Exponential Distribution and Analysis of CLT

F. Alex Crofut

Sunday, April 10, 2016

Overview

The purpose of this project is to investigate the exponential distribution and compare it with the Central Limit Theorem (CLT). The approach is to take 1,000 simulations of an exponential distribution of size 40 and rate of lambda = 0.2. Then the distribution of the means and variance of the simulations is examined and compared to the theoretical distribution and normal curve to determine if the CLT holds.

Load Packages

This analysis utilizes the following packages:
- ggplot2
- plyr
- knitr

The folowing r code will determine if the required package is available, and if not, will install it. The packages are then loaded.

```
if (!"ggplot2" %in% installed.packages()) install.packages("ggplot2")
library("ggplot2")
if (!"plyr" %in% installed.packages()) install.packages("plyr")
library("plyr")
if (!"knitr" %in% installed.packages()) install.packages("knitr")
library("knitr")
```

Simulations

The seed is set in order to maintain reproducibility of the project. One thousand simulations of exponential distributions of size 40 are taken and added to a matrix. The means of each simulation are stored in a vector.

```
set.seed(123) # set the seed
n <- 40 # size of each simulation
lambda <- 0.2 # given value of lambda, the rate parameter
nsims <- 1000 # no. of simulations
sims <- matrix(rexp(nsims * n, rate=lambda), nsims)    # create a matrix with results of random samplin
simsMeans <- apply(sims, 1, mean)   # store the means of each simulation
```

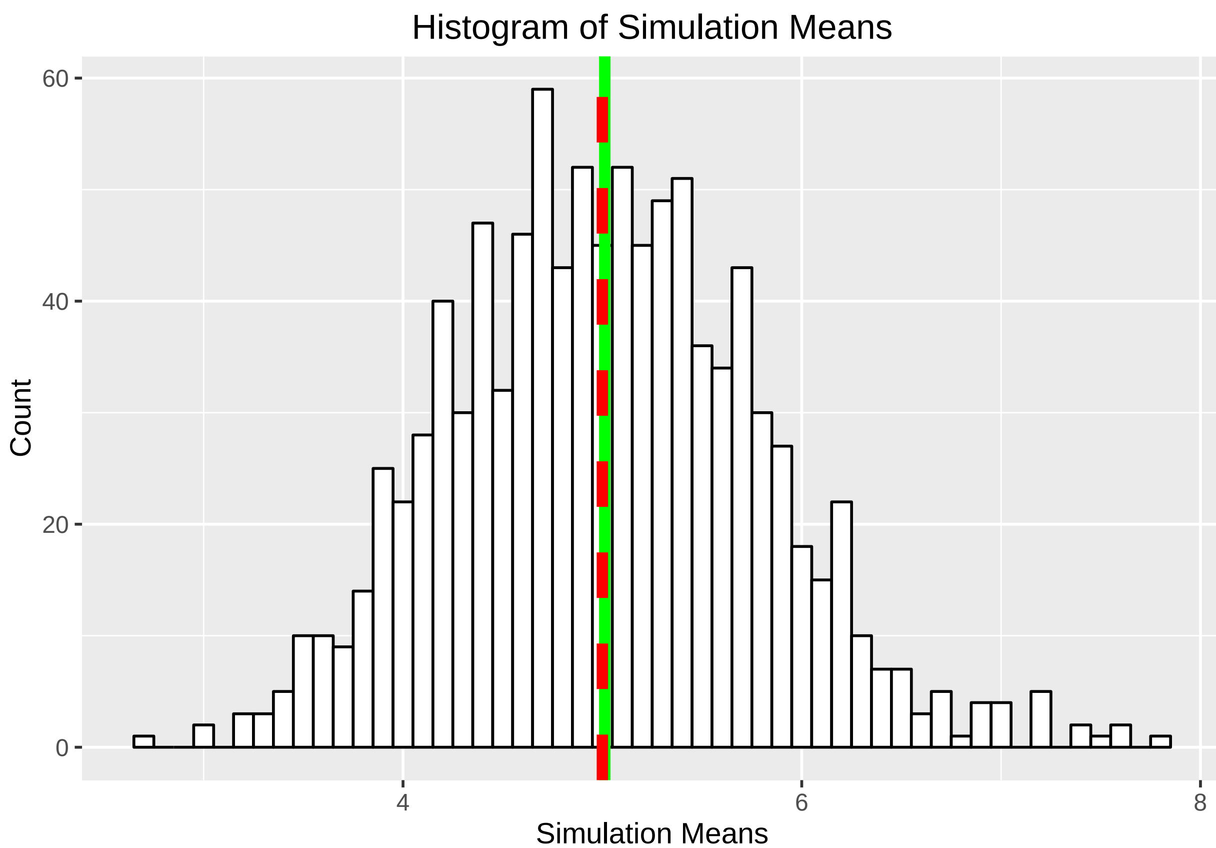
Sample Mean versus Theoretical Mean

With the simulations completed, one can now compare the sample mean to the theoretical mean. According to the CLT, these should be approximately equal.

```
sampleMean <- mean(sims)
theoreticalMean <- 1 / lambda
```

