# Statistical Inference - Simulation Assignment

*VN*

*9 September 2017*

## Overview

This document has been prepared for the submission of the assignment (part 1) of Statistical Interfernce course. This involves an exercise around Simulation. The objective is to investigate the exponential distribution in R and compare it with the Central Limit Theorem.

## Assignment

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 also 1/lambda. Set lambda = 0.2 for all of the simulations. In this simulation, you will investigate the distribution of averages of 40 exponential(0.2)s. Note that you will need to do a thousand or so simulated averages of 40 exponentials.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s.

### Global settings

Set working directory and load ggplot2

### Simulation

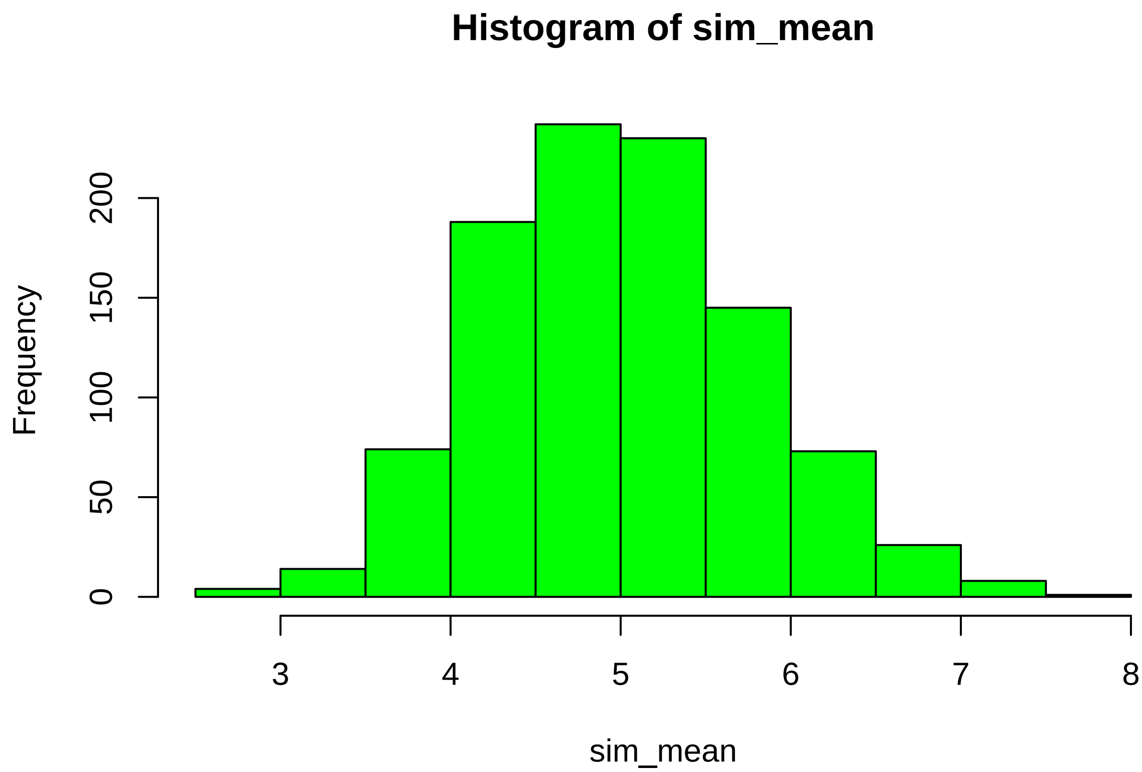
Prepare data after setting the seed value

```
## set parameters
set.seed(11234)
no_sim <- 1000
lambda <- 0.2
n <- 40

## generate data
sim_matrix <- matrix(rexp(no_sim * n, rate=lambda), no_sim, n)
sim_mean <- rowMeans(sim_matrix)
```

Data plot as histogram

```
hist(sim_mean, col = "green")
```




### Sample Mean comparison

The actual mean and theoratical mean of the data is calculated below

```
mean_data <- mean(sim_mean)
theory_mean <- 1/lambda
mean_data ## is the actual mean
```

```
## [1] 5.001142
```

```
theory_mean ## is the theoratical mean
```

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

As computed above both the values are very close

### Variance Comparison

Similary variance calculated below

```
actual_var <- var(sim_mean)
theory_var <- (1/lambda)^2/n
actual_var
```

```
## [1] 0.6138506
```

```
theory_var
```

