Statistical Inference Course Project - 1

Author: Krupa Rajendran

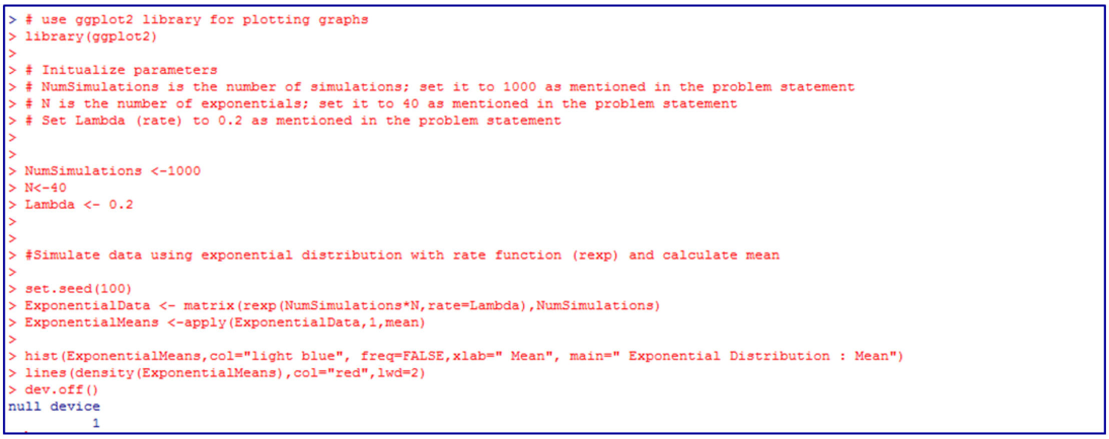
This course project is a simulation exercise and investigate the exponential distribution in R and compare it with the Central Limit Theorem. Though simulation, the following report demonstrates that the calculated mean and variance for the simulated exponentially distributed data is close to theoretical mean and variance. The report also illustrates that for large samples, the Exponential distribution is approximately Normal.

Requirement 1 is to calculate the sample mean and compare it to the theoretical mean of the distribution.

The Exponential distribution is simulated using the function 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. As mentioned in the project requirement, lambda = 0.2 for all of the simulations.

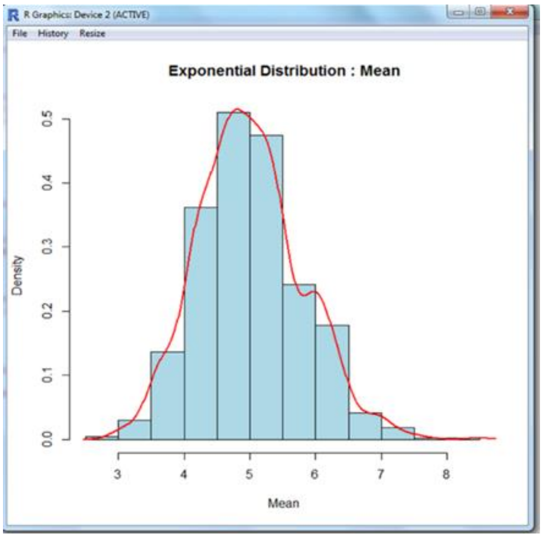
Studied the distribution of averages of 40 exponentials. And number of simulations is taken as 1000.

The R code is as follows:



