

$$E(\overline{P}_2 - \overline{P}_1) = 0$$

$$Var(\overline{P}_2 - \overline{P}_1) = Var(\overline{P}_2) + Var(\overline{P}_1) = \pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$H_0 : \pi_2 - \pi_1 = 0 \implies \pi_2 - \pi_1 = \pi$$

The pooled sample proportion

Problem is that π is not known. An estimator for π is given by the pooled sample proportion

$$\overline{P} = \frac{n_1 \overline{P_1} + n_2 \overline{P_2}}{n_1 + n_2} = \frac{S_1 + S_2}{n_1 + n_2}$$

The test statistic

$$\frac{\overline{P}_2 - \overline{P}_1 - (p_2 - p_1)_{H_0}}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\overline{P}_2 - \overline{P}_1}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx N(0,1)$$

unknown parameter

$$\frac{\overline{P}_2 - \overline{P}_1}{\sqrt{\overline{P}(1-\overline{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Example:
Agresti – An Introduction to Categorical Data Analysis

Aspirin use and heart attacks.

Example: aspirin use and myocardial infarction (MI)

Table 2.3. Cross Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262–264, 1988.

Relationship between aspirin use and myocardial infarction (heart attacks).

A five-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease.

Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo.

Population 1

$$X_i = \begin{cases} 1 & \pi_1 \\ 0 & 1 - \pi_1 \end{cases}$$

Population 2

$$Y_i = \begin{cases} 1 & \pi_2 \\ 0 & 1 - \pi_2 \end{cases}$$

$$\pi_\ell = P(heart \quad attack), \quad \ell = Placebo, Aspirin$$

Example: aspirin use and heart attacks.

Table 2.3. Cross Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262–264, 1988.

Main interest:

$$H_0 : \pi_{Placebo} - \pi_{Aspirin} = 0$$

```
> P.Placebo<-189/(189+10845)
> P.Placebo
[1] 0.01712887
> P.Aspirin<-104/(104+10933)
> P.Aspirin
[1] 0.00942285
> P.Placebo-P.Aspirin
[1] 0.007706024
```

$\hat{\pi}_1 = \hat{P}(heart \quad attack \mid Placebo)$

$\hat{\pi}_2 = \hat{P}(heart \quad attack \mid Aspirin)$

$\hat{\pi}_1 - \hat{\pi}_2$

Proportion under the null and the test statistic

Table 2.3. Cross Classification of Aspirin Use and Myocardial Infarction

Group	Myocardial Infarction		Total
	Yes	No	
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.,* 318: 262–264, 1988.

```
> P.tot<-(189+104)/(189+10845+104+10933)
> P.tot
[1] 0.01327534
```

