Statistical Inference Course Project

Sachin Sharma

10/3/2021

Peer Graded Assignment: Statistical Inference Course Project

Instructions

The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

Part 1: Simulation Exercise Instructions

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be 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. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Question 1 : Show the sample mean and compare it to the theoretical mean distribution

```
n <- 40
Simulations <- 1000
Lambda <- 0.2

SampleMean <- NULL
for(i in 1:Simulations) {
  SampleMean <- c(SampleMean, mean(rexp(n, Lambda)))
}
mean(SampleMean)
```

```
## [1] 4.996443
```

Here we can see that compared to the theoretical mean distribution of 5 , our mean 4.99 is very close to 5 .

Question 2: Show the sample is (via variance) and compare it to the thoretical variance of the distribtution.

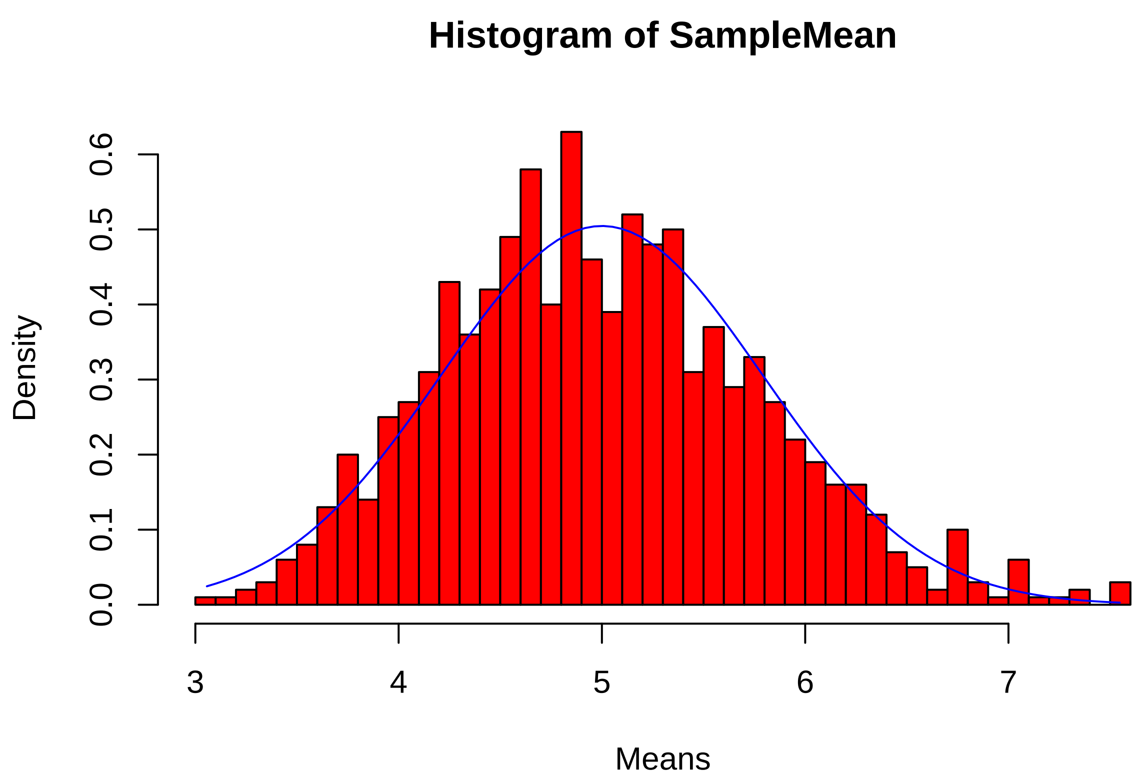
The theoretical standard deviation of the distribution is also 1/lambda , for a lambda of 0.2 , equates to 5 . We know that the variance is the square of the standard deviation, which is 25 .

```
Variance <- var(SampleMean)
```