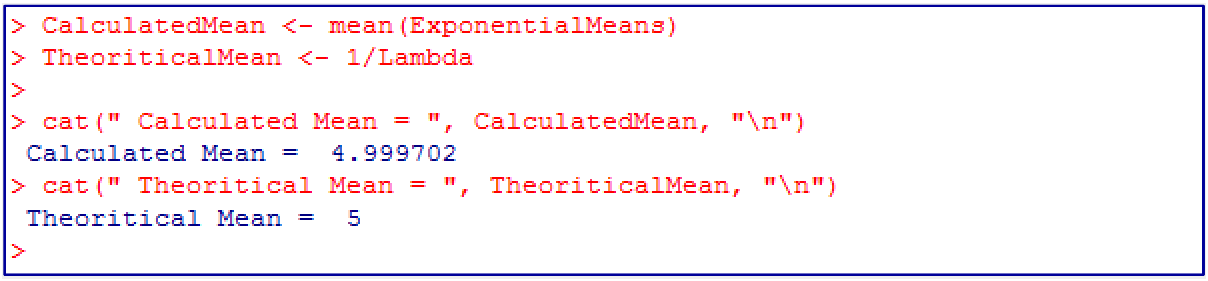
```
> # use ggplot2 library for plotting graphs
> library(ggplot2)
>
> # Initualize parameters
> # NumSimulations is the number of simulations; set it to 1000 as mentioned in the problem statement
> # N is the number of exponentials; set it to 40 as mentioned in the problem statement
> # Set Lambda (rate) to 0.2 as mentioned in the problem statement
>
>
> NumSimulations <-1000
> N<-40
> Lambda <- 0.2
>
>
> #Simulate data using exponential distribution with rate function (rexp) and calculate mean
>
> set.seed(100)
> ExponentialData <- matrix(rexp(NumSimulations*N,rate=Lambda),NumSimulations)
> ExponentialMeans <-apply(ExponentialData,1,mean)
>
> hist(ExponentialMeans,col="light blue", freq=FALSE,xlab=" Mean", main=" Exponential Distribution : Mean")
> lines(density(ExponentialMeans),col="red",lwd=2)
> dev.off()
null device
          1
```

The graph from the above code is as follows:



As visible from the above graph, the mean is close to 5.

Following is the R code and output

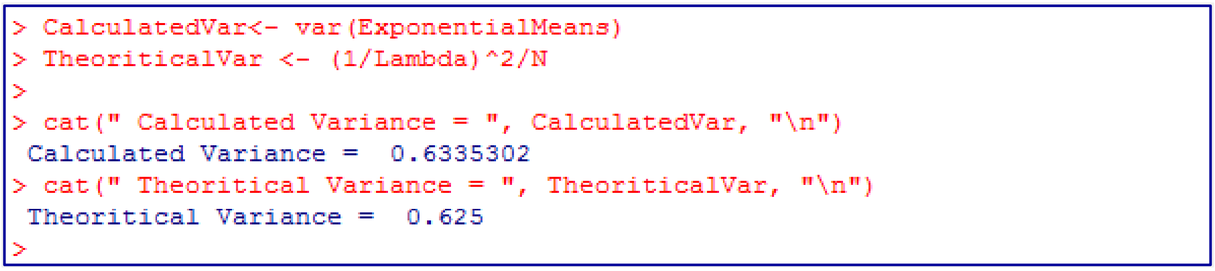


```
> CalculatedMean <- mean(ExponentialMeans)
> TheoriticalMean <- 1/Lambda
>
> cat(" Calculated Mean = ", CalculatedMean, "\n")
 Calculated Mean =  4.999702
> cat(" Theoritical Mean = ", TheoriticalMean, "\n")
 Theoritical Mean =  5
>
```

Mean for the simulated Exponential Distribution (when calculated through R code) is 4.9997 while the theoretical mean of the distribution (1 / Lambda) is 5.

Requirement 2 is to calculate sample variance and compare it to the theoretical mean of the distribution.

The Standard Deviation for Exponential Distribution is 1/Lambda and hence the variance would be (1/Lambda) ^2/N, where N is the size of the sample.



```
> CalculatedVar<- var(ExponentialMeans)
> TheoriticalVar <- (1/Lambda)^2/N
>
> cat(" Calculated Variance = ", CalculatedVar, "\n")
 Calculated Variance =  0.6335302
> cat(" Theoritical Variance = ", TheoriticalVar, "\n")
 Theoritical Variance =  0.625
>
```

