Comparing the exponential distribution in R and the Central Limit Theorem

Daniel

21/1/2021

Overview

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. It will be investigated the distribution of averages of 40 exponential doing a thousand simulations.

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Simulations

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
library(knitr)

#defining variables
n_simulations <- 1000
n <- 40
lambda <-  0.2

sim_data <- matrix(rexp(n= n_simulations*n,rate=lambda), n_simulations, n) #creating a matrix with the
sample_mean <- rowMeans(sim_data) #getting the mean of each row of the matrix
```

Simulated Data (sample_mean) VS Theoretical Mean

We are going to calculate the mean of the data simulated and the theoretical one in order to verify some differences.

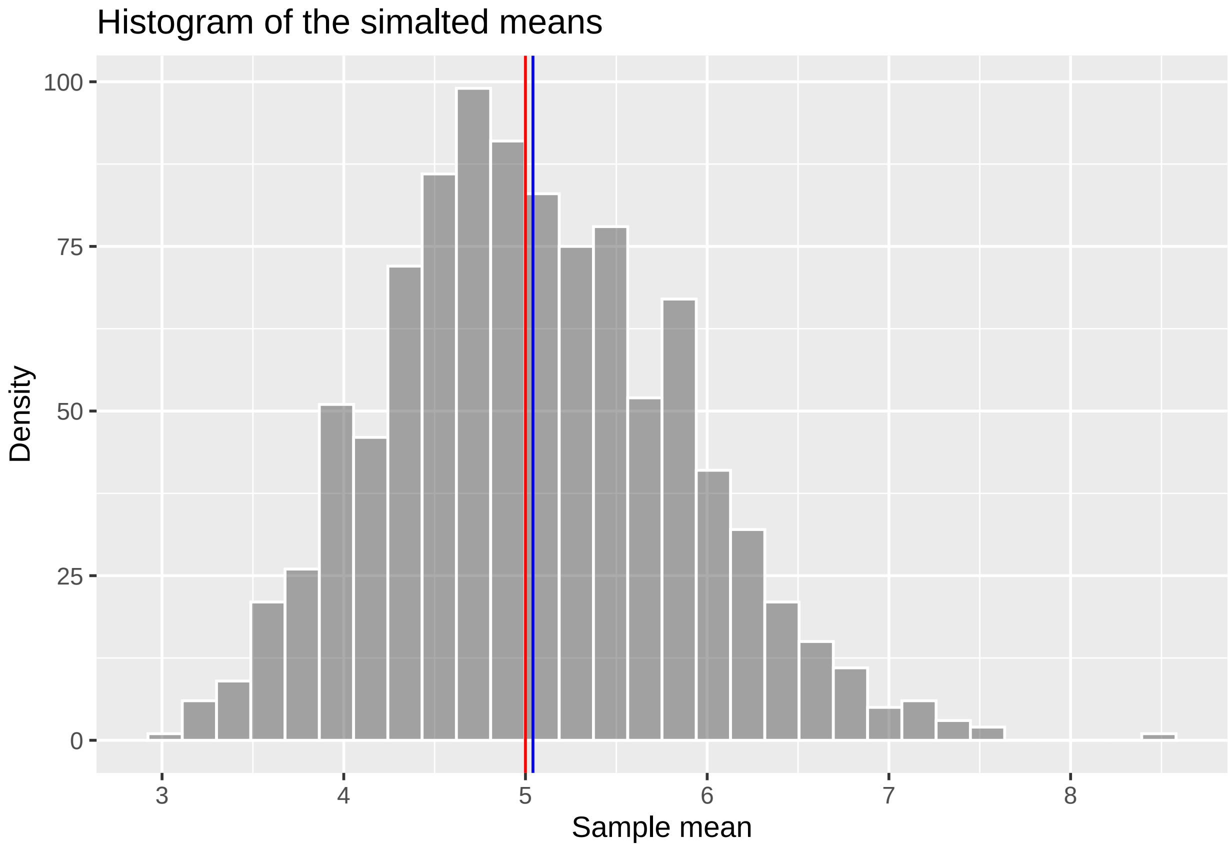
```r
simulated_mean <- mean(sample_mean)
theoretical_mean <- 1/ lambda

