Basic skills in Bootstrap using R

Ziv Shkedy, Hasselt University

Introduction

Skills. . . .

This book contains all R code used to produce the results for the course “computer intensive methods using R: an introduction”. The focus of this book is skills and not only the theory behind. In other words, we focus on the question “How to do it ?” and not just of the queation “why to do it ?”. The approach we take in the book is to program the procedure and not to use R functions/packages to produce the results. All examples are illustrated using the R software.

R ?

Only basic knowledge in R is required. All the models discussed in the book can be fitted using the `lm()` and `glm()` function in R. The datasets used for illustrations are available in R and many of them are part of the `bootstrap` R package. To run the code smoothly, this package need to be installed.

```
library(bootstrap)
```

A for loop in R

A for loop in R is a loop in which we repeatedly ask R to do the same action in each step of the loop. For example, suppose that we would like to draw a sample of 10 observations from $N(0, 1)$. In R this can be done using the code

```
x<-rnorm(10,0,1)
```

The sample is

```
x
```

```
##  [1] -0.04681759 -0.98551173 -0.79878917  1.80749837 -0.45727204  0.24069254
##  [7]  0.21624841 -1.61544945 -0.90715884  0.46153241
```

and the sample mean

```
mx<-mean(x)
mx
```

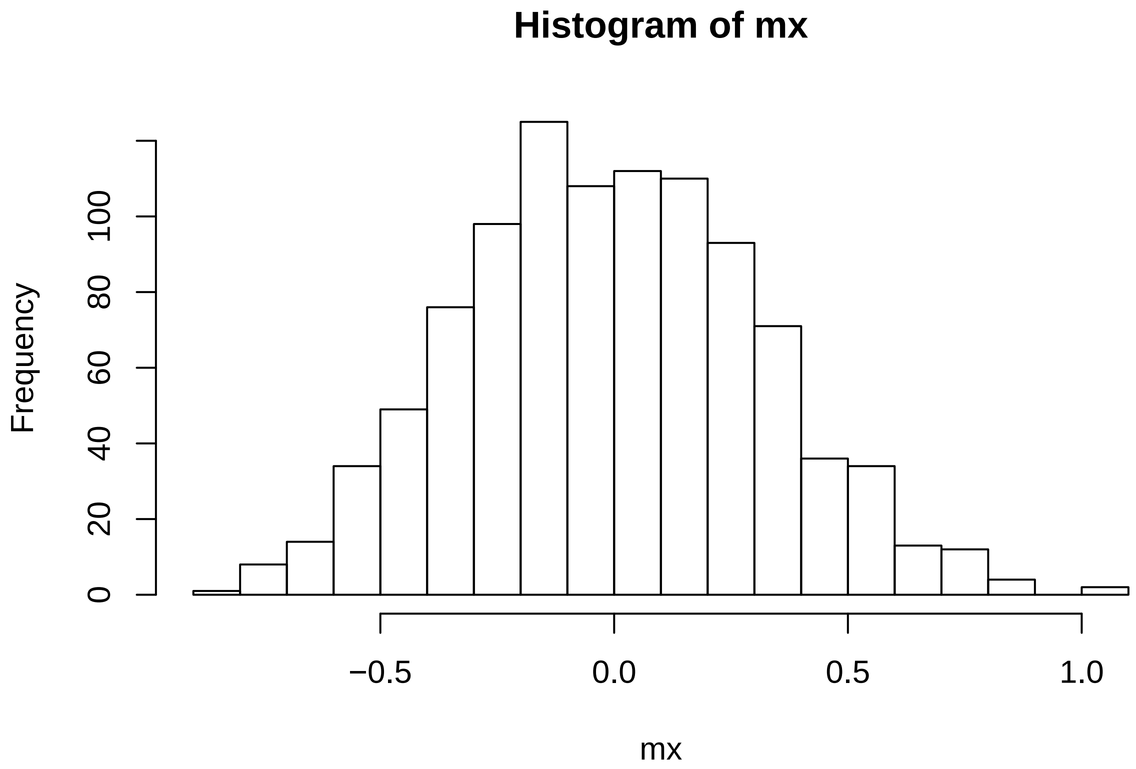
```
## [1] -0.2085027
```

Suppose that we would like to draw a sample of 10 observations from $N(0, 1)$ 1000 times. To do this we can use a “for loop” in the following way:

```
mx<-c(1:1000)
for(i in 1:1000)
{
x<-rnorm(10,0,1)
mx[i]<-mean(x)
}
```

A histogram of the sample means:

```
hist(mx,nclass=20)
```



Notaion

Throughout out the book we keep, as much as we can the notion presented in the book of Efron and Tibshirani (1993) “An Introduction to the Bootstrap”.

