Statistical Inference -Peer Assessment1

Assignment Instrustion

This is the project for the statistical inference class. In it, you will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

You will create a report to answer the questions. Given the nature of the series, ideally you'll use knitr to create the reports and convert to a pdf. (I will post a very simple introduction to knitr). However, feel free to use whatever software that you would like to create your pdf.

Each pdf report should be no more than 3 pages with 3 pages of supporting appendix material if needed (code, figures, etcetera).

Problem Statement

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

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

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.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Envirnoment Setup

```
setwd("C:/Statistical Inference/Assignment1")


require(knitr)
## Loading required package: knitr
require(ggplot2)
## Loading required package: ggplot2
opts_chunk$set(echo=TRUE)
set.seed(1)
```

Define and set the variables as mentioned in assignment instruction. Number of values (n) = 40 lambda = 0.2 number of iterations, as least 1000. numsim=2000

```
n<-40
lambda<-0.2
numsim<-2000
```

Create data using rexp

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

Answer the following questions

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

Question 2 - Show how variable it is and compare it to the theoretical variance of the distribution.

Question 3 - Show that the distribution is approximately normal.

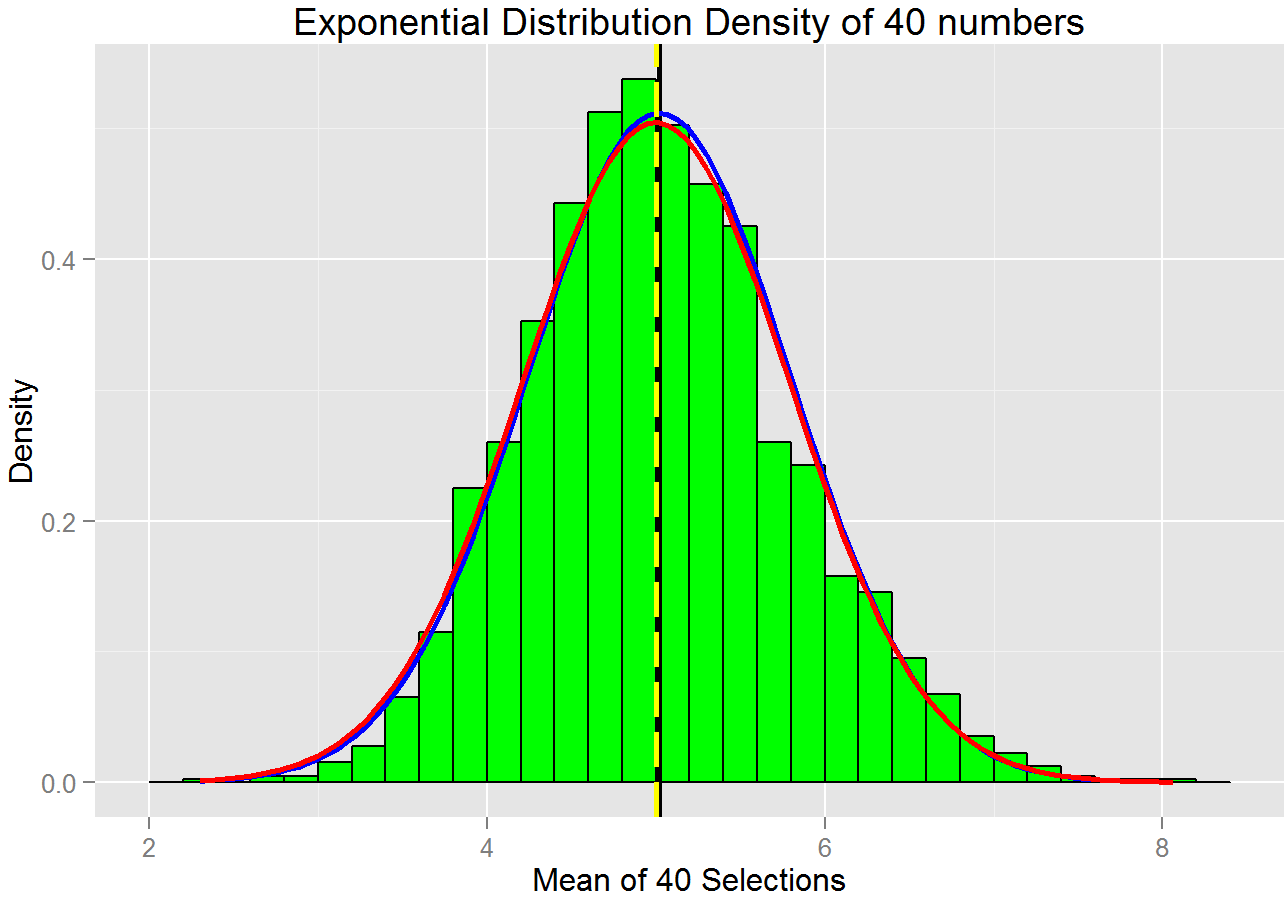
The theoretical mean was given in the problem as 1 / lambda or 1 / 0.2. The actual mean of the generated data should be calculated by using the apply & mean functions to obtain a mean for each row and then taking the mean of those numbers. Let us get the Standard Deviation and Variance as well as it will be required at later stage.

```
TheoryMean<-1/lambda
RowMeans<-apply(dataset,1,mean)
ActualMean<-mean(RowMeans)
TheorySD<-((1/lambda) * (1/sqrt(n)))
ActualSD<-sd(RowMeans)
TheoryVar<-TheorySD^2
ActualVar<-var(RowMeans)
```

ActualMean = 5.002 Theoretical Mean = 5.0 AcutalSD = 0.779 TheoreticalSD = 0.790 ActualVariance = 0.6072 TheoreticalVariance = 0.625

plotting as following

```
dfRowMeans<-data.frame(RowMeans) # convert to data.frame for ggplot
mp<-ggplot(dfRowMeans,aes(x=RowMeans))
mp<-mp+geom_histogram(binwidth = lambda,fill="green",color="black",aes(y = ..density..))
mp<-mp + labs(title="Exponential Distribution Density of 40 numbers", x="Mean of 40 Selections", y="Density")
mp<-mp + geom_vline(xintercept=ActualMean,size=1.0, color="black") # add a line for the actual mean
mp<-mp + stat_function(fun=dnorm,args=list(mean=ActualMean, sd=ActualSD),color = "blue", size = 1.0)
mp<-mp + geom_vline(xintercept=TheoryMean,size=1.0,color="yellow",linetype = "longdash")
mp<-mp + stat_function(fun=dnorm,args=list(mean=TheoryMean, sd=TheorySD),color = "red", size = 1.0)
mp
```



Conclusion

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

It is Clearly demonstrated(Both Number and Graph) the actual mean compared to the theroetical mean

Show how variable it is and compare it to the theoretical variance of the distribution.

It is demonstrated the actual standard deviation and variance compared to their theoretical values(both Number and Graph)

Show that the distribution is approximately normal

It is demonstrated through graphical (including numerical) form that both the calculated curve is very close to the theoretical normal curve which endorces Central Limit Theory.