# Statistical Inference Course Project

*Saurabh Thakur*

*10 June 2018*

## Peer Graded assignment.

```
setwd("E:/R/coursera/Assignments/Statistical Inference")
set.seed(18746)
```

## Instructions

**There are two sections of the project**

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

## 1. A simulation exercise.

**Overview**

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.

**Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.**

**1. Show the sample mean and compare it to the theoretical mean of the 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] 4.982101
```

```
theo_mean <- 1 / Lambda
theo_mean
```

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

Looking at the values of both sample mean and theoretical mean we see that they are very close.

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

Sample Variance is the variance of the Sample mean.

```
variance<- var(SampleMean)
variance
```

```
## [1] 0.6291304
```

Theoretical variance of exponential distribution is (lambda*sqrt(n))^-2

```
theo_var <- (Lambda*(sqrt(n)))^-2
theo_var
```

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

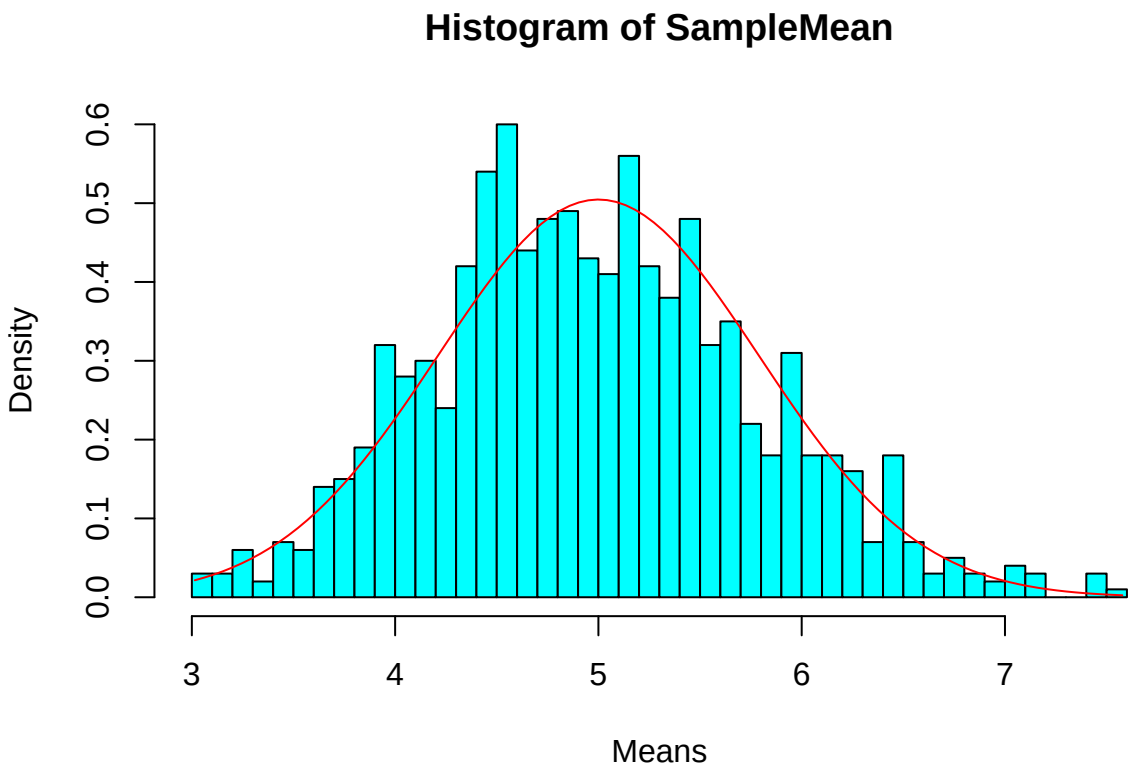
Looking at the values fo both Sample variance and Theoretical variance we see that they both are very close.

**3. Show that the distribution is approximately normal.**

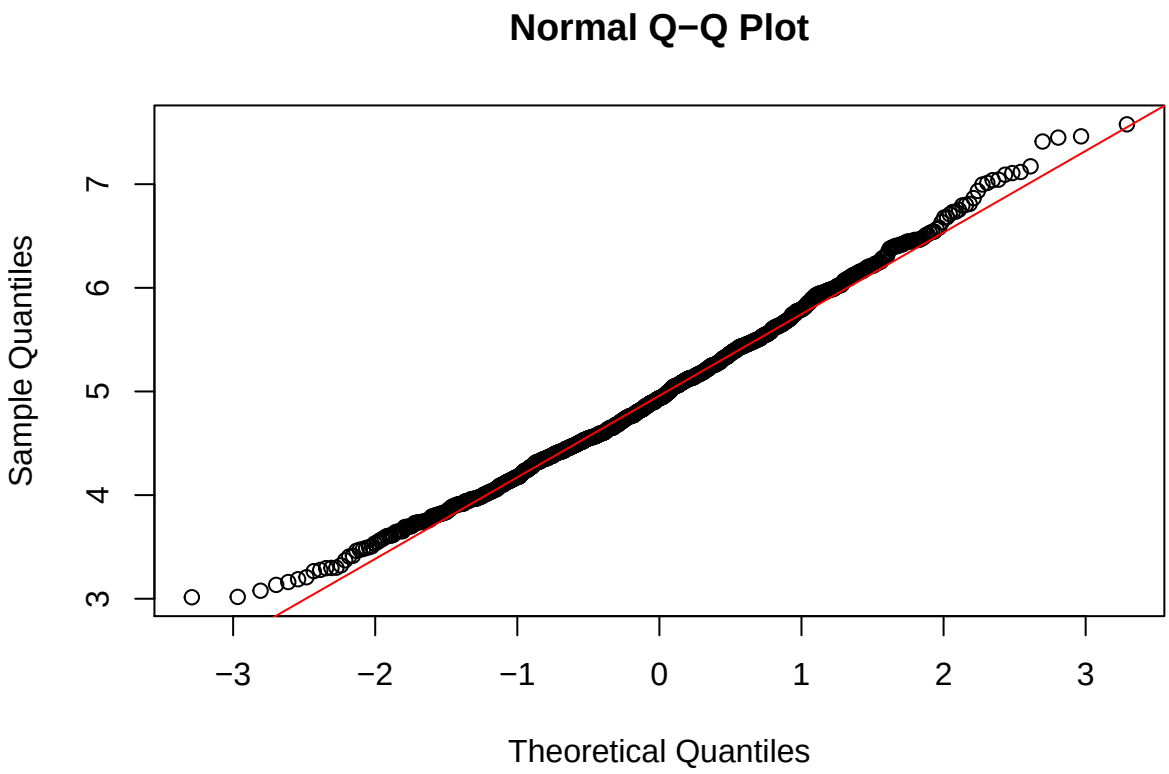
```
hist(SampleMean,
     breaks = n,
     prob = T,
     col = "cyan",
     xlab = "Means")

