Statistical Inference Part - 1

Niraj Nair

6 June, 2020

Overview

A pdf report is created by answering the questions in the problem statement. This report consists of graphs and calculations to answer the questions. The pdf report is made using knitr. The main tasks done here are simualtions to explore inference and simple inferential data analysis.

Problem Statement

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. 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.

Preparation

Loading require packages.

```
library(ggplot2)
set.seed(1)
```

Defining the variables. Sample Size = 40, Mean = .2, Number of simulations = 1000.

```
n <- 40
lambda <- .2
sim <- 1000
```

Creating the dataset using rexp.

```
dataset <- matrix(rexp(n*sim, lambda), sim)
dim(dataset)

## [1] 1000   40
```

Answers

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

The theoretical mean is 1 / lambda ie. 1/0.2 = 5.

```
theoMean <- 1 / lambda
sampleMeans <- apply(dataset, 1, mean)
```

```
calMean <- mean(sampleMeans)
round(calMean,2)

## [1] 4.99
```

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

The theoretical SD is (1 / lambda) * (1/sqrt(n)) Therefore, theoretical variance is square of theroretical SD.

```
theoSD <- (1 / lambda) * (1/sqrt(n))
theoVar <- theoSD^2
round(theoSD, 3)#Theoretical SD

## [1] 0.791

round(theoVar, 3) #Theoretical Variance

## [1] 0.625

calVar <- var(sampleMeans)
calSD <- sd(sampleMeans)
round(calSD, 3) #Calculated SD

## [1] 0.786

round(calVar,3) #Calculated Variance

## [1] 0.618
```

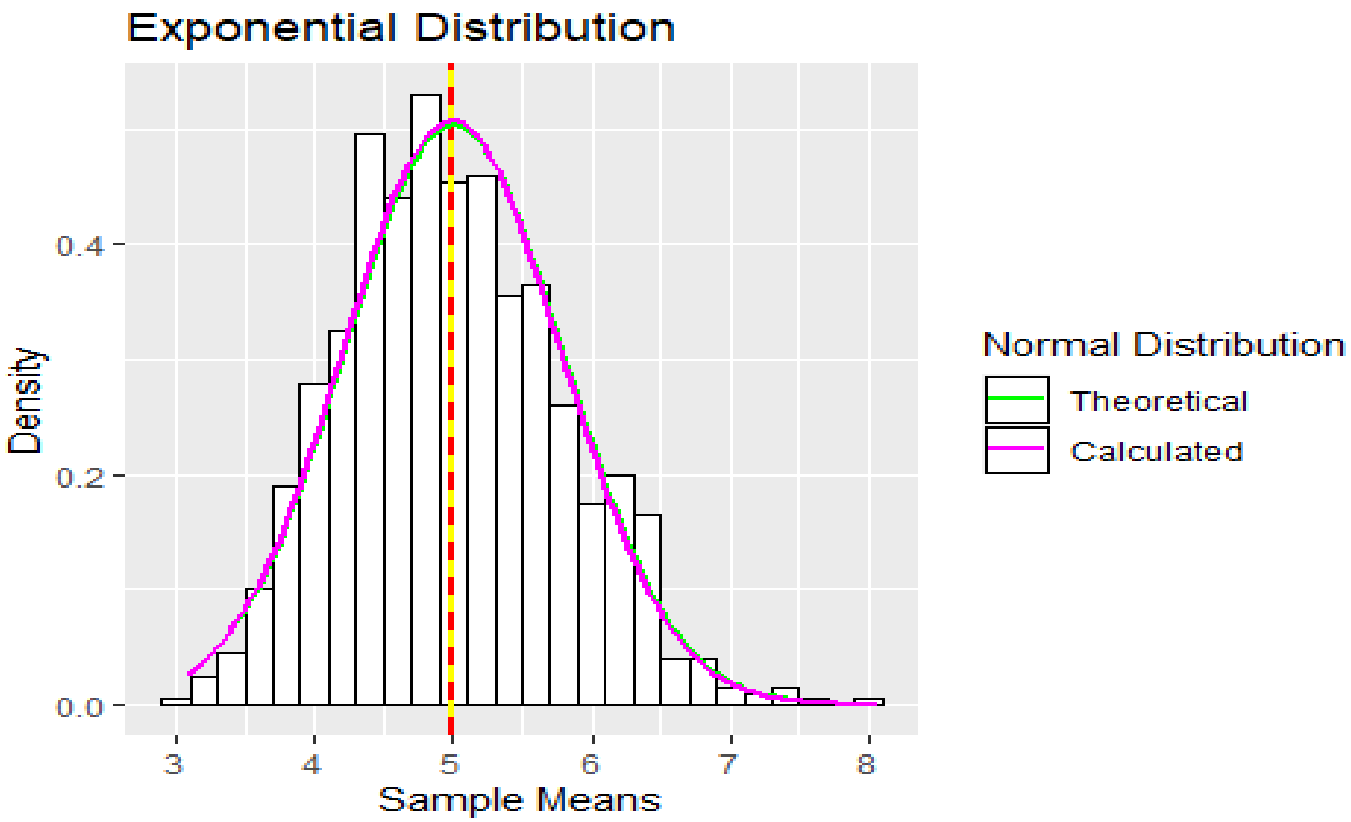
3.Show that the distribution is approximately normal.

```
sampleMeansData <- data.frame(sampleMeans)
g <- ggplot(data = sampleMeansData, aes(x=sampleMeans))
g + geom_histogram(binwidth = lambda, fill = "white", color="black", aes(y =
..density..), show.legend = TRUE, legendPosition = "top") +
    geom_vline(xintercept = theoMean, color = "yellow", size = 1) +
    geom_vline(xintercept = calMean, color = "red", size = .8, linetype =
"dashed") +
    stat_function(fun = dnorm, args = list(mean = theoMean, sd = theoSD),
aes(color = "green"), size = 1) +
    stat_function(fun = dnorm, args = list(mean = calMean, sd = calSD),
aes(color = "magenta"), size = 1) +
    labs(title = "Exponential Distribution", x = "Sample Means", y =
"Density")+
    scale_colour_identity(name = "Normal Distribution",labels =
c("Theoretical", "Calculated"),guide = "legend")

## Warning: Ignoring unknown parameters: legendPosition

## Warning: `mapping` is not used by stat_function()


## Warning: `mapping` is not used by stat_function()
```



Here

1. White line -> Theoretical Mean

2. Red dashed line-> Calculated mean

Conclusion

1. It is clear that that the center of distribution is approximately equal to the theoretical mean.

2. The Standard deviation and Variance of the distribution is compared to their theoretical values and it is found that they are approximately equal.

3. It is clear from the figure that the density distribution forms a bell curve and since the theoretical and calculated density distribution are very close to overlapping we can say that this is a Normal Distribution.