Statistical Inference - Final Project

Alberto Rossi

09/06/2020

Overview

This exercise aims to analyze two bases in R: the first is the exponential distribution and compare it with the Central Limit Theorem. Then, the second analysis will be made a basic analysis of inference on a database called ToothGrowth, present in the standard R.

Part 1: Simulation Exercise Instructions

Let's analyze the similarities between the exponential distribution with the Central Limit Theorem

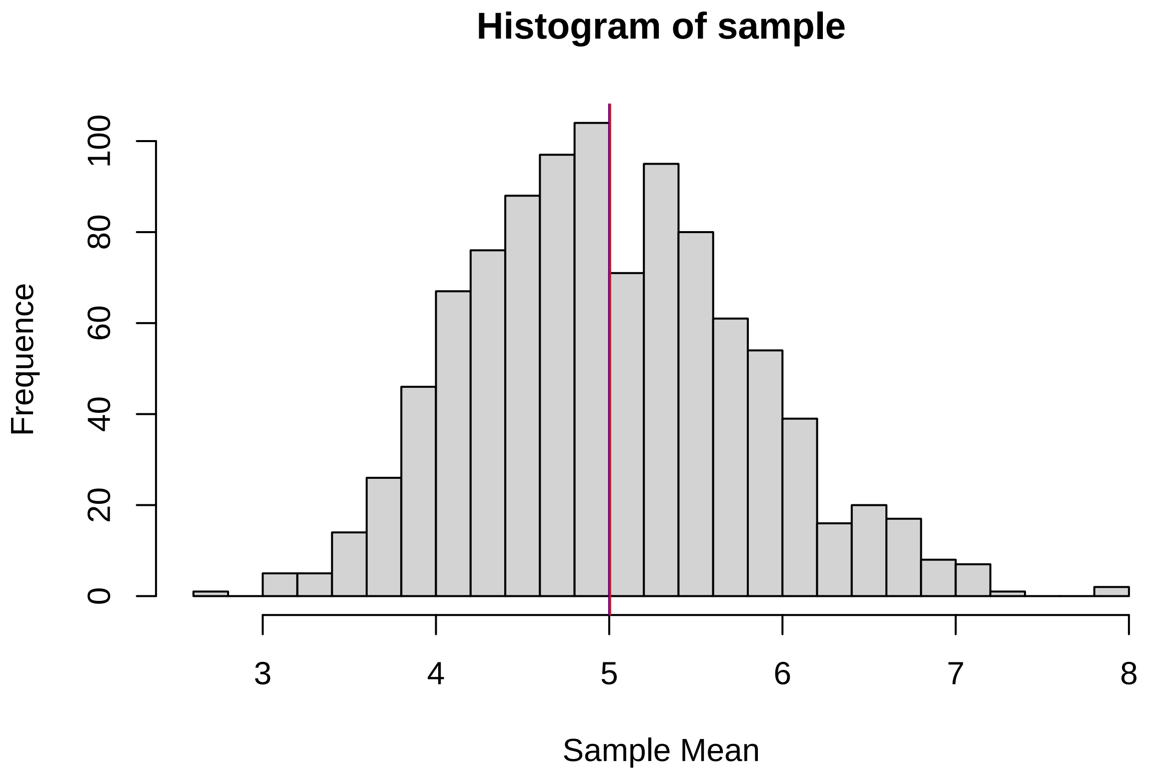
First things first: Load data

```
#Load data and parameters from exercise

#Number os samples
n <- 40

#Lambda = 0.2
lambda <- 0.2

#Number of simulations
b <- 1000

#CI = 95%
z <- 1.96

#As data is from aleatory distribution generation, set seed for reproducibility
set.seed(96)
```

1. Show the sample mean and compare it to the theoretical mean of the distribution.

```
#Create a table to hold data
data <- matrix(rexp(n * b, rate = lambda), b)

#Simulated mean per sample row
```

```r
sample <- rowMeans(data)

#Mean of means
simMeanMean <- mean(sample)

#Theoretical exponential mean
theMean <- 1/lambda

#Histogram of the sample means and theoretical mean
hist(sample, xlab="Sample Mean", ylab = "Frequence", breaks = 30)
abline(v=theMean, col="blue", lwd=1)
abline(v=simMeanMean, col="red", lwd=1)
```



Make a really zoom to see that theoretical mean and sample mean are almost identical.