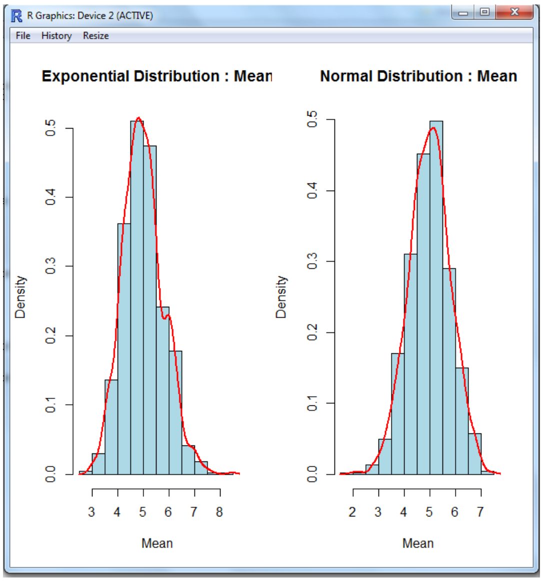
Variance for the simulated Exponential Distribution (when calculated through R code) is 0.6335 while the theoretical variance of the distribution (1 / Lambda)^2/N is 0.625.

Requirement 3 is to show that the distribution is approximately normal:

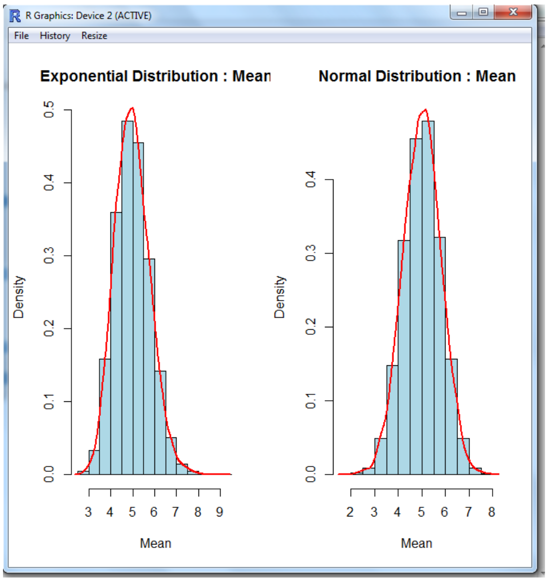
The central limit theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

That is, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution (commonly known as a "bell curve").

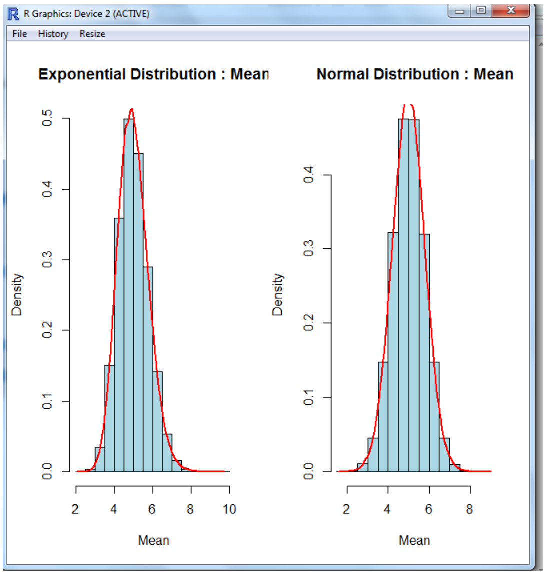
To show that the Exponential distribution is approximately normal for large samples, the exponential and normal distributions are plotted adjacent to each other so that the curves can be compared. This proves the CLT that the mean of a sufficiently large number of iterates of independent random variables, will be approximately normally distributed, regardless of the underlying distribution.