dataframe <-data.frame("Mean"=c(simulated_mean, theoretical_mean),
                     row.names = c("Mean from simulated data ","Theoretical mean"))

dataframe
```

```
##                                Mean
## Mean from simulated data  5.041252
## Theoretical mean          5.000000
```

You can notice there is a minimum difference between the simulated and theoretical means.Lets see this in more detail with the graphic below.

```r
simulatedMean_data <- as.data.frame (sample_mean)

ggplot(simulatedMean_data, aes(sample_mean))+
  geom_histogram(alpha=.5, position="identity", col="white")+
  geom_vline(xintercept = theoretical_mean, colour="red",show.legend=TRUE)+
  geom_vline(xintercept = simulated_mean, colour="blue", show.legend=TRUE)+
  ggtitle ("Histogram of the simalted means ")+xlab("Sample mean")+ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

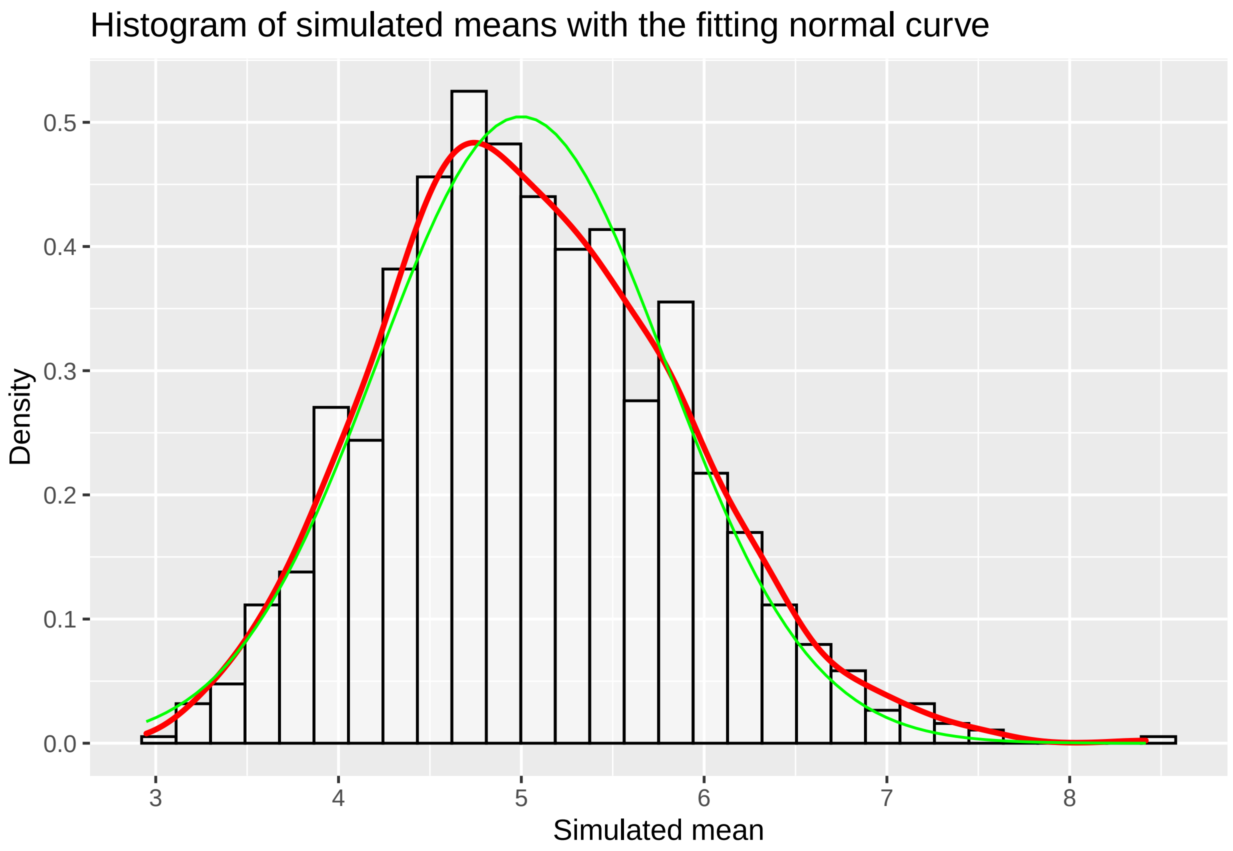


Sample Variance VS Theorical Variance