$$\overline{P} = \frac{n_1 \overline{P_1} + n_2 \overline{P_2}}{n_1 + n_2} = \frac{S_1 + S_2}{n_1 + n_2}$$

```
> z<-(P.Placebo-P.Aspirin)/sqrt(P.tot*(1-P.tot)*(1/n.P+1/n.A))
> z
[1] 5.001388
```
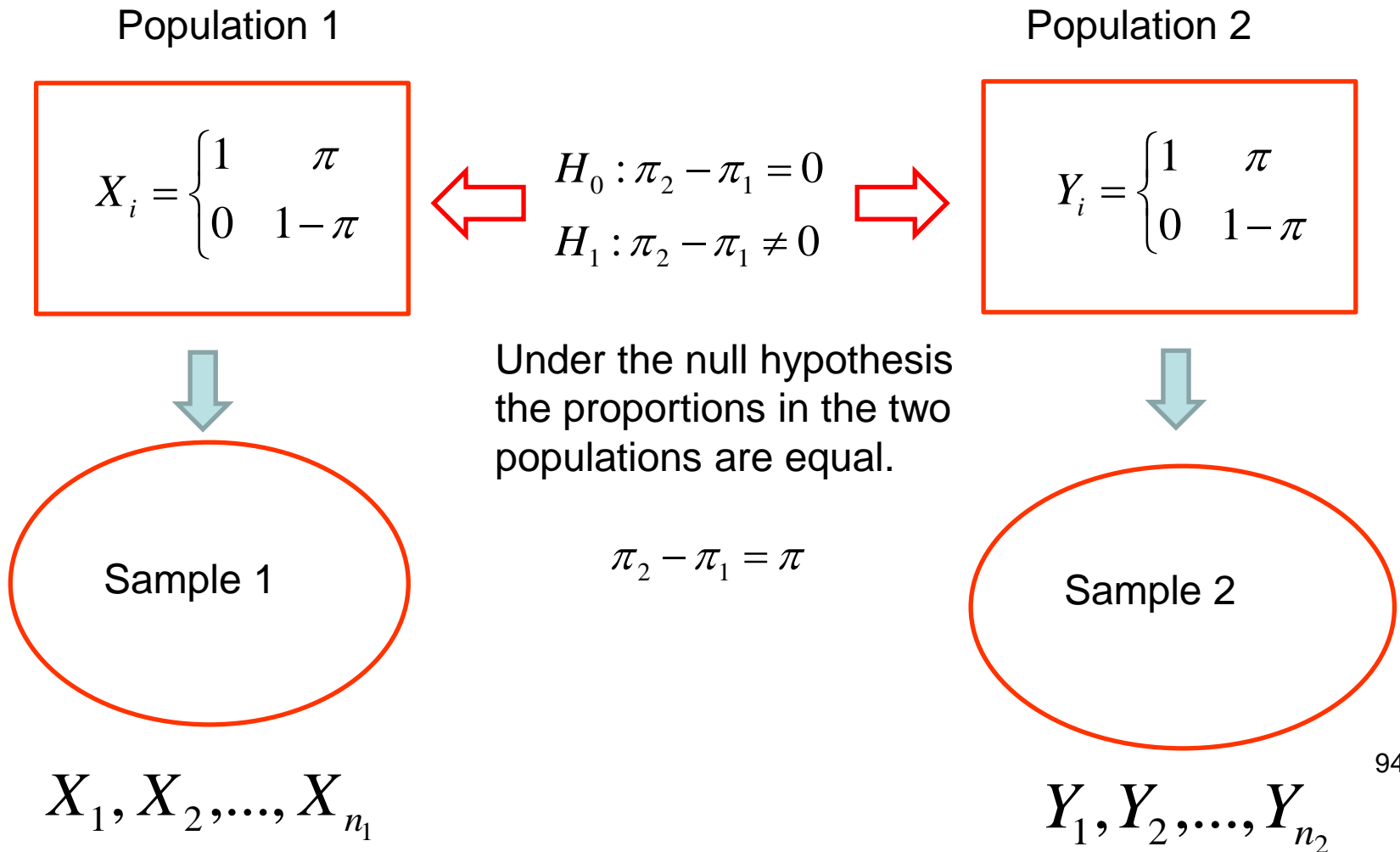
$$z = \frac{\overline{P_2} - \overline{P_1}}{\sqrt{\overline{P}(1-\overline{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

```
> 1-pnorm(z,0,1)
[1] 2.845948e-07
```
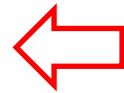$p - value$

The null hypothesis is rejected.
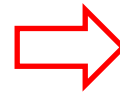
Parametric bootstrap test:
populations and samples

Population 1

Population 2

$$X_i = \begin{cases} 1 & \pi \\ 0 & 1-\pi \end{cases}$$

$$H_0 : \pi_2 - \pi_1 = 0$$

$$H_1 : \pi_2 - \pi_1 \neq 0$$

$$Y_i = \begin{cases} 1 & \pi \\ 0 & 1-\pi \end{cases}$$

Under the null hypothesis the proportions in the two populations are equal.