Check the real numbers:

```
## [1] "Simulated exponential mean: 5.01"
```

```
## [1] "Theoretical mean: 5"
```

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

Again, simulated exponential variance and theoretical variance are really close:

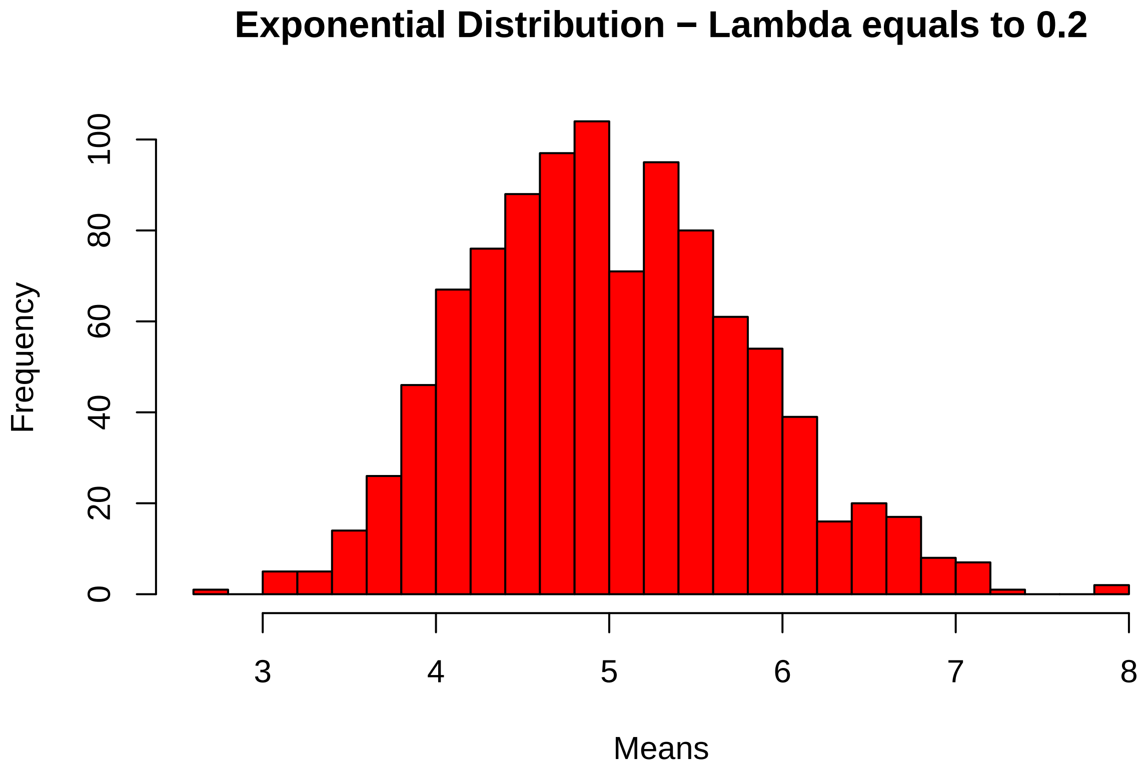
```
## [1] "Simulated exponential variance: 0.64"
```

```
## [1] "Theoretical variance: 0.62"
```