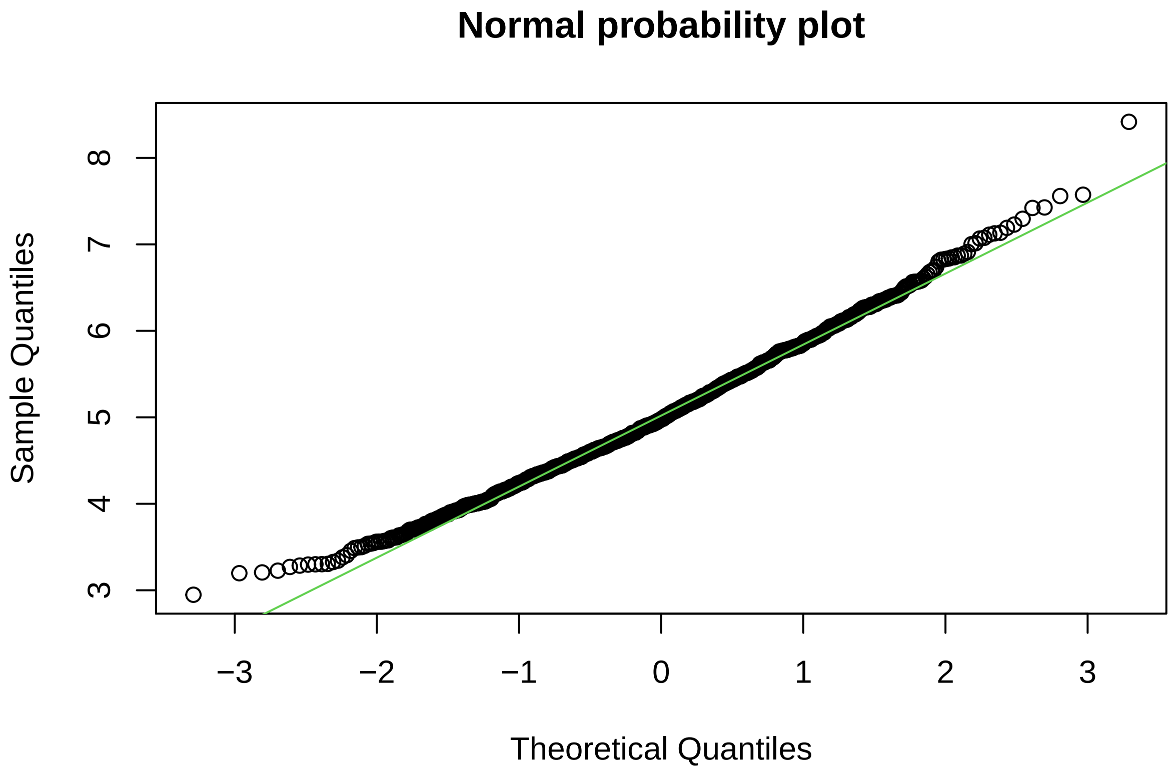
Let´s calculate the the variance based on the simulated data and compare it with the theoretical one.

```
simulated_variance <- var(sample_mean)

theoretical_variance <- (1/ lambda)^2 /n

result2 <-data.frame("Variance"=c(simulated_variance, theoretical_variance),
                     row.names = c("Variance from the sample ","Theoretical variance"))

ggplot(simulatedMean_data, aes(sample_mean))+
  geom_histogram(aes(y=..density..), alpha=.5, position="identity", fill="white", col="black")+
  geom_density(colour="red", size=1)+
  stat_function(fun = dnorm, colour = "green", args = list(mean = theoretical_mean, sd = sqrt(theoretica
  ggtitle ("Histogram of simulated means with the fitting normal curve ")+
  xlab("Simulated mean")+
  ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
qqnorm(sample_mean, main ="Normal probability plot")
qqline(sample_mean,col = "3")
```



