---

title: "Statistical Inference Final Project" author: "Dimuthu Attanayake" date: "5/10/2020" output: html_document

Overview

This is the two part final assignment for Statistical Inference course offered by the Johns Hopkins University on Coursera. The Course is part of the ten course specialization on data science offered by the university.

Part 1: Simulating the exponential distribution in R and then comparing it with the Central Limit Theoram

Part 2:Basic inferential data analysis using the ToothGrowth data in the R datasets package

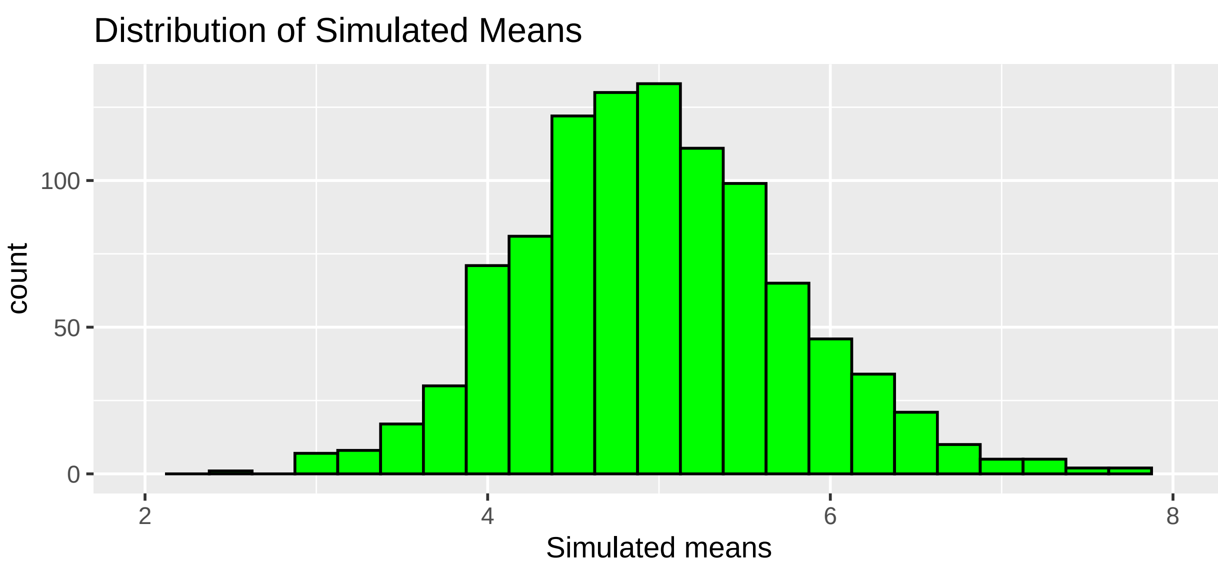
Part 1

Instructions: 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.

Running 1000 simulations on distribution averages of 40 exponentials with lambda = 0.2

```
n <- 40
set.seed(90)
nosim <- 1000
lambda <- 0.2
sim <- replicate(nosim, rexp(n,lambda))
sim_mean <-  apply(sim, 2, mean)
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



1.The sample mean and the theoritical mean of the distribution

```
sam_mean <- mean(sim_mean)
sam_mean
```

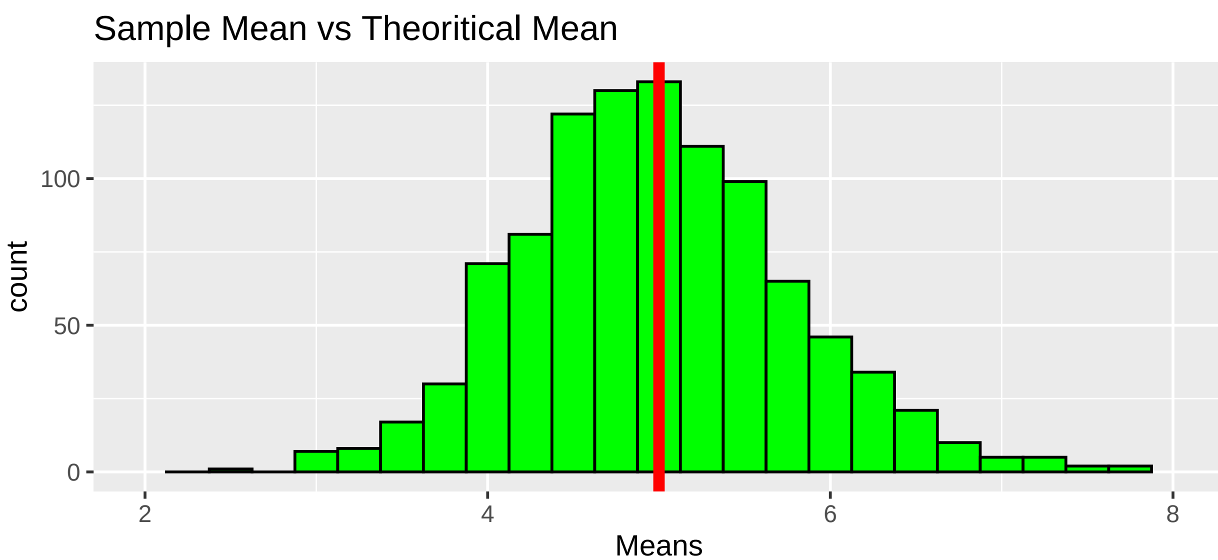
```
## [1] 4.977191
```

