# Part 1: Simulation Exercise

*Erik Johnson*

*10/30/2017*

## Overview

In this report, we investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). We investigate the distribution of averages of 40 exponentials, and show that the distribution of the averages is approximately normal as hypothesized by the CLT.

## Simulations

We define a simulation as drawing 40 random variables from the exponential distribution with parameter lambda = 0.2.

We start by running 1000 such simulations and storing the resultant random variables in a matrix called sims.

```
library(ggplot2)
```

```
# Set the random seed so that the results can be exactly reproduced
set.seed(1234)

# Set the parameters for the simulations
lambda <- 0.2
numberPerSim <- 40
numberOfSims <- 1000

# Run 1000 simulations, drawing 40 random variables from the exponential distribution
# The result is a matrix whose columns represent simulations and rows represent
# random variables in each simulation
sims <-
    vapply(
        1:numberOfSims,
        function(x) rexp(numberPerSim, lambda),
        FUN.VALUE = numeric(40))
```

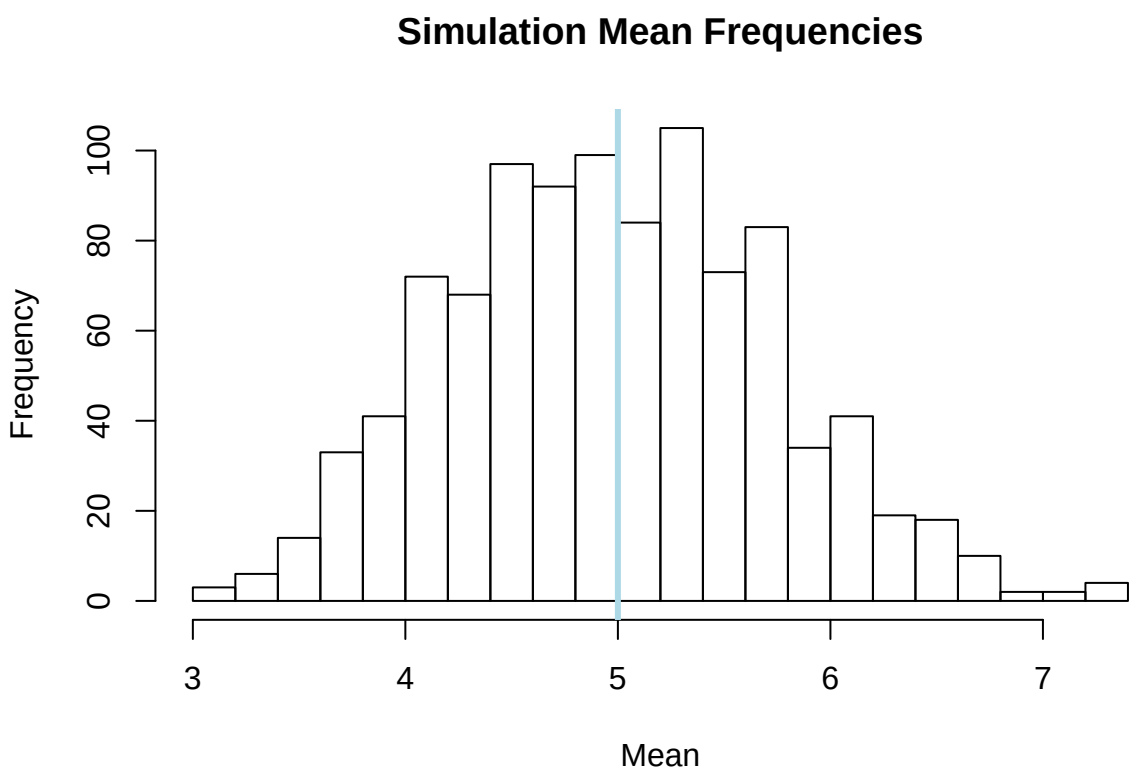
Next, we calculate the means and standard deviations of each simulation.

```
# Calculate the average of each simulation (each column represents a single
# simulation)
simMeans <- vapply(1:numberOfSims, function(i) mean(sims[,i]), FUN.VALUE = numeric(1))

# Calculate the standard deviation of each simulation
simSigmas <- vapply(1:numberOfSims, function(i) sd(sims[,i]), FUN.VALUE = numeric(1))
```