$$\pi_2 - \pi_1 = \pi$$

Sample 1

Sample 2

$$X_1, X_2, ..., X_{n_1}$$

$$Y_1, Y_2, ..., Y_{n_2}$$

Parametric bootstrap test

Population 1

$$X_i = \begin{cases} 1 & \pi \\ 0 & 1-\pi \end{cases}$$

$$H_0 : \pi_2 - \pi_1 = 0$$
$$H_1 : \pi_2 - \pi_1 \neq 0$$

Population 2

$$Y_i = \begin{cases} 1 & \pi \\ 0 & 1-\pi \end{cases}$$

$$\hat{\pi} = \overline{P} = \frac{n_1 \overline{P_1} + n_2 \overline{P_2}}{n_1 + n_2} = \frac{S_1 + S_2}{n_1 + n_2}$$

Table 2.3. Cross Classification of Aspirin Use and Myocardial Infarction

Group	Myocardial Infarction		Total
	Yes	No	
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262–264, 1988.

$$n_1 = 11034$$
$$n_1 = 11037$$

$$\hat{\pi} = \overline{P} = \frac{189 + 104}{11034 + 11037}$$

95

Parametric bootstrap test

For 1:B

Population 1

Population 2

$$x_b^* \sim B(11034, \hat{\pi})$$

$$H_0 : \pi_2 - \pi_1 = 0$$
$$H_1 : \pi_2 - \pi_1 \neq 0$$

$$y_b^* \sim B(11037, \hat{\pi})$$

$$\hat{p}_{1b}^* = \frac{1}{n_1} \sum x$$

$$\hat{p}_{2b}^* = \frac{1}{n_2} \sum y$$

$$p_b^* = \frac{S_{1b}^* + S_{2b}^*}{n_1 + n_2}$$

$$z_b^* \frac{\hat{p}_{1b}^* - \hat{p}_{2b}^*}{\sqrt{p_b^*(1 - p_b^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Test statistic for the b'th bootstrap sample.

The bootstrap loop

96

Two sided parametric bootstrap test

Simulate B samples and
calculate the test statistic values:

$$z_1^*, z_2^*, ..., z_B^*$$

Bootstrap p value (for two sided test):

$$P = \frac{1 + \#\left\{\left|z_b^*\right| > \left|z^{obs}\right|\right\}}{B + 1}$$