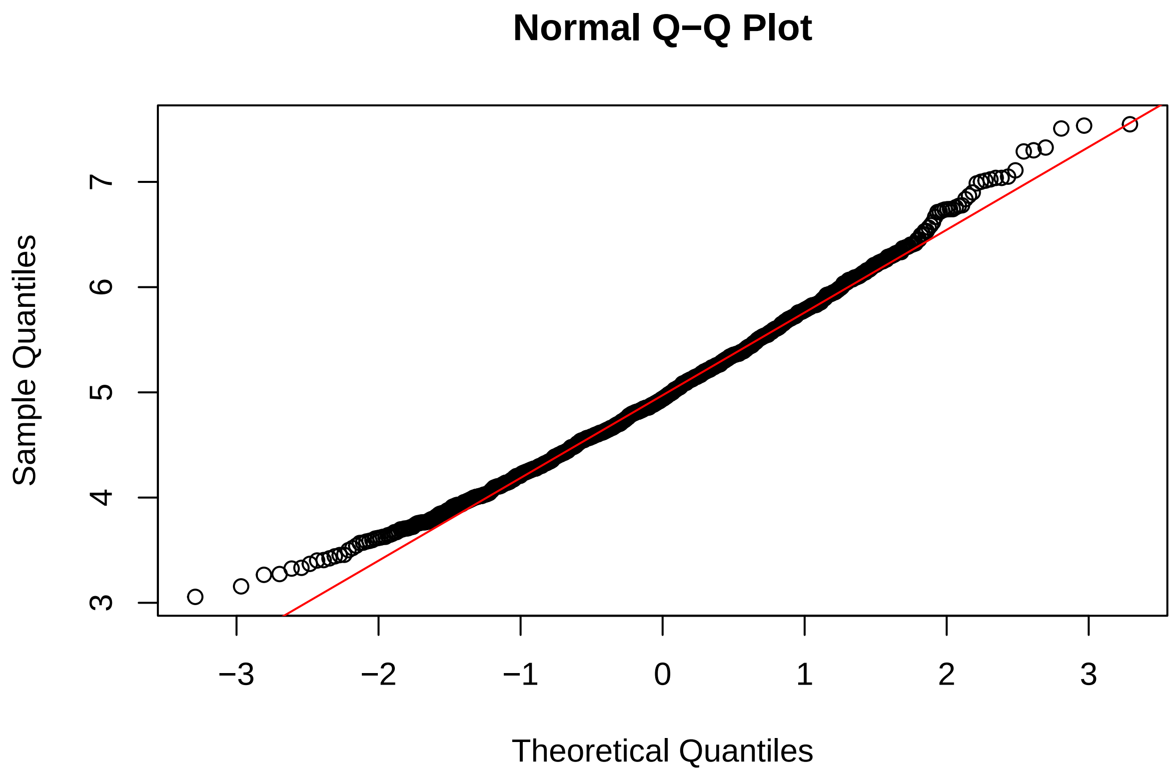
0.6 is close to the theoretical distribution.

Show that the distribution is appoximately normal

```
hist(SampleMean, breaks = n, prob = T, col = "red", xlab = "Means")
x <- seq(min(SampleMean), max(SampleMean), length = 100)
lines(x, dnorm(x, mean = 1/Lambda, sd = (1/Lambda/sqrt(n))), pch = 25, col = "blue")
```




```
qqnorm(SampleMean)
qqline(SampleMean, col = "red")
```



The distribution averages of 40 exponentials is very close to a normal distribution

Part 2: Basic Inferential Data Analysis Instructions

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

Load the ToothGrowth data and perform some basic exploratory data analysis

Importing the data

```
library(datasets)
data(ToothGrowth)
library(ggplot2)

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 ...
```

Visualizing the first few rows of the dataframe by using head() function

```
head(ToothGrowth)
```

```
##    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
## 6 10.0   VC  0.5
```