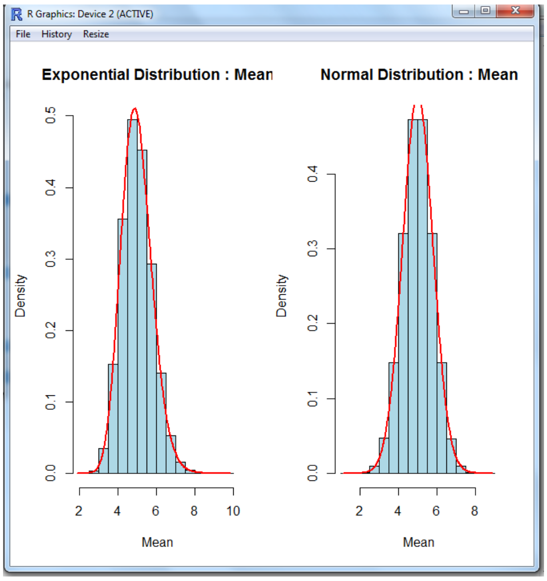
Number of simulations = 1000



Number of simulations = 10000

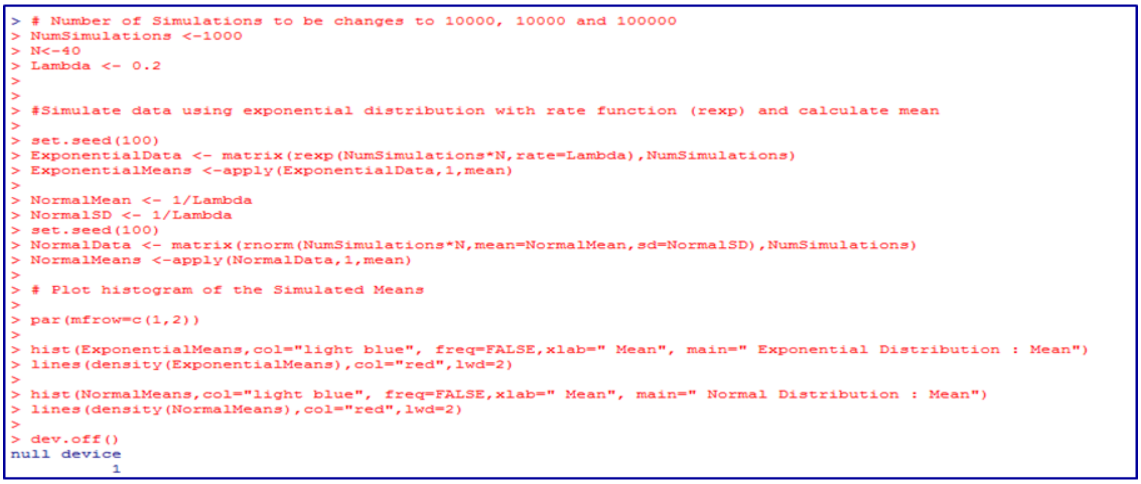


Number of simulations = 100000



Number of simulations = 1000000

The R code for creating these graphs is as follows:



```
> # Number of Simulations to be changes to 10000, 10000 and 100000
> NumSimulations <-1000
> N<-40
> Lambda <- 0.2
>
>
> #Simulate data using exponential distribution with rate function (rexp) and calculate mean
>
> set.seed(100)
> ExponentialData <- matrix(rexp(NumSimulations*N,rate=Lambda),NumSimulations)
> ExponentialMeans <-apply(ExponentialData,1,mean)
>
> NormalMean <- 1/Lambda
> NormalSD <- 1/Lambda
> set.seed(100)
> NormalData <- matrix(rnorm(NumSimulations*N,mean=NormalMean,sd=NormalSD),NumSimulations)
> NormalMeans <-apply(NormalData,1,mean)
>
> # Plot histogram of the Simulated Means
>
> par(mfrow=c(1,2))
>
> hist(ExponentialMeans,col="light blue", freq=FALSE,xlab=" Mean", main=" Exponential Distribution : Mean")
> lines(density(ExponentialMeans),col="red",lwd=2)
>
> hist(NormalMeans,col="light blue", freq=FALSE,xlab=" Mean", main=" Normal Distribution : Mean")
> lines(density(NormalMeans),col="red",lwd=2)
>
> dev.off()
null device
          1
```