```r
theo_mean <- 1/lambda
theo_mean
```

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

Therefore, sample mean = 4.999702 is almost equal to theoritical mean = 5.

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



2.The sample variance and the theoritical variance of the distribution

```r
sam_var <- var(sim_mean)
sam_var
```

```
## [1] 0.6132061
```

```r
sam_sd <- sd(sim_mean)
sam_sd
```

```
## [1] 0.7830748
```

```r
theo_var <- (1 / lambda)^2 / (n)
theo_var
```

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

```r
theo_sd <- (1 / lambda)/(sqrt(n))
theo_sd
```

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

Sample variance = 0.6432422 and sample standard deviation is = 0.8020251.Theoritical variance = 0.625 is closer to the sample variance but slighly smaller. Therefore, the sample displays a comparatively higher variance.

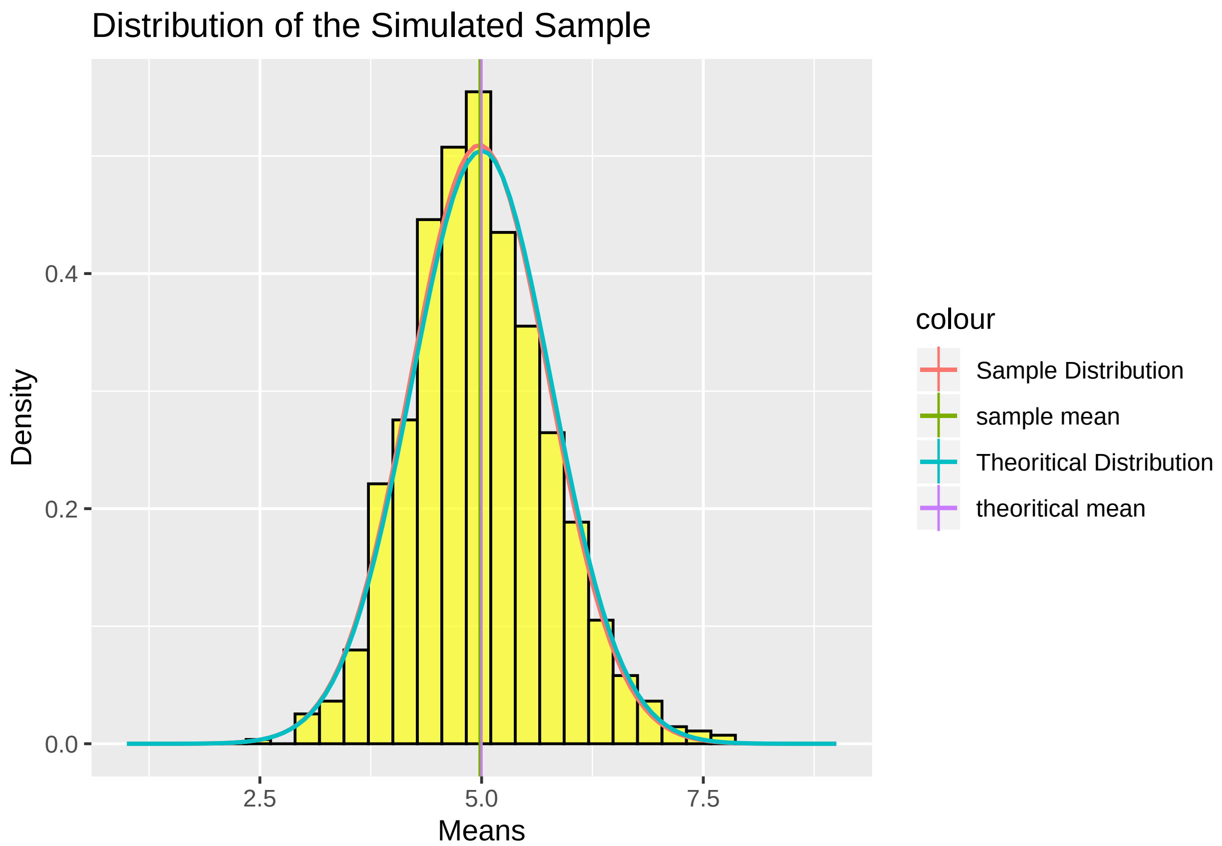
3.To show that the distribution is approximately normal