The histogram and the normal probability plots show a normal distributions.

Part 2

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

```
data(ToothGrowth)
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 ...
```

```
head(ToothGrowth, 5)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
```

2. Provide a basic summary of the data

```
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
```

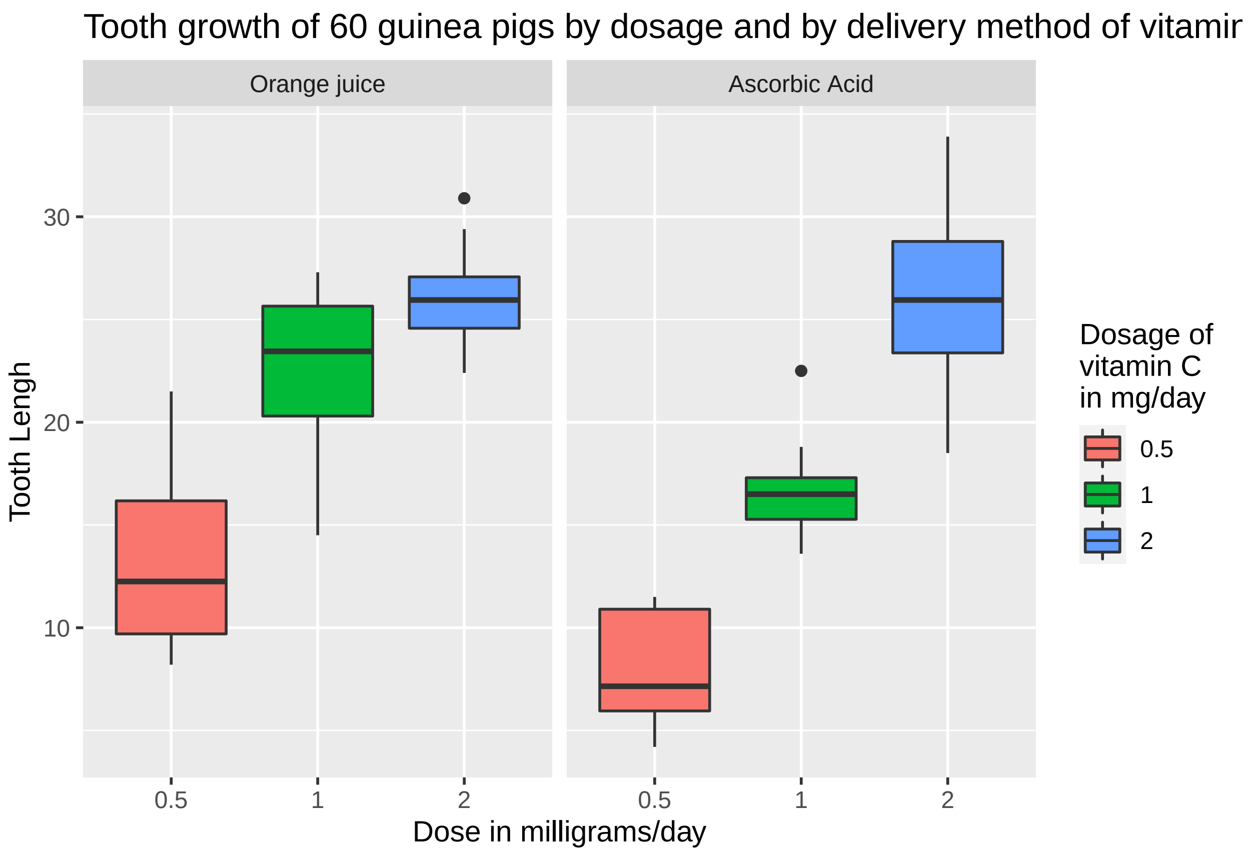
We will applu the “tapply” function taking advantage of the factor variable “supp”. It means, getting the mean of “len” by each “supp”.

```
tapply(ToothGrowth$len,ToothGrowth$supp, mean)
```

```
##       OJ       VC 
## 20.66333 16.96333
```

