Bootstrapping with R

K.O.Obisesan, Ph.D

University of Ibadan

September 28, 2019

Table of Contents

Introduction to Bootstrapping

Statistical Methods

- We are daily bombarded with numbers, charts, graphs and statistical results (Afonja et. al, 2014)
- There are situations in which it is extremely difficult to obtain more data e.g (small sample size, non-normal distributions)
- In these cases, what do we do?

Introduction Contd'

- Bootstrapping is a very essential tool for statisticians and mathematicians. It is a tool that allows for the estimation of confidence intervals, hypothesis testing, standard errors and statistic (mean, proportion etc).
- It is a non parametric statistical technique that falls under a broader heading known as resampling.
- It is a techniques that allows you turn statistic into random variables.

Bootstrapping Concept

Procedures for Bootstrapping

- take a sample from your data
- create hundreds of new samples, called **bootstrap samples** by sampling with replacement from your sample data
- calculate the statistic for each resamples
- compute the confidence interval using standard error method.

Why Bootstrapping?

Advantages of Bootstrapping

- Fewer assumptions: for example, bootstrapping methods do not require that distributions be Normal or that sample size be large
- Greater accuracy
- Generality of method.

Bootstrapping the Mean

- To illustrate the use of bootstrapping, we make use of the dollar amounts spent by 20 consecutive shoppers at a supermarket. We are willing to regard this as an SRS of all shoppers at this market.

```
 3.11  8.88  9.26 10.81 12.69 13.78 15.23 15.62 17.00
17.39 18.36 18.43 19.27 19.50 19.54 20.16 20.59 22.22
23.04 24.47
```

R Code

```
data<-c(3.11,8.88,9.26,10.81,12.69,13.78,15.23,15.62,
17.00,17.39,18.36,18.43,19.27,19.50,19.54,20.16,20.59,
22.22,23.04,24.47)
### getting the mean of data
mean(data)
means<-c()
```

R Code cont:

```
for(i in 1:1000){
samp<-sample(data, length(data), replace=T)
means<-c(means, mean(samp)
                }
```

A Deeper Look

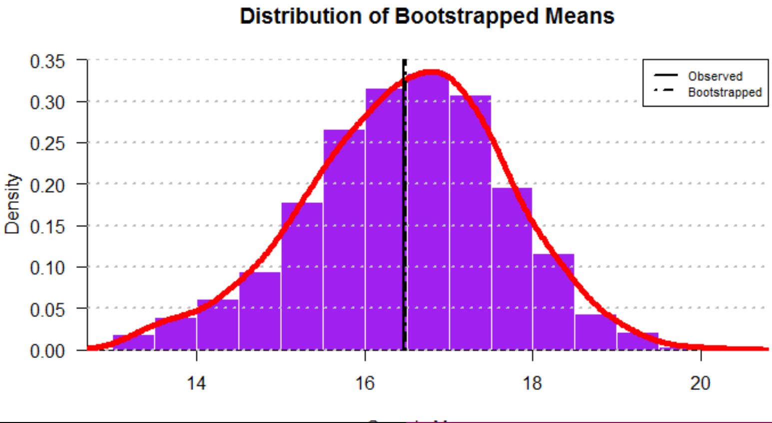
What the Code is Doing

- the data set is loaded
- the mean of the data set is computed
- an empty vector named means is created
- in the loop a sample of same size with data(20) is taken from the data with replacement and the sample means are computed
- the step above is repeated 1000 times.

Some Descriptives

```
## first 10 means computed
head(means,10)
## last 10 means computed
tail(means,10)
## median of the computed means
median(means)
## mean of the computed means
mean(means)
```

Histogram and Density of Bootstrapped Means



Bootstrap Standard Error and Confidence Interval

- The bootstrap standard error of a statistic is the standard deviation of the bootstrap distribution.

$$SE_{boot,\overline{x}} = \sqrt{\frac{1}{B-1}\sum\left(\overline{x}^{*} - \frac{1}{B}\sum\overline{x}^{*}\right)^{2}}$$

- In this expression, $\overline{x}^{*}$ is the mean value of the individual resample with B the number of resamples.

```r
SE <- sqrt(var(means))
SE
```

[1] 1.186762

```r
#confidence interval
mean(means) +c(-1,1)*1.96*SE
```

[1] 14.16106 18.81316

Bootstrapping and Sampling Distribution

- Let us make a quick comparison of bootstrapping with Sampling Distribution of statistics.

```
SE
```

```
[1] 1.202745
```

```
#confidence interval
```

```
[1] 14.11072 18.82488
```

Bootstrapping the Median

- The median of an sample data is an important measure of location but being a statistic without distribution, the confidence interval or standard error of the median cannot be computed except through the use of bootstrap. In this example we are going to compute the confidence interval and standard error of the medain of the **mpg** variable in the **mtcars** in the **datasets** package.

R Code

```r
library(datasets)
data(mtcars)
### getting the median mpg
median(mtcars$mpg)
medians <- c()
```

R Code cont:

```
for(i in 1:700){
  samp <- sample(mtcars$mpg, 20,replace=T)
  medians <- c(medians,median(samp))
}
```

A Deeper Look

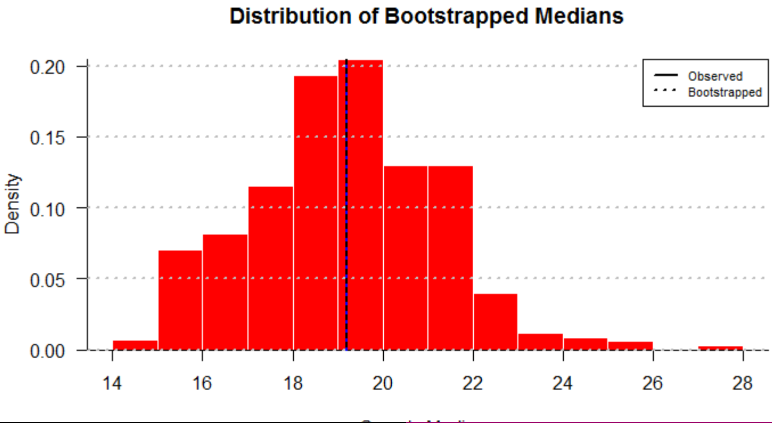
What is the Code Doing?