Just do it

If you did not have a problem to understand the R code above, you will not have any problem to understand the R code that we use to produce all the output for the examples discuss in this book. If you had a difficulty to understand the “for loop” example above you need a short training, at a beginner level, in R.

YouTube tutorial: the for loop in R

For a short online YouTube introduction by Richard Webster about a for loop in R see YTBootstrap 1.

R course online

An introductory course for R is able online in the >eR-BioStat website. See Rcourse.

Part I

Introduction

The empirical distribution function and the plug-in principle

In the first part of this book we cover basic concept that will be used in different chapters I the book:

- Sampling from a population.
- The plug-in principle.
- The empirical distribution.
- The accuracy of the sample mean a, the standard error and estimated standard error.

Slides

Slides for the first part that covers the topics of sampling for a population, the accuracy of the sample means and the glug-in principle can be found here: Slides1.

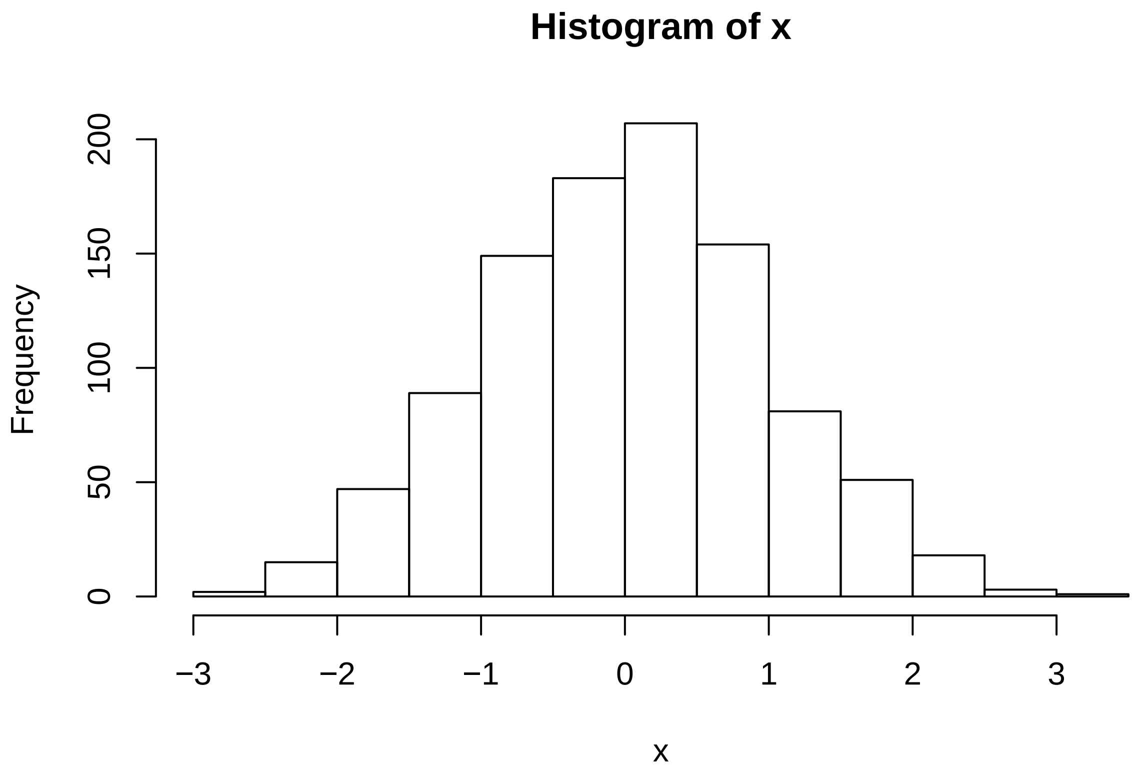
Sampling for a population

Let $x_1, \dots, x_n$ be a random sample from a probability distribution F and let $\theta = \theta(F)$ be a parameter of interest. For example, let F be a standard normal distribution, $F = N(\mu = 0, \sigma^2 = 1)$ and consider a sample of 1000 observations,

```
x<-rnorm(1000,0,1)
```

