Verifying Central Limit Theorem

MOHAMMAD SHADAN

September 8, 2016

OVERVIEW

Central Limit Theorm (CLT) states that the distribution of averages of IID variables, properly normalized, becomes that of a standard normal as the sample size increases. (Refer URL, Slide 7/31)

In this project we investigate
* The Exponential Distribution in R and compare it with the Central Limit Theorem (CLT)
* The Exponential Distribution data is simulated in R with rexp(n, lambda) where lambda is the rate parameter.
* For Exponential Distribution, mean = 1/lambda and also standard deviation = 1/lambda

The analysis should elaborate on the below three points :
1. Show the sample mean and compare it to the theoretical mean of the distribution
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution
3. Show that the distribution is approximately normal

DATA PROCESSING

```
#Setting up the Working Directory, e.g. setwd("~/R/SI") and loading the required packages
library(ggplot2)
```

Detail about Number of Simulation, Sample Size and Rate Parameter which is already provided

```
num_of_sim  <- 1000    #Number of Simulations
sample_size <- 40      #Sample Size
lambda      <- 0.2     #Rate Parameter
```

Calculating the the **Theoretical** Mean, Standard Deviation and Variance based on Data Provided

```
theo_mean <- 1/lambda                     #Theoretical Mean of the Distribution
sigma     <- 1/lambda
theo_sd   <- sigma*(1/sqrt(sample_size))  #Theoretical Stan. Dev. of the Distribution
theo_var  <- sigma^2*(1/sample_size)      #Theoretical Variance of the Distribution
```

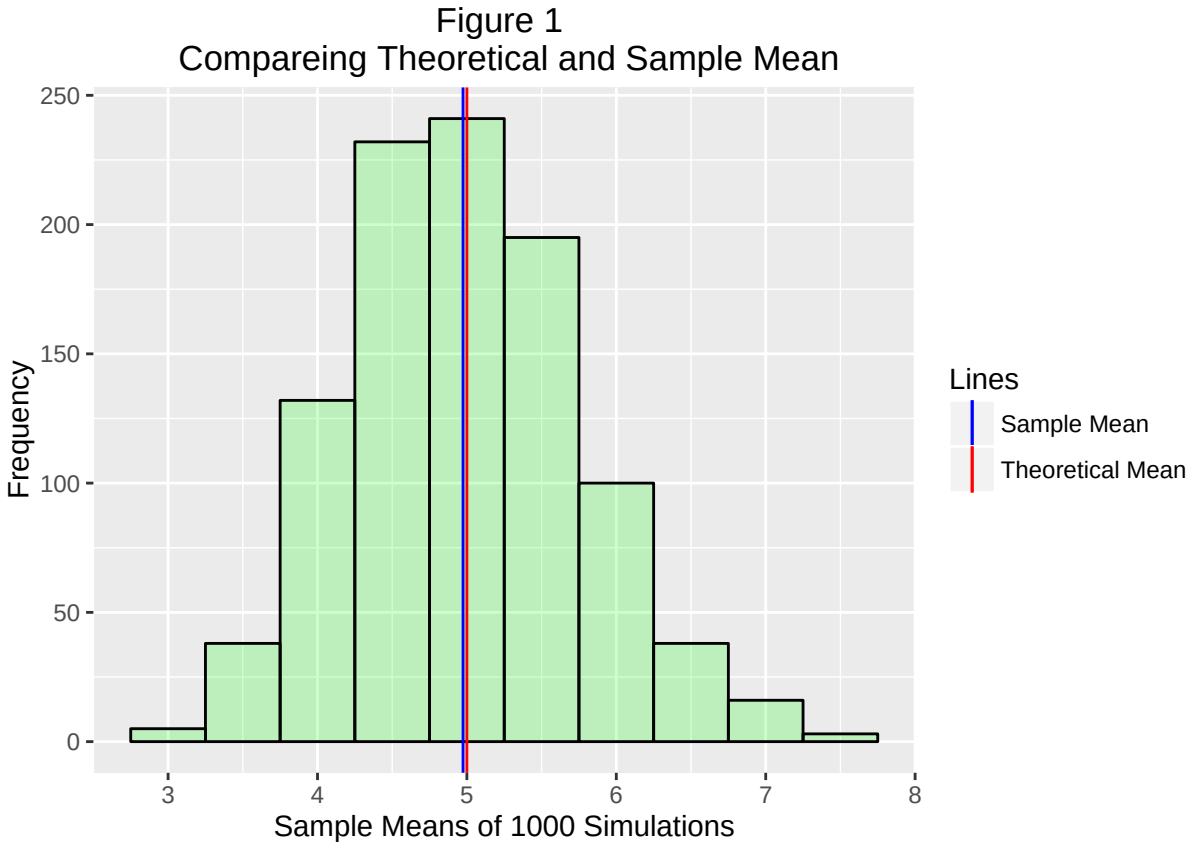
Creating the sample data using `rexp()` as adviced in the assignment

```
set.seed(1234) #To regenrate the same Random Numbers
#Generating Data for Exponential Distribution using rexp()
exp_data <- rexp(n = num_of_sim * sample_size, rate = lambda)
#Creating a matrix with 1000 (number of simulation) rows and 40 (sample size) columns
test_data <- matrix(exp_data, num_of_sim, sample_size)
#Converting the matrix into data frame
test_data <- as.data.frame(test_data)
#Adding a Column "row_mean" to the data set for mean of each row
test_data$row_mean <- apply(test_data,1,mean)
```