- the mtcars data set is loaded
- the median of the mpg varaible in the mtcars dataset is computed
- an empty vector named medians is created
- in the loop a sample of 20 is taken from mpg and the sample median is computed
- the step above is repeated 700 times.

Some Descriptives

```
## first 10 medians computed
head(medians,10)
## last 10 medians computed
tail(medians,10)
## median of the computed medians
median(medians)
## mean of the computed medians
mean(medians)
```

Histogram of Bootstrapped Medians



Standard Error and Confidence Interval

```
SE <- sqrt(var(medians))
SE
```

[1] 2.102774

```
confidence interval
median(medians) +c(-1,1)*1.96*SE
```

[1] 15.07856 23.32144

Bootstrapping Regression Coefficient

Bootstrapping in Regression

- Bootstrapping isn't limited to the median or mean, it is actually applicable to any estimation procedure.

```
x <- runif(100,0,1) ### 100 random uniform
error <- rnorm(100,0,1) ### 100 random normal
y <- 2 + 0.87*x + error ### y samples
regmod <- lm(y~x) ### the linear model
```

Bootstrapping Regression Coefficient Cont:

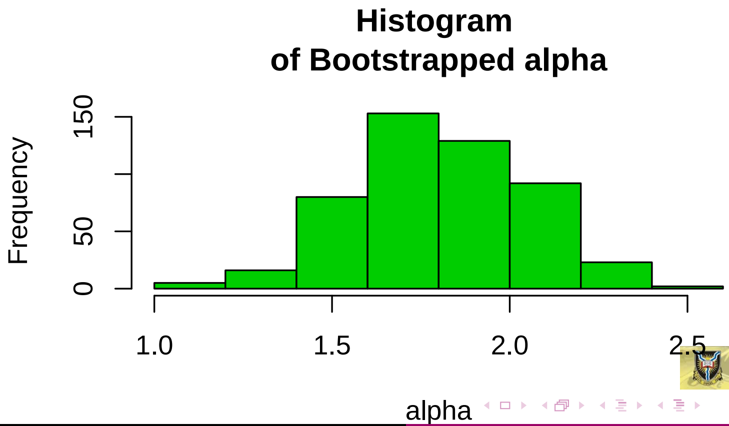
```
nboot <- 500 ### number of intended bootstraps
coefboot <- array(0,dim=c(nboot,2))
### The bootstrap loop
for(i in 1:nboot)
  {
    ystar <- y + sample(error,replace=T)
    bootfit <- lm(ystar~x)
    coefboot[i,] <- bootfit$coefficients
  }
```

Bootstrapping Regression Coefficient Cont:

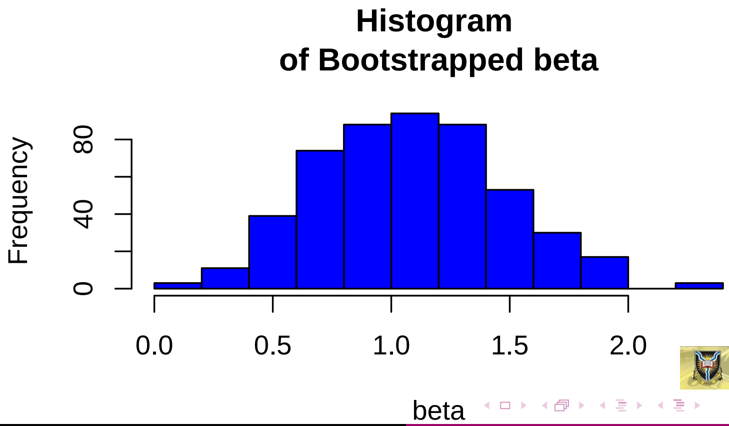
What Did I do?

- we simulated 100 samples from a uniform distribution
- simulated the errors from a normal distribution, 100 of them as well
- we obtained y
- the bootstrap loop is formed and we estimated 500 different coefficients

Histogram of Bootstrapped alpha



Histogram of Bootstrapped Beta



Disadvantages

Why Not Bootstrapping?

- Not as rigid conditions as Central Limit Theorem based methods
- A representative sample is required for generalizabiity. If the sample is biased, the estimate resulting from this sample will also be baised.

Final Note

Use R!

- This is not the whole idea behind Bootstrapping but it definitely is a fine start.
- Further reading on bootstrapping is suggested.
- Till we meet again, **Use R!!!**

References

1. Hesterberg et. al.,(2003): Bootstrap methods and Permutation Test. *Companion chapter 18 to the practice of Business Statistics*.
2. Afonja B. Olubusoye O.E., Ossai E. and Arinola J.(2014): Introductory Statistics: A Learners' Motivated Approach.
3. Barum W.J. and Duncan J.M (2007): A First Course in Statistical Programming: *Cambridge University Press*.
4. Dalgard (2008): Introductory Statistics with R. *Springer*.