Histogram of z.b
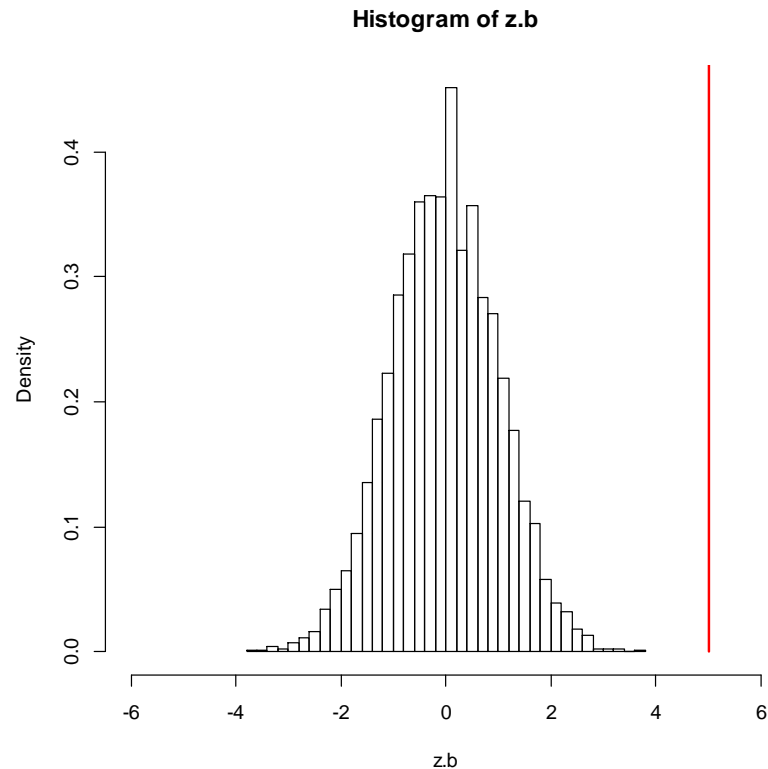
R code: parametric bootstrap

```
B<-10000
z.b<-c(1:B)
for(i in 1:B)
{
S1<-sum(rbinom(n.P,1,P.tot))
P.P.b<-S1/(189+10845)
S2<-sum(rbinom(n.A,1,P.tot))
P.A.b<-S2/(104+10933)
P.tot.b<-(S1+S2)/(n.P+n.A)
z.b[i]<-(P.P.b-P.A.b)/sqrt(P.tot.b*(1-P.tot.b)*(1/n.P+1/n.A))
}
```
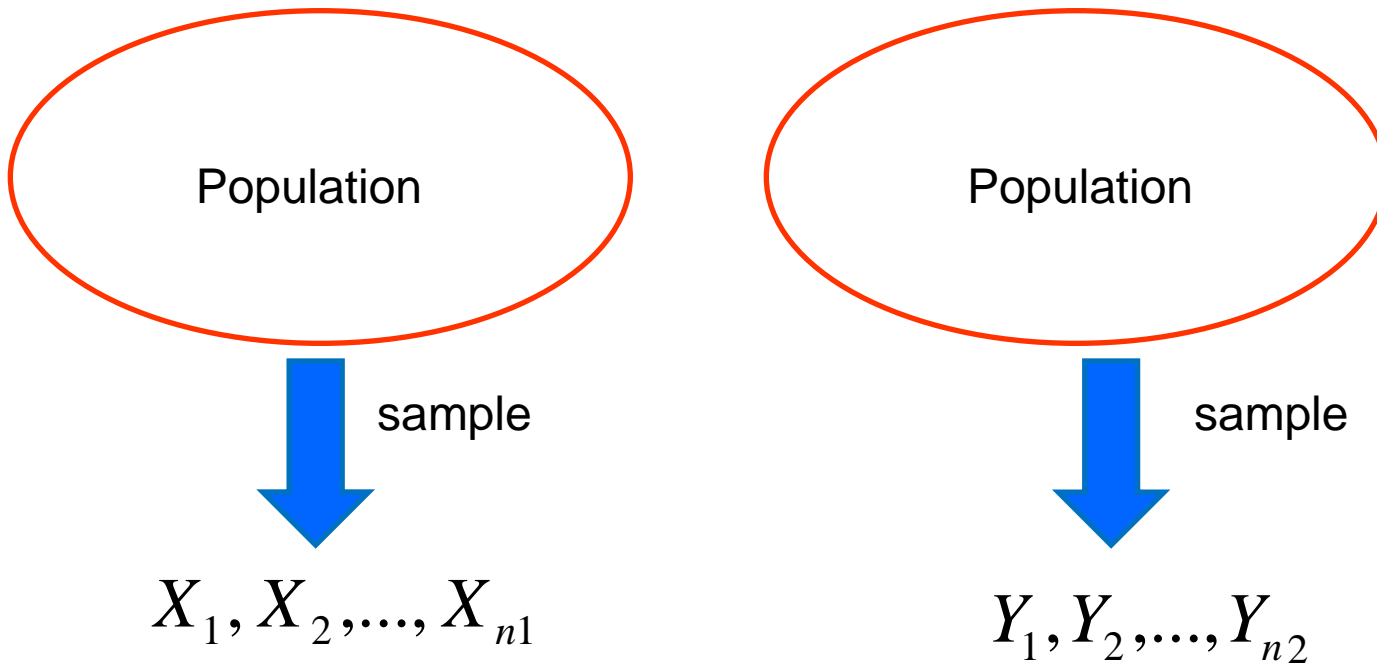
$$x_b^* \sim B(11034, \hat{\pi})$$

$$y_b^* \sim B(11037, \hat{\pi})$$

$$z_b^* \frac{\hat{p}_{1b}^* - \hat{p}_{2b}^*}{\sqrt{p_b^*(1-p_b^*)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