The distribution of the sample is shown below

```
hist(x)
```



The parameter of interest of the mean, $\mu = 0$ and the parameter estimate $\hat{\mu} = \bar{x}$ is equal to

```
mean(x)
```

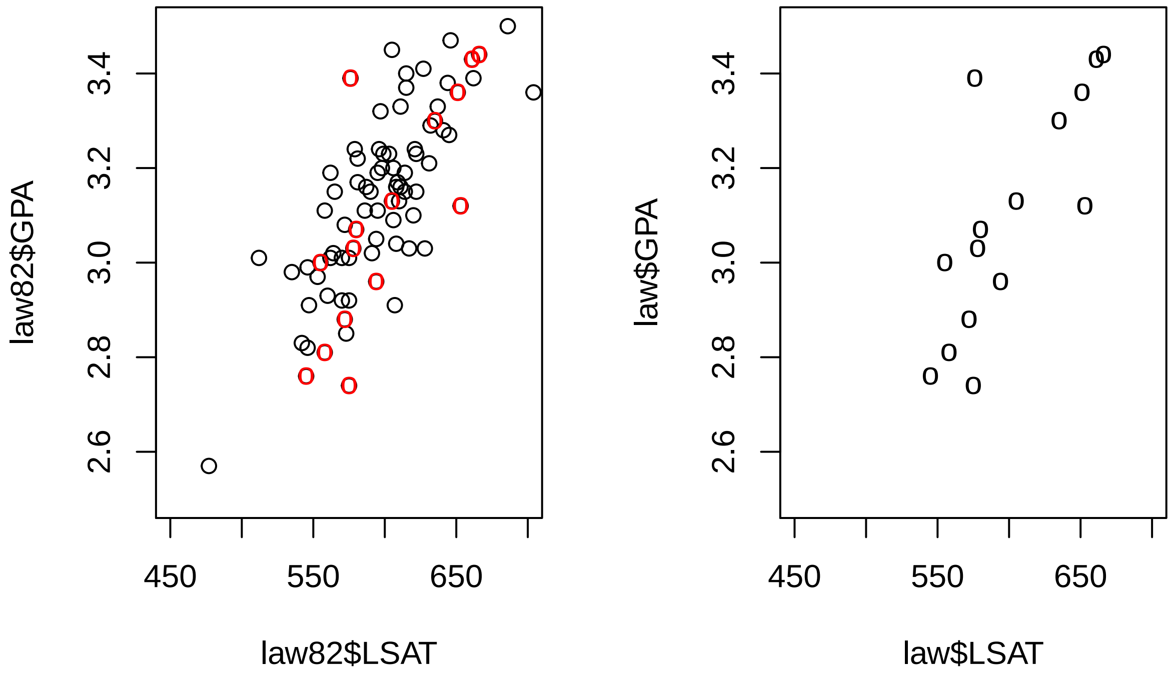
```
## [1] 0.01950897
```

We note that $\bar{x}$ is defined on teh sample, that is $\hat{\mu} = \bar{x} = T(x_1, \ldots, x_n)$.

Example: the low school data

The law school data contain information about a random sample of size n=15 from the population of 82 USA law schools. The data consists of two measurements: LSAT (average score on a national law test) and GPA (average undergraduate grade-point average). The R object $law82S contains data for the whole population of 82 law schools. The left panel in the Figure below present the population of 82 scools while the right panel the random sample of 15 schools out the population.

```
par(mfrow=c(1,2))
plot(law82$LSAT,law82$GPA,xlim=c(450,700),ylim=c(2.5,3.5))
points(law$LSAT,law$GPA,pch="o",col=2)
plot(law$LSAT,law$GPA,pch="o",xlim=c(450,700),ylim=c(2.5,3.5))
```



The population correlation and the sample correlations.

```
cor(law82$LSAT,law82$GPA)
```

```
## [1] 0.7599979
```

```
cor(law$LSAT,law$GPA)
```

```
## [1] 0.7763745
```