Statistical Inference Course Project

JB

2023-10-04

Overview

This report investigates the exponential distribution inR in comparison to Central Limit Theorem in part 1 as well as shows basic Inferential Data analysis on the ToothGrowth data set in part 2.

```
# Loading necessary packages
library(datasets)
library(ggplot2)
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

Part 1: Simulation exercise.

This exercise starts with generating 1000 simulations of 40-size samples of exponential distribution. The simulation is done with a rexp() function in R. The mean of each simulation is calculated using the apply() function.

```
lambda<- 0.2
n<- 40
simData<- replicate(1000, rexp(n,lambda))

simMeans<- apply(simData, 2, mean)
```

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

```r
sampleMean<- mean(simMeans)
sampleMean
```

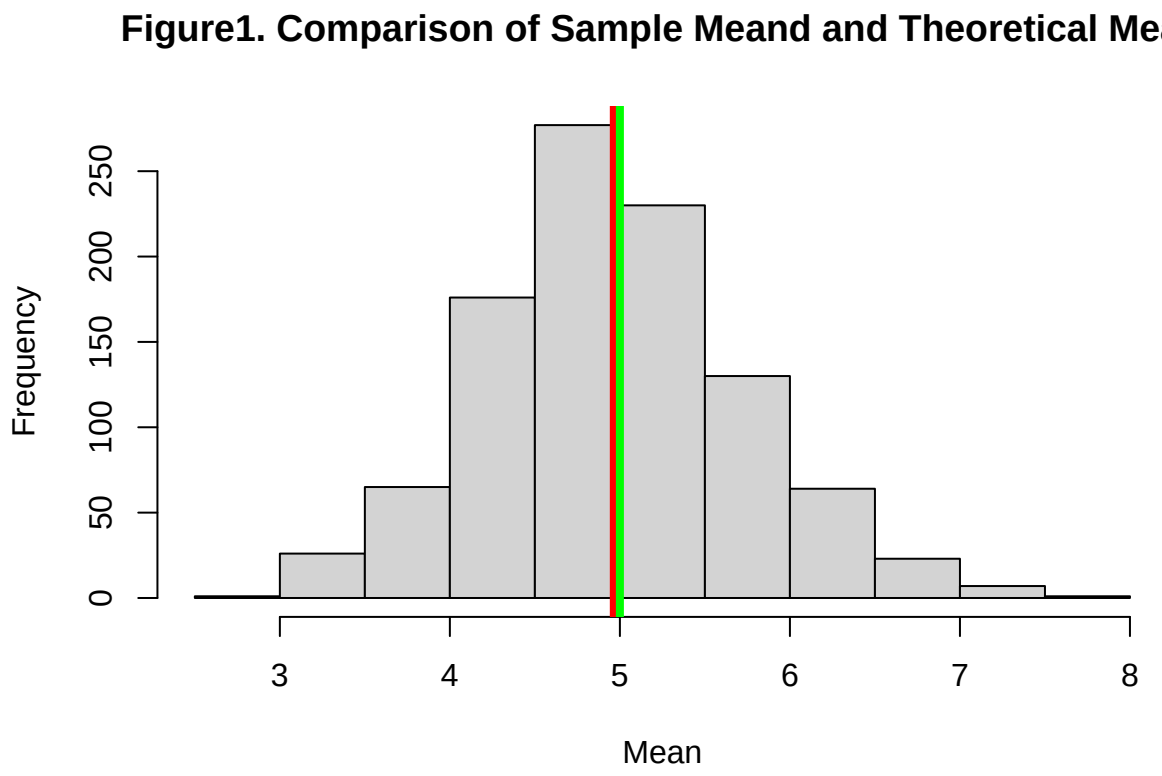
```
## [1] 4.964276
```

```r
theoreticalMean<- 1/lambda
theoreticalMean
```

```
## [1] 5
```

2. Plot Histogram showing the distribution of 1000 mean values calculated for each simulation.

```r
hist(simMeans, main = "Figure1. Comparison of Sample Meand and Theoretical Mean", xlab="Mean")
abline(v= sampleMean, lw = 4, col= "red")
abline(v= theoreticalMean, lw = 4, col = "green")
```



The red line indicates the simulated data mean, whereas the green the theoretical one.

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