Non parametric bootstrap:
populations and samples

$$(Z_1, Z_2, ..., Z_{n1}, Z_{n1+1}, ..., Z_{n1+n2}) = (X_1, X_2, ..., X_{n1}, Y_1, ..., Y_{n2})$$

Application of algorithm 16.1

$$z = (y, x)$$

Simulate B samples of size n_1+n_2 with replacement from z.

The first n_1 are considered as observations x* and the remaining n_2 as observations y*.

Calculate the test statistic for each bootstrap sample:

$$t(z^*)_1, t(z^*)_2, ..., t(z^*)_B$$

Test statistic for the b'th bootstrap sample:

$$t(z^*)_b = z_b^* \frac{\hat{p}_{1b}^* - \hat{p}_{2b}^*}{\sqrt{p_b^*(1-p_b^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
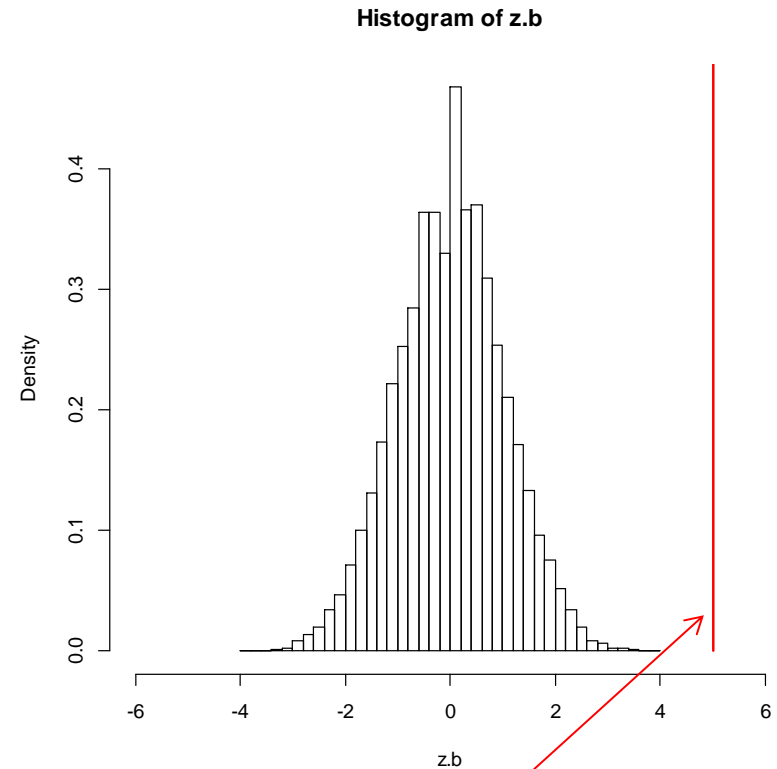
Two sided non parametric bootstrap test

Simulate B samples and
calculate the test statistic values:

$$z_1^*, z_2^*, \ldots, z_B^*$$

Bootstrap p value (for two sided test):

$$P = \frac{1 + \#\left\{\left|z_b^*\right| > \left|z^{obs}\right|\right\}}{B+1}$$