Now, look some tendencies in more detail with the graphic below.

```
ggplot(ToothGrowth, aes(factor(dose), len, fill = factor(dose))) +
      geom_boxplot() +
      facet_grid(.~supp, labeller = as_labeller(
            c("OJ" = "Orange juice",
              "VC" = "Ascorbic Acid"))) +
      labs(title = "Tooth growth of 60 guinea pigs by dosage and by delivery method of vitamin C",
           x = "Dose in milligrams/day",
           y = "Tooth Lengh") +
      scale_fill_discrete(name = "Dosage of\nvitamin C\nin mg/day")
```



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

We are going to apply the “t.test” function to each dose (0.5, 1 y 2) by each “supp” (OJ and VC) as it is shown below.

```r
# Comparison by delivery method for the same dosage
t05 <- t.test(len ~ supp,
       data = rbind(ToothGrowth[(ToothGrowth$dose == 0.5) &
                                  (ToothGrowth$supp == "OJ"),],
                    ToothGrowth[(ToothGrowth$dose == 0.5) &
                                  (ToothGrowth$supp == "VC"),]),
       var.equal = FALSE)

t1 <- t.test(len ~ supp,
       data = rbind(ToothGrowth[(ToothGrowth$dose == 1) &
                                  (ToothGrowth$supp == "OJ"),],
                    ToothGrowth[(ToothGrowth$dose == 1) &
                                  (ToothGrowth$supp == "VC"),]),
       var.equal = FALSE)

t2 <- t.test(len ~ supp,
       data = rbind(ToothGrowth[(ToothGrowth$dose == 2) &
                                  (ToothGrowth$supp == "OJ"),],
                    ToothGrowth[(ToothGrowth$dose == 2) &
                                  (ToothGrowth$supp == "VC"),]),
       var.equal = FALSE)

# Summarizing the three tests as follows

summary <- data.frame(
      "p-value" = c(t05$p.value, t1$p.value, t2$p.value),
      "Conf.Low" = c(t05$conf.int[1],t1$conf.int[1], t2$conf.int[1]),
      "Conf.High" = c(t05$conf.int[2],t1$conf.int[2], t2$conf.int[2]),
      row.names = c("Dosage .05","Dosage 1","Dosage 2"))

summary
```

```
##                 p.value  Conf.Low Conf.High
## Dosage .05 0.006358607  1.719057  8.780943
## Dosage 1   0.001038376  2.802148  9.057852
## Dosage 2   0.963851589 -3.798070  3.638070
```

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

We can reject de null hypothesis (the means between two groups are equal), stating there is no difference between the 0.5 and 1 mgs/day for tooth growth. This means the the delivery method does matter to define the tooth growth.

We fail to reject the null hypothesis, stating that there is no difference in the tooth growth by the delivery method for 2 milligrams/day. We observe p-values more than the treshold of .05 and the confidence levels include 0. So, for dosage of 2 milligrams/day the delivery method doesn't matter.