x <- seq(min(SampleMean), max(SampleMean), length = 100)

lines(x,
      dnorm(x,
            mean = 1/Lambda,
            sd = (1/Lambda/sqrt(n))),
      pch = 25,
      col = "red")
```




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




## 2. Basic inferential data analysis.

### Overview

The purpose of the this data analysis is to analyze the ToothGrowth data set by comparing the guinea tooth growth by supplement and dose. I will do the exploratory data analysis on the data set and then will compare with confidence intervals in order to make conclusions about the tooth growth.

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

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

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

### ..2 Provide a basic summary of the data.

```
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 ...
```

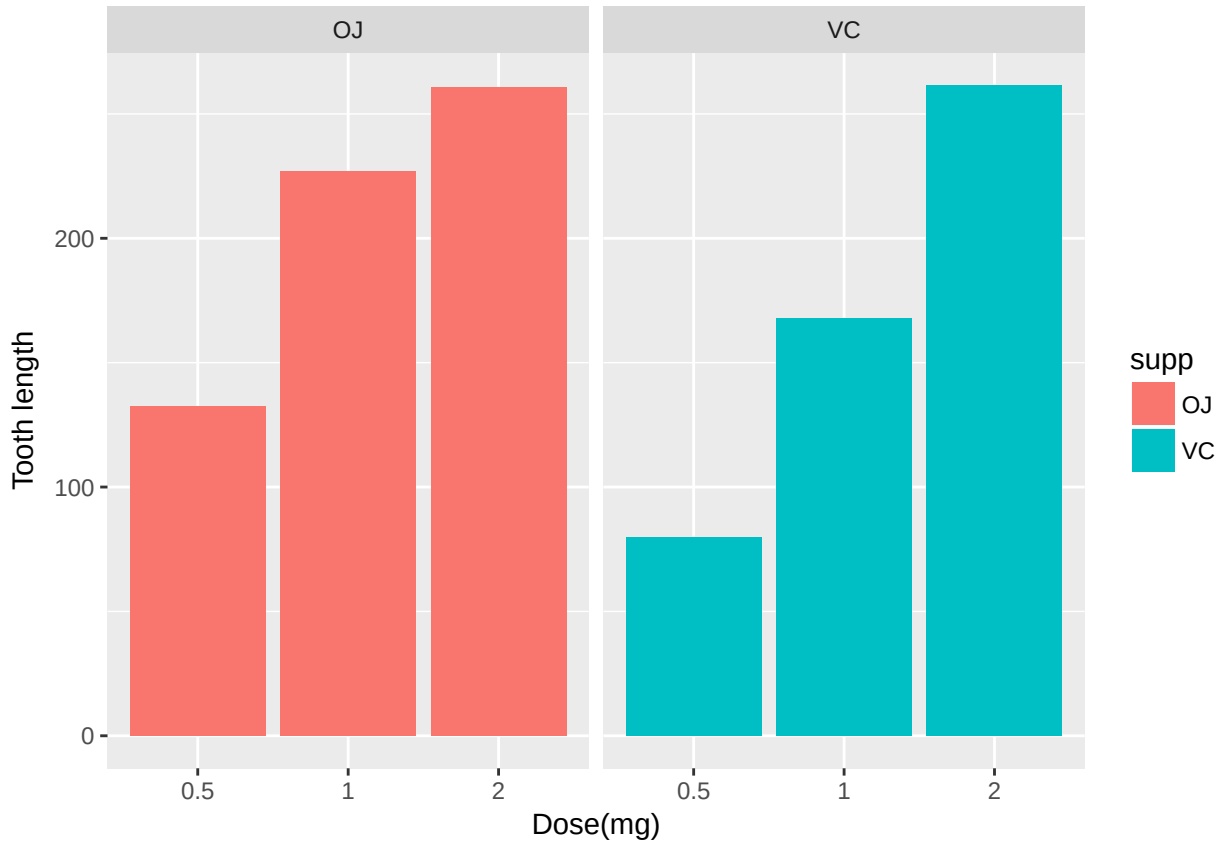
```
head(ToothGrowth)
```

```
##    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
```

```
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
```

```
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")
```




The box plots seem to show, increasing the dosage increases the tooth growth. Orange juice is more effective than ascorbic acid for tooth growth when the dosage is .5 to 1.0 milligrams per day. Both types of supplements are equally as effective when the dosage is 2.0 milligrams per day.

**..3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering).**

**Hypothesis #1**

**Orange juice & ascorbic acid deliver the same tooth growth across the data set.**

```
hypoth1 <- t.test(len ~ supp, data = ToothGrowth)
hypoth1$conf.int
```

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

```
hypoth1$p.value
```

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

*The confidence intervals includes 0 and the p-value is greater than the threshold of 0.05. The null hypothesis cannot be rejected.*

**Hypothesis #2**

**For the dosage of 0.5 mg/day, the two supplements deliver the same tooth growth.**

```
hypoth2<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 0.5))
hypoth2$conf.int
```

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

```
hypoth2$p.value
```

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

*The confidence interval does not include 0 and the p-value is below the 0.05 threshold. The null hypothesis can be rejected. The alternative hypothesis that 0.5 mg/day dosage of orange juice delivers more tooth growth than ascorbic acid is accepted.*

### Hypothesis #3

**For the dosage of 1 mg/day, the two supplements deliver the same tooth growth**

```
hypoth3<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 1))
hypoth3$conf.int
```

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

```
hypoth3$p.value
```

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

*The confidence interval does not include 0 and the p-value is smaller than the 0.05 threshold. The null hypothesis can be rejected. The alternative hypothesis that 1 mg/day dosage of orange juice delivers more tooth growth than ascorbic acid is accepted.*

### Hypothesis #4

**For the dosage of 2 mg/day, the two supplements deliver the same tooth growth**

```
hypoth4<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 2))
hypoth4$conf.int
```

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

```
hypoth4$p.value
```

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

*The confidence interval does include 0 and the p-value is larger than the 0.05 threshold. The null hypothesis cannot be rejected.*

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

### Conclusions & assumptions

- Orange juice delivers more tooth growth than ascorbic acid for dosages 0.5 & 1.0.
- Orange juice and ascorbic acid deliver the same amount of tooth growth for dose amount 2.0 mg/day.

- For the entire data set we cannot conclude orange juice is more effective that ascorbic acid.

**Assumptions**

- Normal distribution of the tooth lengths
- No other unmeasured factors are affecting tooth length

**END**