The observed test statistic is larger than all bootstrap statistics

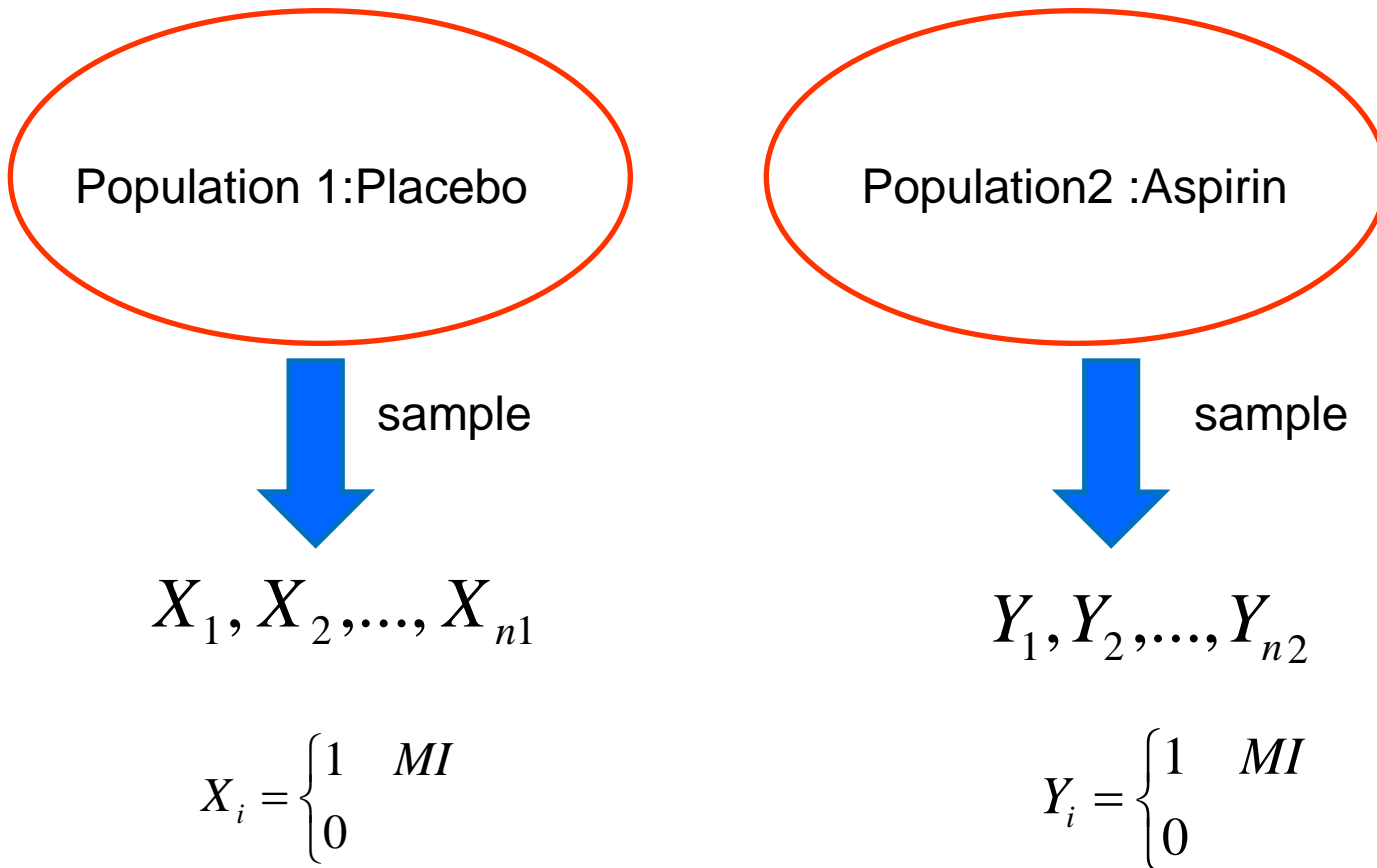$$P = \frac{1}{B+1}$$

101

R code for non parametric bootstrap

```
z.i<-c(Placebo,Aspirin)
nz<-length(z.i)
nz
```

$$(Z_1, Z_2, ..., Z_{n1}, Z_{n1+1}, ..., Z_{n1+n2}) = (X_1, X_2, ..., X_{n1}, Y_1, ..., Y_{n2})$$