The sample mean is 5.012 while the theoretical mean is 5, a difference of only 0.012. In the figure below, one can see first that the distribution appears to be approximately normal. One can also see that the mean of the simulations (green line) is indeed approximately the same as the theoretical mean (red dashed line).

```
meansPlot <- ggplot() + aes(simsMeans) + geom_histogram(binwidth=.1, colour="black", fill="white") +
    geom_vline(xintercept=sampleMean, color="green", size=2) +
    geom_vline(xintercept=theoreticalMean, color="red", size=2, linetype=2) +
    labs(x="Simulation Means", y="Count", title="Histogram of Simulation Means")
meansPlot
```



Sample Variance versus Theoretical Variance

The sample variance and theoretical variance must also be compared in order to determine that the CLT holds.

```
sampleVar <- var(simsMeans)
theoreticalVar <- (1/lambda)^2/n
```

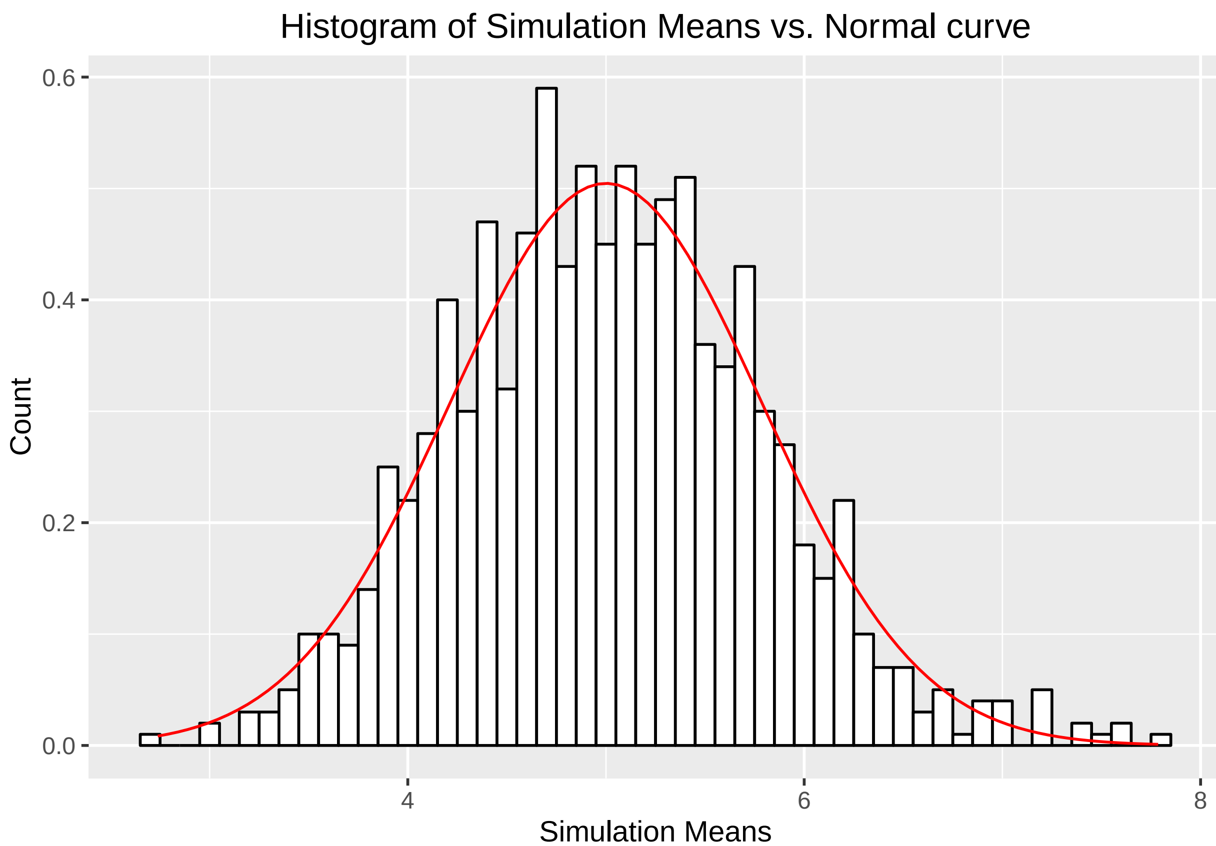
Sample Variance	Theoretical Variance	Difference
0.609	0.625	-0.016

As seen in the table above, the variance of the simulation means is quite close to the theoretical variance.

Distribution

In order to determine if the simulation means appear to be normal, one can look once more at the histogram, this time super-imposing a normal curve (red line) with the theoretical mean and standard deviation.

```
dat <- data.frame(simsMeans)
distPlot <- ggplot(dat, aes(x=simsMeans)) + geom_histogram(binwidth=0.1, aes(y=..density..), color="bla
      stat_function(fun=dnorm, color="red", args=list(mean=theoreticalMean, sd=sqrt(theoreticalVar))) +
      labs(x="Simulation Means", y="Count", title="Histogram of Simulation Means vs. Normal curve")
distPlot
```



One can visually determine from the above figure that the distribution closely follows the normal curve. One can also determine whether or not the theoretical mean of 5 lies within the 95% confidence interval of our simulation means distribution.

```
sample95Int <- sampleMean + c(-1,1) * qnorm(0.975) * sqrt(sampleVar) / sqrt(n)
```

The sample 95% confidence interval is 4.77 to 5.254, which includes the theoretical mean of 5.

Conclusion

This project has shown that the mean of the sample means from 1,000 simulations of n=40 is approximately equal to the theoretical mean. Furthermore, the distribution of these means is approximately normal with variance approximately equal to the theoretical variation. Therefore, the distribution of the sample means follows the Central Limit Theorem.