A Simulation Exercise - Part1 - "Statistical Inference"

Date: 09-26-2015 **Course Student:** DataRacer11 **References: Course book** by Brian Caffo, PhD, John Hopkins University

Overview - Information about the simulation exercise: (Exercise questions are highlighted in yellow on each page)

The exponential distribution of the mean of 40 exponentials in R is compared with the Central Limit Theorem. Exponential Distribution is described as: Density, distribution function, quantile function and random generation for the exponential distribution with rate (i.e., mean 1/rate). In this simulation exercise we will compare the observed or empirical mean of the sample data with the theoretical or expected mean of the sample data. **A common application of the exponential distribution is the time between events such as the inter-arrival time between earthquakes.** The R-code rexp(n=12, rate=6) will generate twelve inter-arrival times when the rate is 6. For example, 6 quakes per hour. The R code `rexp(n, lambda)` will be utilized in this project to simulate this exponential distribution. A sample of n= 40, where lambda is now the rate parameter will be used. An exponential distribution can be randomly generated in R. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. For all simulations, we will set lambda = 0.2. The distribution of averages will require one thousand simulations. **Next, The Central Limit Theorem (CLT)** for purposes of our exercise states that the distribution of averages of `iid` variable (properly normalized) becomes that of a standard normal as the sample size increases. The CLT applies in an endless variety of settings. This fact holds especially true for sample sizes over 30. As more samples are taken, especially large ones, a graph of the sample means will look more like a normal distribution. A common simple test for CLT is rolling a fair die.
The more times you roll the die, the more likely the shape of the distribution of the means tends to look like a normal distribution graph.

Simulations: **The following illustrates via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.**

```
lambda <- 0.2                                                    #Set lambda as rate parameter
sample_exp <- 40                                                 #Set 40 exponentials as sample size
req_sims  <- 1:1000                                              #Set required simulations as 1000
empMean <- mean(x_mean$x)                                        #Set the Empirical Mean
theorMean <-mean(1/lambda)                                       #Set the Theoretical Mean as 1/lambda
empVar <- var(x_mean$x)                                          #Set the Variances for Empirical*
theorVar <- ((1/lambda)/sqrt(40))^2                              #Set the Variances for Theoretical*
aggreMean <- cumsum(x_mean$x) / seq_along(x_mean$x)              #Set the Aggregate / Cumulative Mean*
aggreVar <- cumsum((x_mean$x - empMean)^2)/(seq_along(x_mean$x)-1)  #Set the Aggregate / Cumulative Variance*

set.seed(6000) #Set the seed value equal to 6000. (The number 6000 is not special. We could have used any positive integer)
```

Set the data.frame per the simulation requirements as described in paragraph two above.

```
x_mean <- data.frame(x = sapply(req_sims, function(x) {mean(rexp(sample_exp,lambda))})
```

Review the descriptive statistics like mean, median, mode, standard deviation, distribution of values and summary of the key characteristics of the distribution of the data:

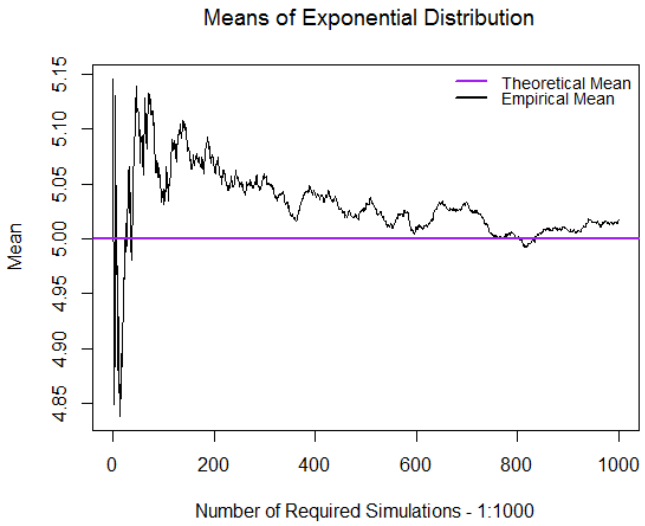
```
head(x_mean)                                                     #Outputs only the first few rows of sample data.
tail(x_mean)                                                     #Outputs only the last few rows of sample data.
```

Empirical Mean versus Theoretical Mean: 1. Shows the sample mean and compares it to the theoretical mean of the distribution.