$n_1 + n_2$

```
B<-10000
z.b<-c(1:B)
for(i in 1:B)
{
z.boot<-sample(z.i,size=nz,replace =TRUE)
x.boot<-z.boot[1:n.P]
y.boot<-z.boot[(n.P+1):nz]
S1<-sum(x.boot)
P.P.b<-S1/(189+10845)
S2<-sum(y.boot)
P.A.b<-S2/(104+10933)
P.tot.b<-(S1+S2)/(n.P+n.A)
z.b[i]<-(P.P.b-P.A.b)/sqrt(P.tot.b*(1-P.tot.b)*(1/n.P+1/n.A))
}

hist(z.b,nclass=50,xlim=c(-6,6),probability=TRUE)
lines(c(z,z),c(0,100),col=2,lwd=2)
```

$$z_b^* \frac{\hat{p}_{1b}^* - \hat{p}_{2b}^*}{\sqrt{p_b^*(1-p_b^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Permutations test:
populations and samples

Population 1:Placebo

Population2 :Aspirin

sample

sample

$$X_1, X_2, ..., X_{n1}$$

$$Y_1, Y_2, ..., Y_{n2}$$

$$X_i = \begin{cases} 1 & MI \\ 0 & \end{cases}$$

$$Y_i = \begin{cases} 1 & MI \\ 0 & \end{cases}$$

$$v = (X_1, X_2, ..., X_{n1}, Y_1, ..., Y_{n2})$$

Permutations

We define two vectors:

Response (MI)

$$v = (v_1, v_2, ..., v_N)$$

$$v_i = \begin{cases} x_i & observation \in P \\ y_i & observation \in A \end{cases}$$

$$v_i = \begin{cases} 1 & MI \\ 0 & not \quad MI \end{cases}$$
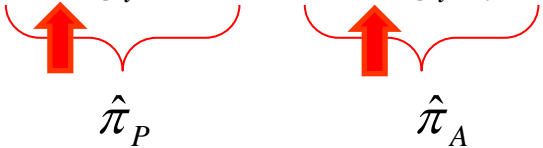
Membership

$$g = (g_1, g_2, ..., g_N)$$

$$g_i = \begin{cases} x & observation \in P \\ y & observation \in A \end{cases}$$

Test Statistic

$$\hat{\theta} = S(g, v)$$

$$\hat{\theta} = \frac{1}{n_1} \sum_{g_i = x} v_i - \frac{1}{n_2} \sum_{g_i = y} v_i$$

$\underbrace{}_{\hat{\pi}_P} \qquad \underbrace{}_{\hat{\pi}_A}$

The test statistic is a function of the two vectors.

The mean difference $(\hat{\pi}_P - \hat{\pi}_A)$.

Permutations

$$\hat{\theta}^*(b) = S(g^*(b), v)$$

$$z_b^* = \hat{\theta}^*(b) = \frac{\hat{p}_{1b}^* - \hat{p}_{2b}^*}{\sqrt{p_b^*(1 - p_b^*)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

$$ALS = \frac{\#\left(\hat{\theta}^*(b) \geq \hat{\theta}\right)}{B}$$

We draw B samples without replacement (why ?) from the vector v.

