Peer Assessments / Statistical Inference Course Project - Part 1 of 2

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).

OVERVIEW:

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.

SIMULATIONS:

Set global options

```r
library(knitr)

opts_chunk$set(echo=TRUE)

set.seed(1)
```

Set variables

```r
lambda <- 0.2

n  <- 40

numsim <- 1000
```

Generate dataset

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

Calculate descriptive statistics

```r
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)
```

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.

```
ActualMean
## [1] 4.990025
TheoryMean
## [1] 5
```

Actual Mean = 4.990
Theoretical Mean = 5

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

```
ActualVar
## [1] 0.6177072
TheoryVar
## [1] 0.625
```

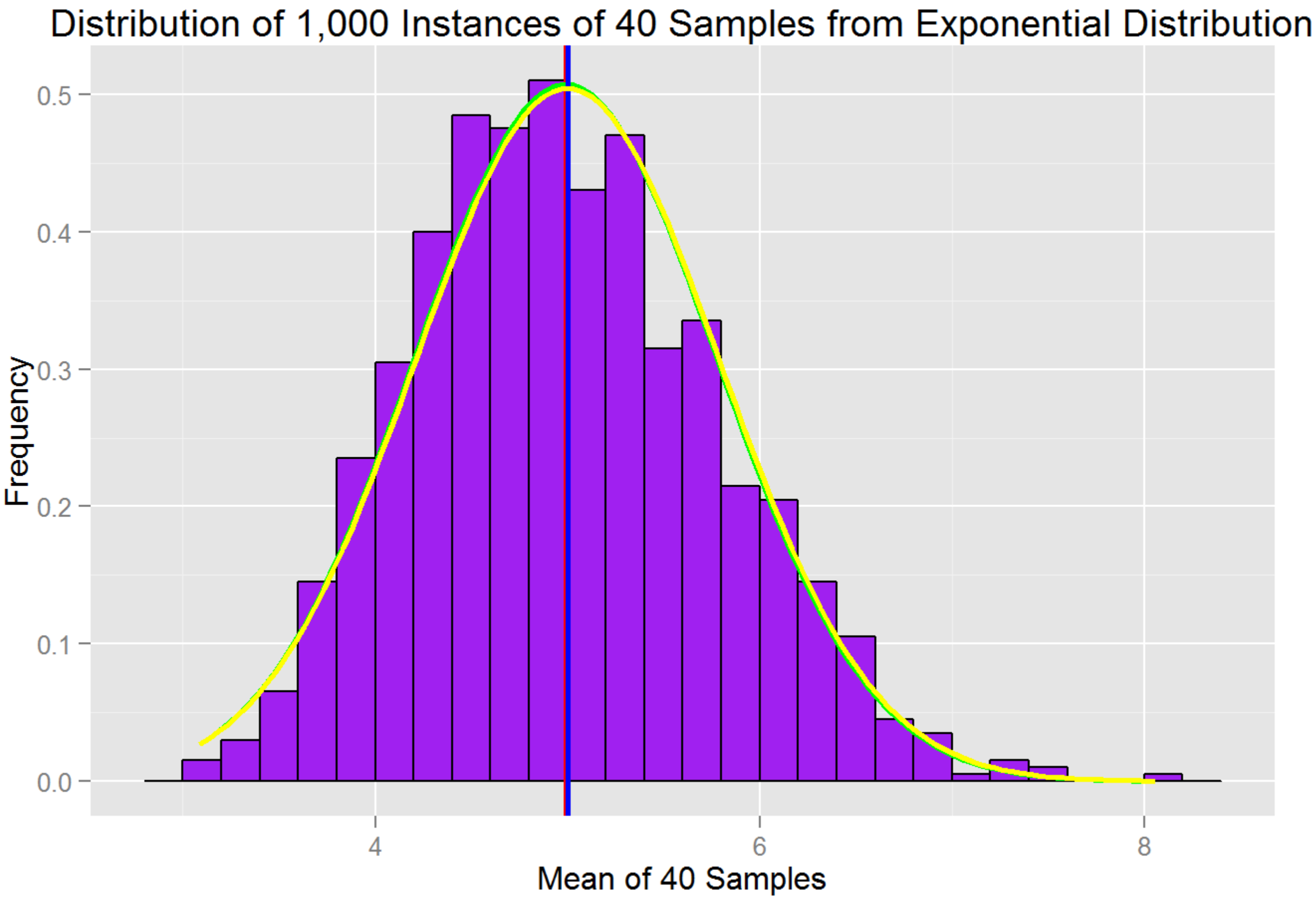
Actual Variance = 0.6177
Theoretical Variance = 0.625

3. Show that the distribution is approximately normal.

Plot distribution

```
library(ggplot2)
dfRowMeans <- data.frame(RowMeans)
mp <- ggplot(dfRowMeans,aes(x=RowMeans))
mp <- mp+geom_histogram(binwidth = lambda,fill="purple",color="black",aes(y = ..density..))
mp <- mp + labs(title="Distribution of 1,000 Instances of 40 Samples from Exponential Distribution", x="Mean of 40 Samples", y="Frequency")
mp <- mp + geom_vline(xintercept=ActualMean, size=1.0, color="red")
mp <- mp + stat_function(fun=dnorm,args=list(mean=ActualMean, sd=ActualSD),color = "green", size = 1.0)
mp <- mp + geom_vline(xintercept=TheoryMean,size=1.0,color="blue")
mp <- mp + stat_function(fun=dnorm,args=list(mean=TheoryMean, sd=TheorySD),color = "yellow", size = 1.0)
```

```
mp
```



- The actual mean is shown by a **red** line.
- The theoretical mean is shown by a **blue** line
- The actual curve formed by the mean and standard deviation is shown in **green**.
- The normal curve formed by the the theoretical mean and standard deviation is shown in **yellow.**

The actual data is approximately normally distributed as predicted by the **Central Limit Theorem**.