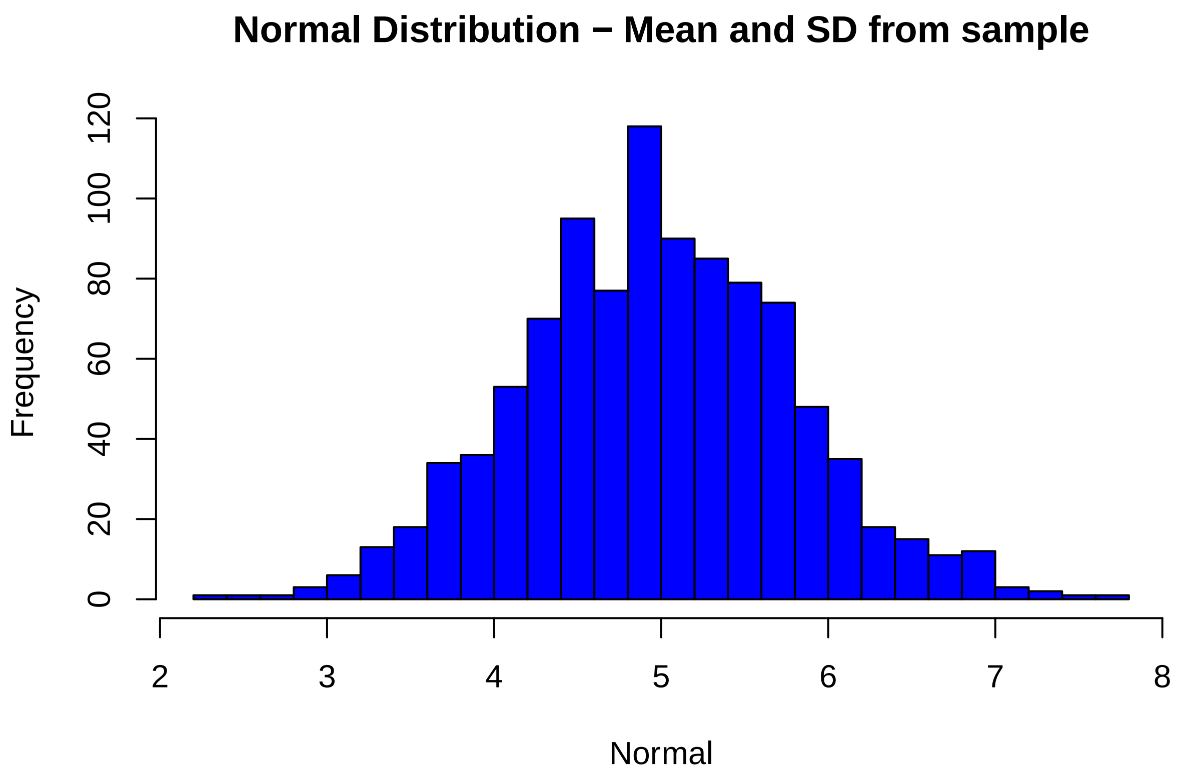
3. Show that the distribution is approximately normal.

It is enough to show both distributions, so that the similarities are evident:

```
simNorm <- rnorm(1000, mean = simMeanMean, sd = sd(sample))
hist(sample, breaks = 20, main = "Exponential Distribution - Lambda equals to 0.2", xlab = "Means", c
```



```
hist(simNorm, breaks = 20, main = "Normal Distribution - Mean and SD from sample", xlab = "Normal", c
```



Part 2: Basic Inferential Data Analysis Instructions

1. Load the ToothGrowth data and perform some basic exploratory data analyses

This dataset contains data from a study on the Effect of Vitamin C on Tooth Growth in Guinea Pigs.

```
library(ggplot2)
library(stats)
data(ToothGrowth)
```

2. Provide a basic summary of the data.

Basically we have 3 variables with 60 observations:

```
summary(ToothGrowth)
```