```r
sampleSD<- sd(simMeans)
sampleSD
```

```
## [1] 0.7602755
```

```
theoreticalSD<- (1/lambda)/sqrt(n)
theoreticalSD
```

```
## [1] 0.7905694
```

```
sampleVar<- sampleSD^2
sampleVar
```

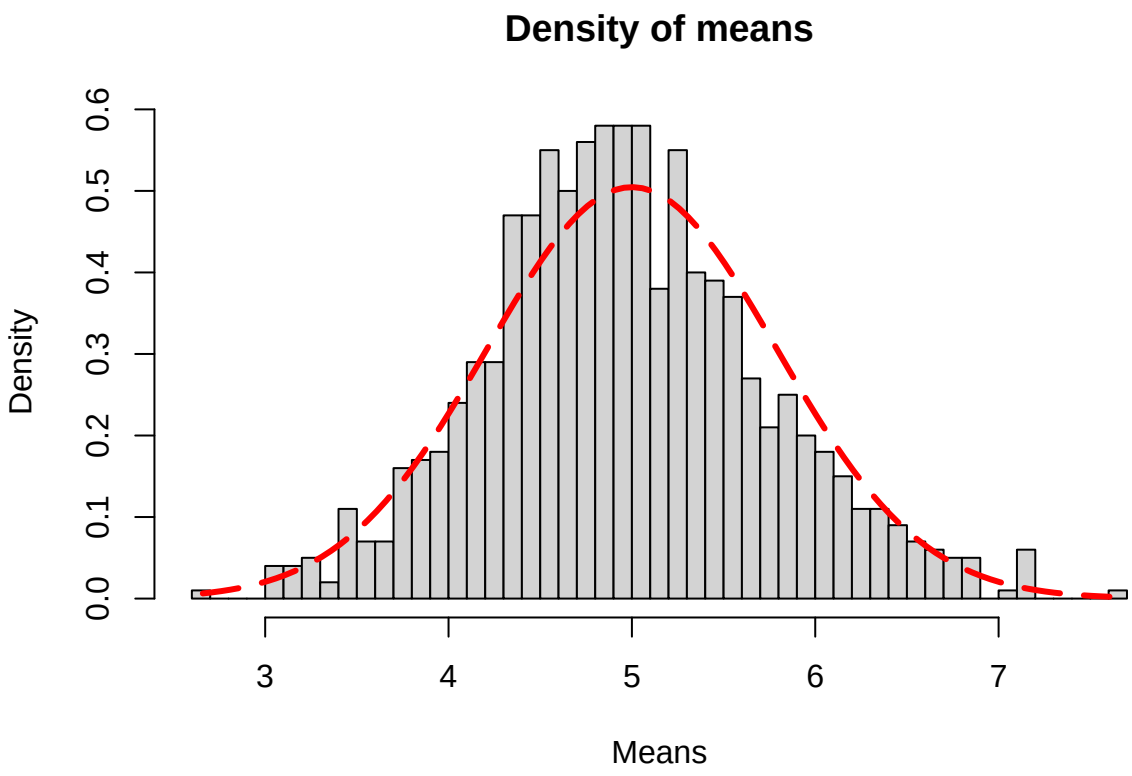
```
## [1] 0.5780188
```

```
theoreticalVAr<- theoreticalSD^2
theoreticalVAr
```

```
## [1] 0.625
```

4. Show that the distribution is approximately normal.

```
xfit<- seq(min(simMeans), max(simMeans), length=100)
yfit<- dnorm(xfit, mean = 1/lambda, sd= (1/lambda)/sqrt(n))
hist(simMeans, breaks=n, prob = T, main = "Density of means", xlab = "Means", ylab = "Density" )
lines(xfit, yfit, pch = 22, lw=3, lty = 5, col="red")
```



The red curve indicates the normal distribution

Summary between the two distribution can be seen below

```
##             Variable Simulated Theoretical
## 1               Mean 4.9642756   5.0000000
## 2 Standard Deviation 0.7602755   0.7905694
## 3           Variance 0.5780188   0.6250000
```

Conclucion As we can see from the table all of mean, standard deviation and variance are closely similar. Due to the CLT the distribution of averages of 40 exponential is closely similar to a normal distribution.

Part 2: Basic Inferential Data Analysis.