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

```
#Mean of the sample means (i.e. mean of the row means in our case)
sample_mean <- mean(test_data$row_mean)
#Displaying Theoretical Mean, Sample Mean and Difference between the two
cat("Theoretical Mean        :",theo_mean,"\nSample Mean             :", sample_mean,
    "\nDiff (Theo - Sample)    :", theo_mean - sample_mean)
```

```
## Theoretical Mean        : 5
## Sample Mean             : 4.974239
## Diff (Theo - Sample)    : 0.02576123
```



R Script for the above plot is available in APPENDIX under heading Figure 1

Observation : The Theoretical Mean and Sample Mean are almost same with a difference of only **0.02576123**. Same can be verified from the above plot as Theoretical Mean (vertical line in red) is almost merging with Sample Mean (vertical line in blue)

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

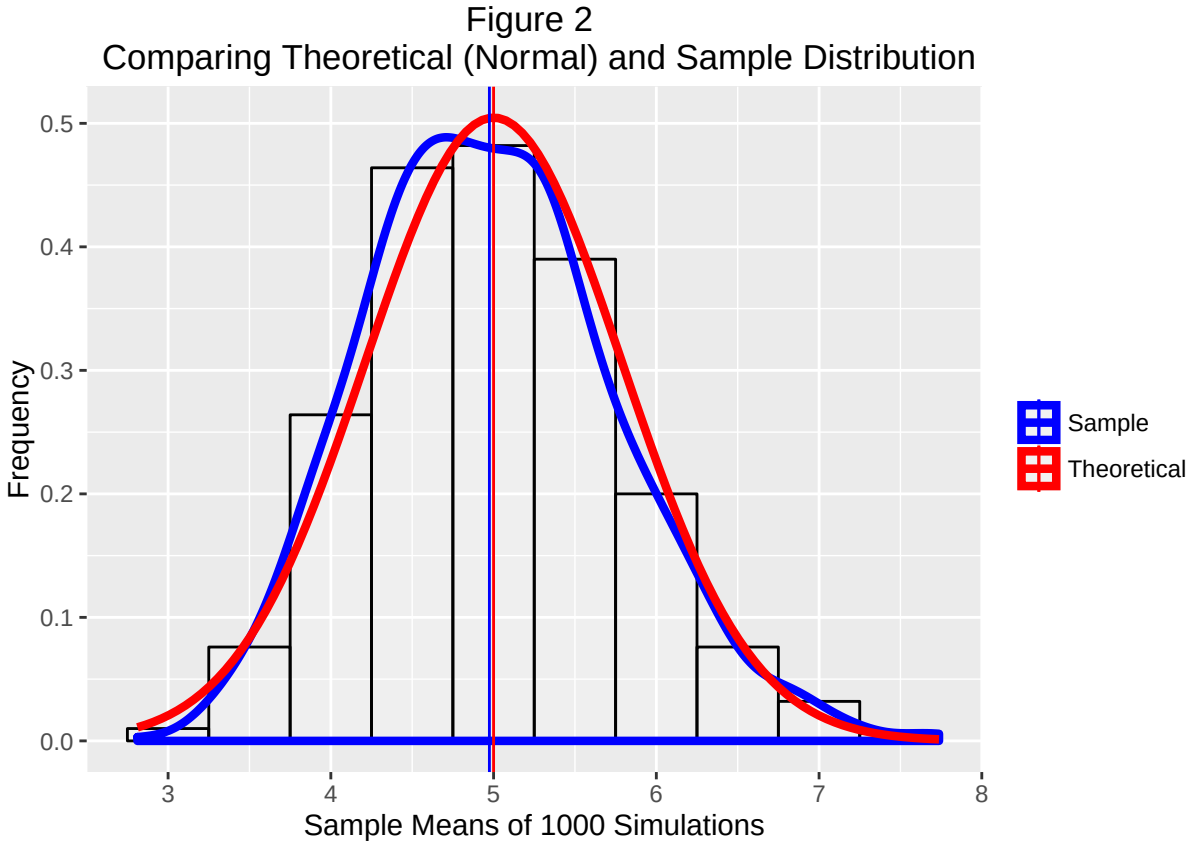
```
#Variance of the sample means
sample_var <- var(test_data$row_mean)

#Displaying Theoretical Mean, Sample Variance and Difference between the two
cat("Theoretical Variance    :",theo_var,"\nSample Variance         :", sample_var,
    "\nDiff (Theo - Sample)    :", theo_var - sample_var)
```

```
## Theoretical Variance    : 0.625
## Sample Variance         : 0.5949702
## Diff (Theo - Sample)    : 0.03002984
```