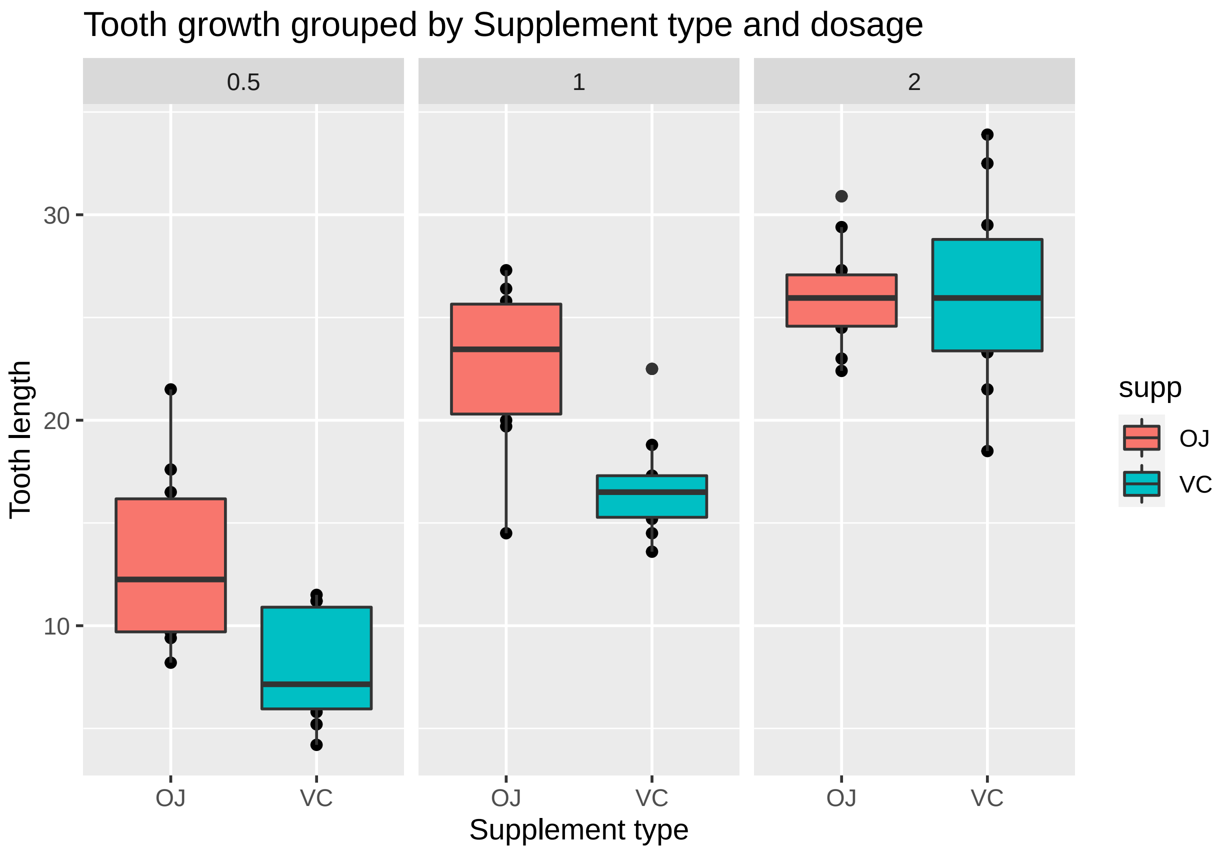
```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000  
```

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Visually, with grouped data:

```
qplot(x=supp,y=len,data=ToothGrowth, facets=~dose, main="Tooth growth grouped by Supplement type and d
```



3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

If we assume the data is normally distributed, we have a null hypothesis that there is no difference between the mean under each kind of supplements, or each dose of the supplements:

```
t.test(x = ToothGrowth$len, data = ToothGrowth, paired = FALSE, conf.level = 0.95)$conf.
```

```
## [1] 16.83731 20.78936
## attr(,"conf.level")
## [1] 0.95
```

We will be able to construct a confidence interval that 95% of the time, an interval between 16.84 and 20.79 will contain the true mean of the population.

Then we calculate the mean under each kind of supplements, and each dose of the supplements, “OJ” and “VC”:

```
mean(ToothGrowth[ToothGrowth$supp == "OJ", ]$len)
```

```
## [1] 20.66333
```

```
mean(ToothGrowth[ToothGrowth$supp == "VC", ]$len)
```

```
## [1] 16.96333
```

Thus the mean of teeth growth after taking OJ is 20.66; the mean of teeth growth after taking VC is 16.96. Both the two are within the confidence interval.