```
data("ToothGrowth")
dt<- ToothGrowth
unique(dt$dose)
```

```
## [1] 0.5 1.0 2.0
```

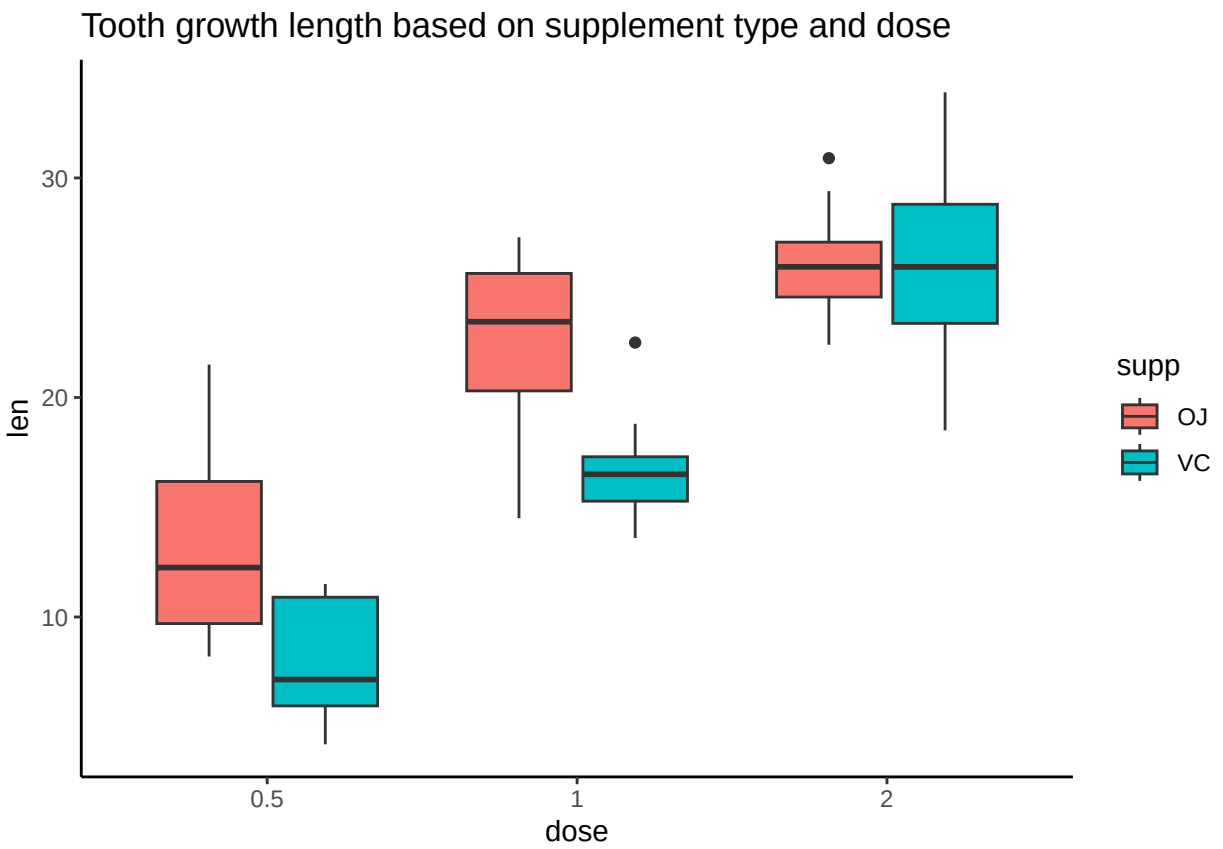
There are three unique values for dose which I will convert to factors.

```
dt$dose<- factor(dt$dose)
str(dt)
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

1. Plotting box plot to visualise the difference between two supplement types and dosage.

```
ggplot(dt, aes(x=dose, y=len, fill = supp))+
          geom_boxplot()+
          ggtitle("Tooth growth length based on supplement type and dose")+
          theme_classic()
```



From the plot we see that for the first two doses there seem to be an observable difference in tooth growth between the two supplemtents. We also see that with increased dose the tooth growth length increases. Last conclusiion we can make is that for 0.5 and 1 dose OJ supplememnt seems to be performing better whereas for 2 dose there isnt much difference between OJ and VC.

2. Calculating mean difference between supplements for all three doses.

```
dt %>%
          group_by(supp, dose)%>%
          summarise(mean = mean(len), .groups = "drop")%>%
          spread(supp, mean) %>%
          mutate(diff = abs(VC-OJ))
```

```
## # A tibble: 3 x 4
##   dose     OJ    VC   diff
##   <fct> <dbl> <dbl>  <dbl>
## 1 0.5    13.2  7.98 5.25
## 2 1      22.7 16.8  5.93
## 3 2      26.1 26.1  0.0800
```

Only for dose 2 the difference between two supplememnts is really smallmeaning its harder to comapre their effectiveness.

3. T test hypothesis for all doses.

Null hypothesis is that there is no significant difference between OJ and VC. Althernative hypothesis says there is a difference between the two drugs. We set alpha for standard 0.05.

```
# Filtering data for testing
dose_half<- filter(dt, dose==0.5)
dose_one<- filter(dt, dose==1)
dose_two<- filter(dt, dose==2)
# t-test for 0.5 dose
t.test(len~supp, dose_half)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
# t-test for 1 dose
t.test(len~supp, dose_one)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

```
# t-test for 2 dose
t.test(len~supp, dose_two)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

```r
# Table summarising the tests results
table_data <- data.frame(
          Dose = c(0.5, 1, 2),
          p_value = c(0.006359, 0.001038, 0.9639),
          Conf.Int = c("1.719057 8.780943", "2.802148 9.057852", "-3.79807 3.63807"),
          Decision = c("Reject Null", "Reject Null", "Do not Reject Null")
)

print(table_data)
```

```
##   Dose  p_value          Conf.Int           Decision
## 1  0.5 0.006359 1.719057 8.780943        Reject Null
## 2  1.0 0.001038 2.802148 9.057852        Reject Null
## 3  2.0 0.963900  -3.79807 3.63807 Do not Reject Null
```

Conclusion As expected the p-value for dose two is significantly bigger than for dose 1 and 0.5 meaning we cannot reject the hypothesis that supplements OJ and VC are different. We can't strictly say that those two supplements have a different effect on tooth growth. The conclusion is made under assumption that the data isn't paired and the variance are different.