Observations :
* The Theoretical and Sample Variance are almost same with a difference of only **0.03002984**

3. Show that the distribution is approximately normal.

Details about the coloured distribution curves :
- Red Distribution Curve : Normal Distribution Curve with Mean as Theoretical Mean (theo_mean) and Standard Deviation as the Theoretical Standard Deviation (theo_sd)
- Blue Distribution Curve : The Sample Distribution Curve



R Script for the above plot is available in APPENDIX under heading Figure 2

Observation :
* As can be seen in the above plot the Sample Distribution (in blue) is almost merging with the Normal Distribution (in red), so we can conclude that the distribution is approximately normal

set.seed() is used with argument "1234" to generate the same random numbers. Running the script with different argument in set.seed() might change the Sample Distribution (Mean and Variance) to some extent but the Observations would remain same and Sample Distribution would be Approximately Normal.

Figure 1 (*used in Page 2*)

R Script to plot the sample mean and compare it to the theoretical mean of the distribution

```
#R Code to compare the Theoretical and Sample Mean
g1 <- ggplot(test_data, aes(row_mean))
g1 <- g1 + geom_histogram(binwidth = .5, fill="green", color = "black",alpha = .2)
g1 <- g1 + geom_vline( aes(xintercept = theo_mean, colour="Theoretical Mean"))
g1 <- g1 + geom_vline( aes(xintercept = mean(test_data$row_mean), colour="Sample Mean"))
g1 <- g1 + scale_colour_manual(name='Lines', values = c("Theoretical Mean"="red",
                                                        "Sample Mean"="blue"))
g1 <- g1 + labs(x = "Sample Means of 1000 Simulations")
g1 <- g1 + labs(y = "Frequency")
g1 <- g1 + labs(title = "Figure 1 \n Compareing Theoretical and Sample Mean")
```

Figure 2 (*used in Page 3*)

R Script to plot and show that the distribution is approximately normal

```
#Compareing Theoretical(Normal) and Sample Distribution
g3 <- ggplot(test_data, aes(row_mean))
g3 <- g3 + geom_histogram(aes(y=..density..), binwidth = .5, fill="white",
                          color = "black",alpha = .2)
g3 <- g3 + geom_density(aes(colour="Sample"), size=1.5)
g3 <- g3 + stat_function(fun=dnorm, args=list(mean=theo_mean, sd=theo_sd),
                         aes(colour="Theoretical"), size = 1.5)
g3 <- g3 + geom_vline( aes(xintercept = theo_mean, colour="Theoretical"))
g3 <- g3 + geom_vline( aes(xintercept = mean(test_data$row_mean), colour="Sample"))
g3 <- g3 + scale_colour_manual(name='', values = c("Theoretical"="red", "Sample"="blue"))
g3 <- g3 + labs(x = "Sample Means of 1000 Simulations")
g3 <- g3 + labs(y = "Frequency")
g3 <- g3 + labs(title = "Figure 2 \n Compareing Theoretical(Normal) and Sample Distribution")
```