```r
library(ggplot2)
histdata <- data.frame(sim_mean)
g <- ggplot(histdata,aes(x=sim_mean) ) +
        geom_histogram(aes(y= ..density..), color="black" , fill = "yellow", alpha= 0.65 ) + xlim(c(1,9
```

```r
g <- g + labs( title = "Distribution of the Simulated Sample", x= "Means", y= "Density")
g <- g + stat_function(fun = dnorm, args = list(mean = sam_mean, sd = sam_sd), aes(color = "Sample Dist
g <- g + stat_function(fun = dnorm, args = list(mean = theo_mean, sd = theo_sd), aes(color = "Theoritica
g <- g + geom_vline(aes(xintercept = sam_mean, color= "sample mean"), size = 0.40)
g <- g + geom_vline(aes(xintercept = theo_mean, color= "theoritical mean"), size = 0.40)
g
```

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

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



As shown by the diagram, the distribution of the simulated sample follows an approximately normal distribution, and is very similar to the theoritical distribution. The mean of the sample and the theoritical distibutions lie closer together, at the approximate midpoint of the distrubution. Therefore, the distribution of the mean of 40 exponentials, simulated 1000 times is approximately normal for the given lambda, in line with the Central Limit Theoram which states that"distribution of avaerages of normalized iid variables becomes that of a standard normal as the sample size increases."

##Part 2

1.Load the ToothGrowth data, perform some basic exploratory data analyses and provides a basic summary of the data.

Tooth growth data set provides "the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid)" to measure the effect of vitamin C on the toothgrowth of guinea pigs.

```r
data(ToothGrowth)
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
```

```r
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
```

```r
nrow(ToothGrowth)
```

```
## [1] 60
```

```r
ncol(ToothGrowth)
```

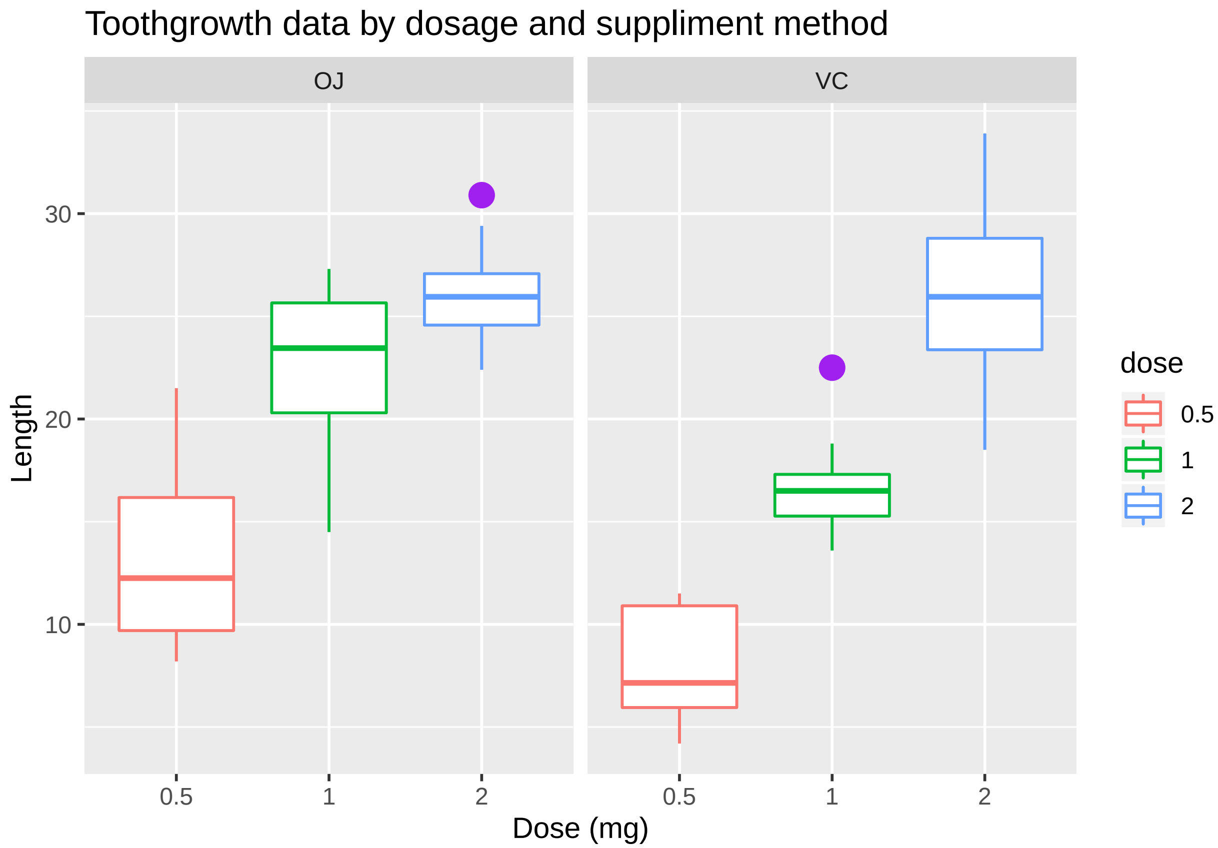
```
## [1] 3
```