This first plot describes the cumulative of the means which have been sampled from the 40 simulations for exponential distribution. The empirical mean obtained from summary was 5.02. The theoretical mean 1/lambda where lambda=.02 is 5. These are approximately the same values.

```
summary(x_mean)        X      Min.: 2.990 1st Qu.: 4.434 Median: 4.963 Mean: 5.017 3rd Qu.: 5.544 Max.: 7.833
```

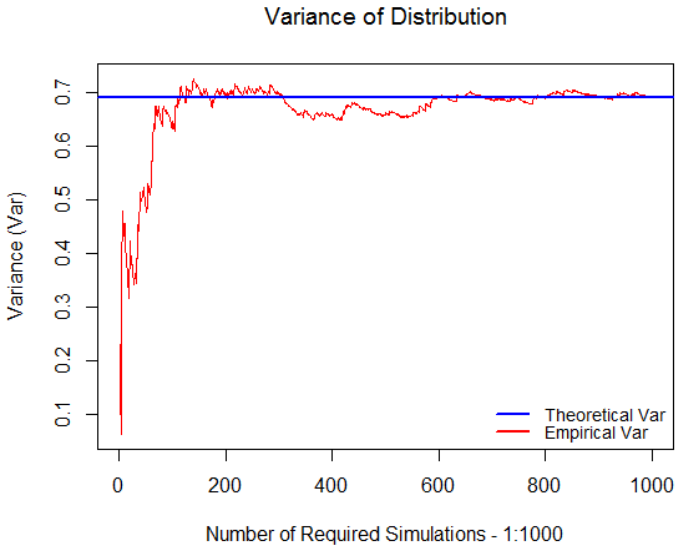
```
plot(seq_along(x_mean$x), aggreMean, type="l", lty=1, lwd=1, main=expression(" Means of Distribution"),xlab="Number of Required Simula
abline(h=theorMean, col="purple",lwd=2)
legend("topright", legend=c("Theoretical Mean","Empirical Mean"), col=c("purple", "black"), lty=c(1,1), lwd=c(2,2), cex=.9, bty="n")
```



Sample Variance versus Theoretical Variance: 2. Shows how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

The second plot shows the variances of the means. The more we increase the number of iterations the closer the variances also become.

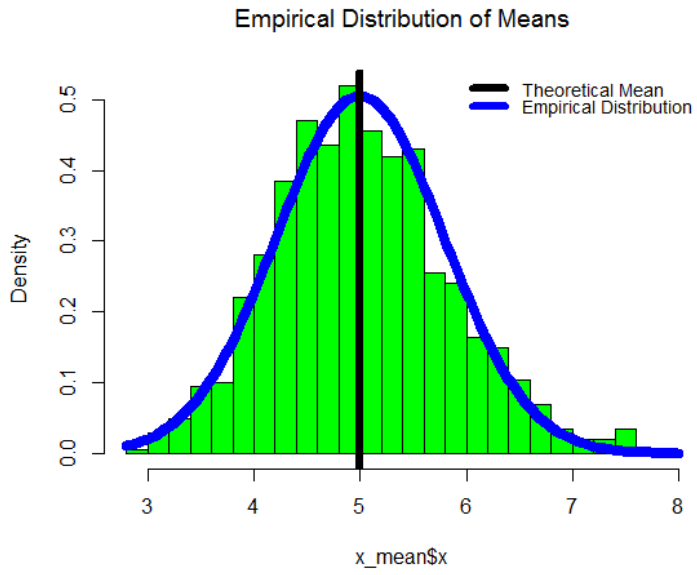
```
plot(seq_along(x_mean$x), aggreVar, type="l", lty=1, lwd=1, col="red", main=expression("Variance of Distribution"),xlab="Number of Req
abline(h=empVar, col="blue",lwd=2)
legend("bottomright", legend=c("Theoretical Var","Empirical Var"), col=c("blue", "red"), lty=c(1,1), lwd=c(2,2), cex=.9, bty="n")
```



The third plot shows the histogram with each distribution. 3. A. The difference between the distribution of a large collection of random exponentials is that the larger the collection and the greater the amount of data the more closer to the empirical is to the theoretical mean. 3.B. Smaller data samples will have greater dispersion. The values in the original datasets depends on the dispersion or variability in the original data sets. Datasets are said to have high dispersion when they cotain values considerably higher or lower than the mean value.

```
x=seq(0,8,0.5)
hist(x_mean$x, breaks=20,freq=FALSE,col="green", main=expression("Empirical Distribution of Means"))
```

```
curve(dnorm(x, mean=theorMean, sd=sqrt(theorVar)),
add=TRUE,lwd=8,col="red")
abline(v=theorMean,lwd=6,col="black")
legend("topright", legend=c("Theoretical Mean","Empirical Distribution"),
col=c("black", "red"), lty=c(1,1), lwd=c(6,6), cex=.9, bty="n")
```



Distribution: 3. Show that the distribution is approximately normal. ***With the fourth*** **plot one can tell the distribution is approximately normal with a q-q plot. The quantiles of the theoretical data set against the quantiles of empirical dataset is shown. A 45 degree ref. line depicts two data sets have similar distribution shapes.**

```
qqnorm(x_mean$x,pch=16,cex=.8,col="purple")
qqline(x_mean$x,col="blue",lwd=2)
legend("bottomright",legend=c("Empirical","Theoretical"),
col=c("purple","blue"),pch=c(16,NA),
lwd=c(NA,2),lty=c(NA,1),cex=.8,bty="n")
```