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

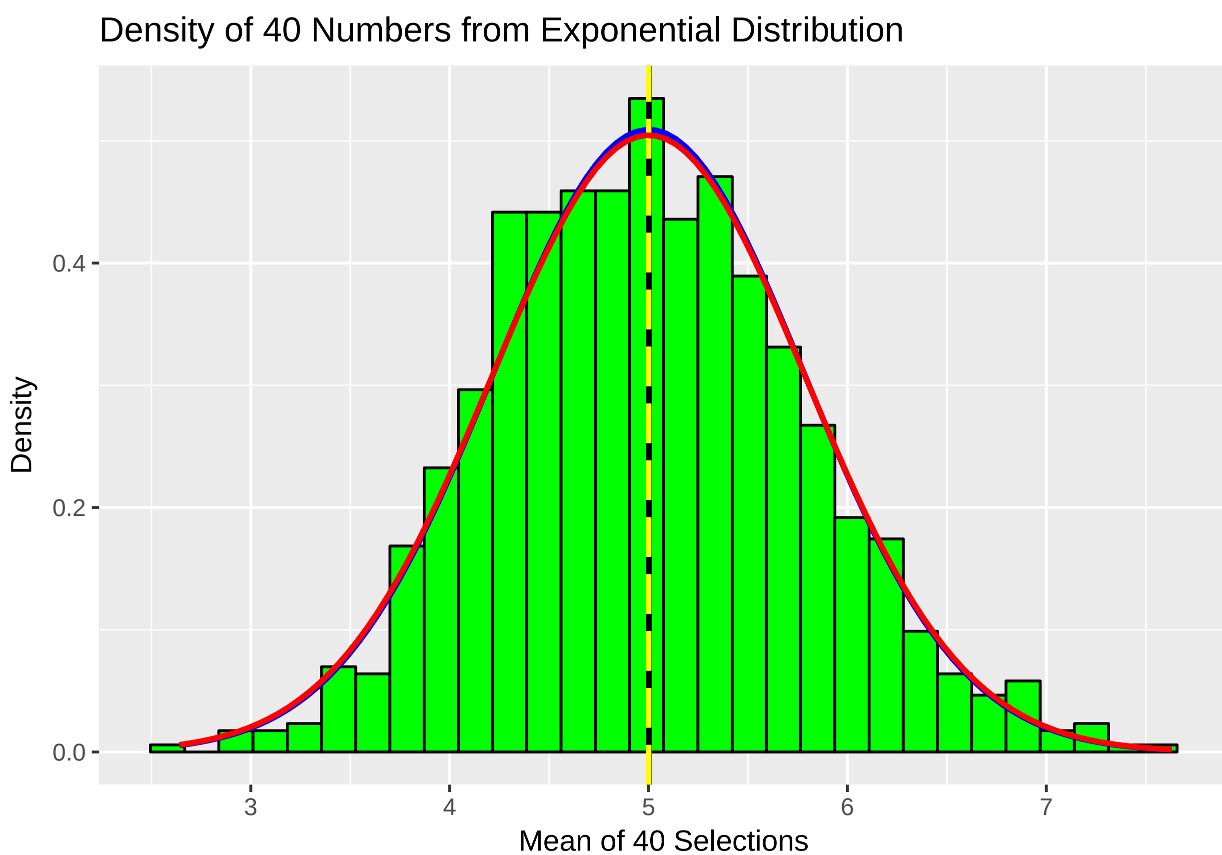
As with mean the variance above are also very close

### Plots

The data simulated has been plotted as below

```
plotdata <- data.frame(sim_mean)
m <- ggplot(plotdata, aes(x =sim_mean))
m <- m + geom_histogram(aes(y=..density..), colour="black",
fill = "green")
m<- m + labs(title="Density of 40 Numbers from Exponential Distribution", x="Mean of 40 Selections", y=
m<- m + geom_vline(xintercept=mean_data,size=1.0, color="black") # add a line for the actual mean
m<- m + stat_function(fun=dnorm,args=list(mean=mean_data, sd=sqrt(actual_var)),color = "blue", size = 1
m<- m + geom_vline(xintercept=theory_mean,size=1.0,color="yellow",linetype = "longdash")
m<- m + stat_function(fun=dnorm,args=list(mean=theory_mean, sd=sqrt(theory_var)),color = "red", size =
m
```

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




Key points from the plot
- GREEN bars indicate the density of the data
- BLACK line represents the Actual mean of the data
- BLACK DASHED line represents the Theoratical mean of the data
- RED normal curve with the theortical mean and sd
- BLUE normal curve with the actual mean and sd

## Summary

The simulation and plots help with answers to the questions :
- *1* actual mean is close to the theroetical mean.
- *2* actual variance is close to theoretical variance
- *3* last of the plots show that the actual curve is very close to the theoretical normal curve therefore proving the Central Limit Theory.