We fail to reject the null hypothesis that there is not a difference in teeth growth after taking the two types of supplements.

Now lets check if there is a difference in teeth growth between each dose of supplements:

```
  mean(ToothGrowth[ToothGrowth$dose == 0.5,]$len)
```

```
## [1] 10.605
```

```
  mean(ToothGrowth[ToothGrowth$dose == 1,]$len)
```

```
## [1] 19.735
```

```
  mean(ToothGrowth[ToothGrowth$dose == 2,]$len)
```

```
## [1] 26.1
```

Thus the mean of teeth growth after taking dose of 0.5 is 10.6; the mean of teeth growth after taking dose of 1.0 is 19.74; the mean of teeth growth after taking dose of 2.0 is 26.1.

We are able to reject the null hypothesis, and there is a difference in teeth growth between each dose of supplements.

Now we know the data is not normally distributed under each dose, along with this conclusion, we may assume the data is normally distributed within each dose. Then we will be able to compare the teeth growth between each supplements under each dose.

```
  dose05 <- ToothGrowth[ToothGrowth$dose == 0.5, ]
  t.test(x = dose05$len, paired = FALSE, conf.level = 0.95)$conf
```

```
## [1]  8.499046 12.710954
## attr(,"conf.level")
## [1] 0.95
```

```r
  mean(dose05[dose05$supp == "VC", ]$len)
```

```
## [1] 7.98
```

```r
  mean(dose05[dose05$supp == "OJ", ]$len)
```

```
## [1] 13.23
```

Thus under the dose of 0.5, there are 95% of the time that a confidence interval between 8.50 and 12.71 will contain the true population mean. We also know that the mean of teeth growth after taking 0.5 dose of VC is 7.98 and the mean of teeth growth after taking 0.5 dose of OJ is 13.23. We rejected the null hypothesis.

```r
  dose10 <- ToothGrowth[ToothGrowth$dose == 1, ]
t.test(x = dose10$len, paired = FALSE, conf.level = 0.95)$conf
```

```
## [1] 17.66851 21.80149
## attr(,"conf.level")
## [1] 0.95
```

```r
  mean(dose10[dose10$supp == "VC", ]$len)
```

```
## [1] 16.77
```

```r
  mean(dose10[dose10$supp == "OJ", ]$len)
```

```
## [1] 22.7
```

Thus under the dose of 1.0, there are 95% of the time that a confidence interval between 17.67 and 21.80 will contain the true population mean. We also know that the mean of teeth growth after taking 1.0 dose of VC is 16.77 and the mean of teeth growth after taking 1.0 dose of OJ is 22.7. We rejected the null hypothesis

```r
  dose20 <- ToothGrowth[ToothGrowth$dose == 2, ]
t.test(x = dose20$len, paired = FALSE, conf.level = 0.95)$conf
```

```
## [1] 24.33364 27.86636
## attr(,"conf.level")
## [1] 0.95
```

```r
  mean(dose20[dose20$supp == "VC", ]$len)
```

```
## [1] 26.14
```

```r
  mean(dose20[dose20$supp == "OJ", ]$len)
```

```
## [1] 26.06
```

Thus under the dose of 2.0, there are 95% of the time that a confidence interval between 24.33 and 27.87 will contain the true population mean. We also know that the mean of teeth growth after taking 2.0 dose of VC is 26.14 and the mean of teeth growth after taking 2.0 dose of OJ is 26.06. Both of them are within the confidence interval. We fail to reject the null hypothesis.

4. State your conclusions and the assumptions needed for your conclusions.

The conclusion is when the dose is 0.5 or 1.0 there is a difference between the teeth growth after taking OJ and VC, while when the dose is 2.0, there is no difference between the teeth growth after taking OJ and VC. The assumption needed is we first assumed the whole population is normally distributed, then we assumed the population is normally distributed under each dose.