The first n are assigned to the first group the last m are assigned to the second group.
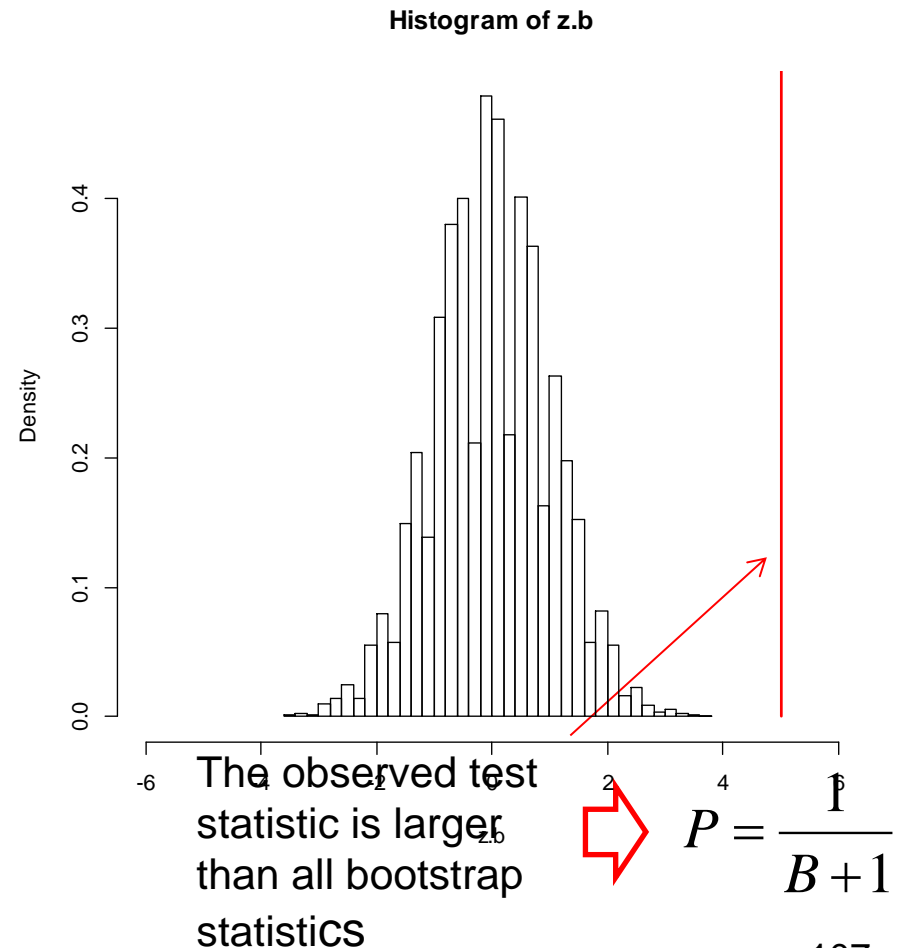
Two sided non parametric bootstrap test

Simulate B samples and
calculate the test statistic values:

$$z_1^*, z_2^*, \ldots, z_B^*$$

Bootstrap p value (for two sided test):

$$P = \frac{1 + \#\left\{\left|z_b^*\right| > \left|z^{obs}\right|\right\}}{B+1}$$

Histogram of z.b

Density

The observed test statistic is larger than all bootstrap statistics

$$P = \frac{1}{B+1}$$

107

R code for non parametric bootstrap

```
z.i<-c(Placebo,Aspirin)
nz<-length(z.i)
nz
```

$$v = (X_1, X_2, ..., X_{n1}, Y_1, ..., Y_{n2})$$

nz $\longrightarrow$ $n_1 + n_2$

```
for(i in 1:B)
{
z.boot<-sample(z.i,size=nz,replace =FALSE)
x.boot<-z.boot[1:n.P]
y.boot<-z.boot[(n.P+1):nz]
S1<-sum(x.boot)
P.P.b<-S1/(189+10845)
S2<-sum(y.boot)
P.A.b<-S2/(104+10933)
P.tot.b<-(S1+S2)/(n.P+n.A)
z.b[i]<-(P.P.b-P.A.b)/sqrt(P.tot.b*(1-P.tot.b)*(1/n.P+1/n.A))
}
```

$$z_b^* \frac{\hat{p}_{1b}^* - \hat{p}_{2b}^*}{\sqrt{p_b^*(1-p_b^*)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$