```r
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 ...
```

The data set has three variables len(mean length of tooth growth), supp(type of suppliment) and dose(dosage of each suppliment). The data set has 60 observations, 60 rows and 3 columns.

```r
#convert dose into a factor variable
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
#boxplot indicating mean length of tooth growth vs dosage for different suppliment methods
library(ggplot2)
g <- ggplot(ToothGrowth, aes(x=dose, y=len, color =dose))+ geom_boxplot(aes(fill = dose))
g <- g + geom_boxplot(outlier.colour="purple",outlier.size=4) + facet_wrap(~ supp)
g <- g + labs(title="Toothgrowth data by dosage and suppliment method",x="Dose (mg)", y = "Length")
g
```



The boxplot shows that as the dosage increases, the mean tooth growth increases. At lower doses, orange juice appears to be a more effective suppliment, while at higher dose =2, vitamin C appears to be more effective.

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

Hypothesis testing will be used to determine whether the difference between mean tooth growth based on suppliment type and dosage is statistically significant.

Hypothesis test to compare mean tooth growth by supplement type(orange juice/vitamin C)

Ho: The difference in means equals zero vs HA: The difference in means is not equal to zero.

```
t.test(len~supp, data=ToothGrowth, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

Since p value is greater than 0.05, do not reject the null hypothesis.

Hypothesis test to compare tooth growth by dose

Difference in mean tooth growth is compared between maximum and minimum doses. From the summary data, it can be seen that the maximum dosage = 2, and minimum dosage =0.5

Ho: The difference in means equals zero vs HA: The difference in means is not equal to zero.

```
max_dose <- subset(ToothGrowth, ToothGrowth$dose == 02)
min_dose <- subset(ToothGrowth, ToothGrowth$dose == 0.5)
t.test(max_dose$len,min_dose$len,paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  max_dose$len and min_dose$len
## t = 11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.83383 18.15617
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

Since P value is lower than 0.05, we reject the null hypothesis.

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

conclusion

Therefore, as a result of the two hypothesis tests performed, we can conclude that there is no significant difference in tooth growth based on the method of suppliment(orange juice or vitamin C). However, there is a significant difference between the mean tooth length based on the dose used.

Assumptions

The tooth growth data has a continous, normal distribution, the two samples(for both the suppliment methods, and high/low dosages) are independant, random samples.

Appendix

Part 1

Code for plot displaying distribution of simulated sample

```
library(ggplot2)
histdata <- data.frame(sim_mean)
g <- ggplot(histdata,aes(x=sim_mean) ) +
        geom_histogram( binwidth = .25,color="black" , fill = "green" )
g <- g + labs( title = "Distribution of Simulated Means", x= "Simulated means")+ xlim(c(2,8))
g
```

Code for plot displaying sample mean Vs theoritical mean

```
library(ggplot2)
histdata <- data.frame(sim_mean)
g <- ggplot(histdata,aes(x=sim_mean) ) +
```

```
        geom_histogram( binwidth = .25,color="black" , fill = "green" )
g <- g + labs( title = "Sample Mean vs Theoritical Mean", x= "Means")+ xlim(c(2,8)) + geom_vline(xinter
g
```