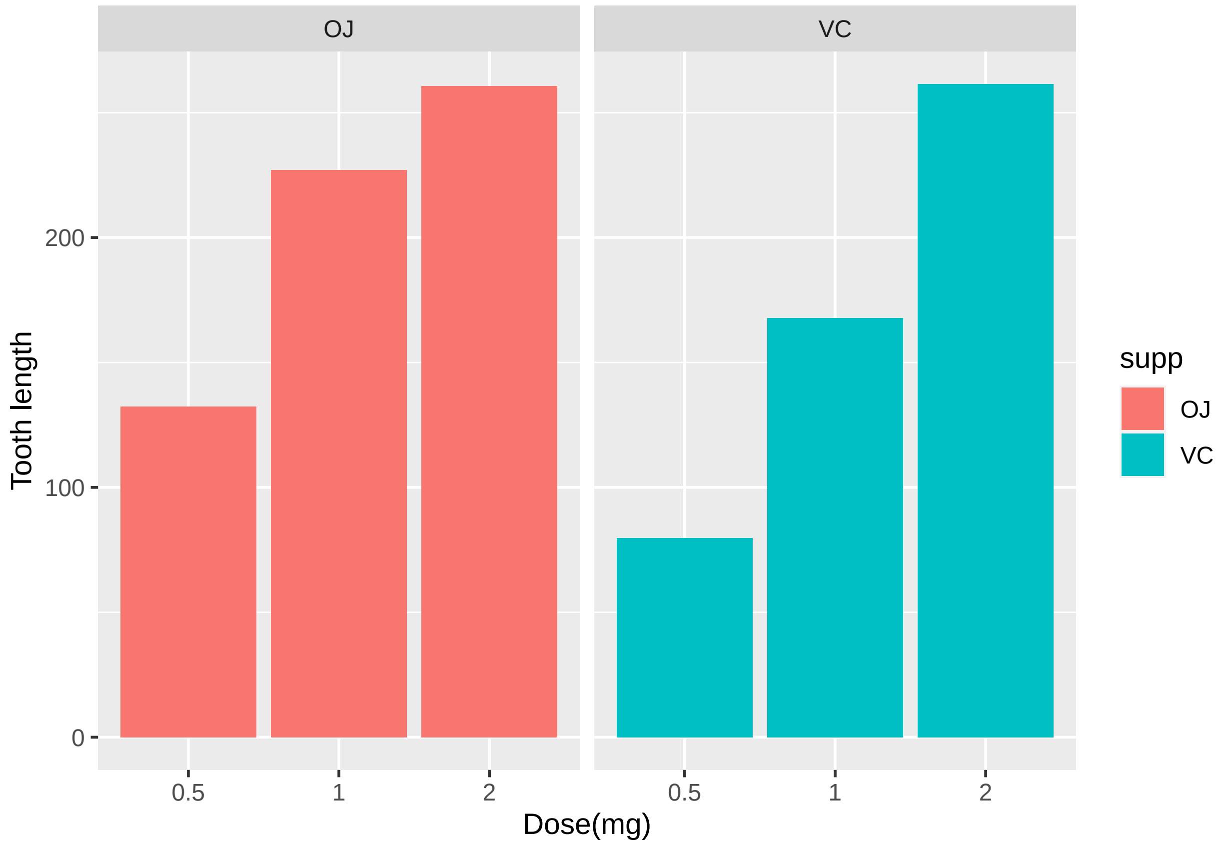
Checking the summary of the dataframe using summary() function

```
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
```

Visualizing the dataframe using ggplot() with the help of bar chart

```
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
    geom_bar(stat="identity") +
    facet_grid(. ~ supp) +
    xlab("Dose(mg)") +
    ylab("Tooth length")
```



Doing hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)

```
hypo_thesis <- t.test(len ~ supp, data = ToothGrowth)
hypo_thesis$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

```
hypo_thesis$p.value
```

```
## [1] 0.06063451
```

```
hypo_thesis_1<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 0.5))
hypo_thesis_1$conf.int
```

```
## [1] 1.719057 8.780943
## attr(,"conf.level")
## [1] 0.95
```

```
hypo_thesis_1$p.value
```

```
## [1] 0.006358607
```

```
hypo_thesis_2<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 1))
hypo_thesis_2$conf.int
```

```
## [1] 2.802148 9.057852
## attr(,"conf.level")
## [1] 0.95
```

```
hypo_thesis_2$p.value
```

```
## [1] 0.001038376
```

```
hypo_thesis_3<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 2))
hypo_thesis_3$conf.int
```

```
## [1] -3.79807  3.63807
## attr(,"conf.level")
## [1] 0.95
```

```
hypo_thesis_3$p.value
```

```
## [1] 0.9638516
```

Conclusions

1. **OJ ensures more tooth growth than VC for dosages 0.5 & 1.0.**
2. **OJ and VC givesthe same amount of tooth growth for dose amount 2.0 mg/day.**
3. **For the entire trail we cannot conclude OJ is more effective that VC for all scenarios.**