## Sample Mean vs Theoretical Mean

We compare the sample mean vs the theoretical mean in two ways. First, we demonstrate that the sample means are tightly clustered around the theoretical mean. Second, we show that the difference between the mean of the means and the theoretical mean is very small.

```r
# Calculate the theoretical mean
theoreticalMean <- 1 / lambda

# Show a histogram of the simulation means
# Add a vertical blue line at the theoretical mean
hist(
    simMeans,
    breaks = 20,
    xlab = "Mean",
    ylab = "Frequency",
    main = "Simulation Mean Frequencies")
abline(v = theoreticalMean, col = "lightblue", lwd = 3)
```




```r
# Show the difference between the average sample mean and the theoretical mean
mean(simMeans) - theoreticalMean
```

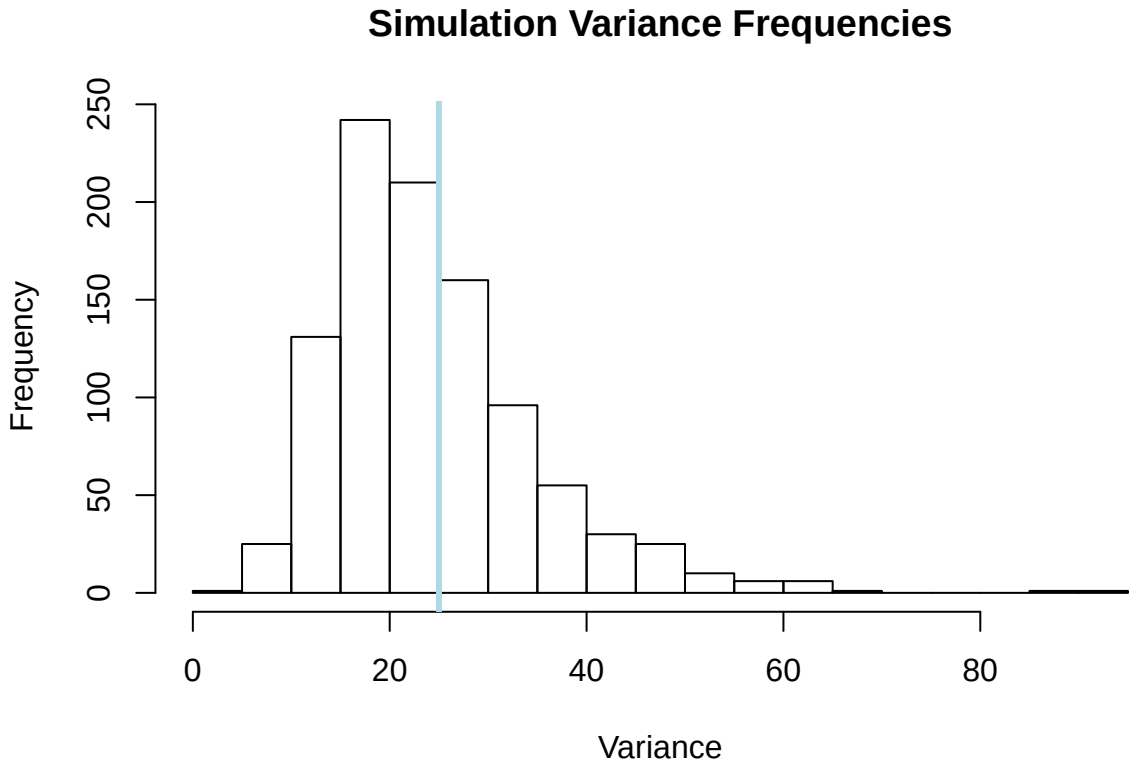
```
## [1] -0.02576123
```

### Sample Variance vs Theoretical Variance

We compare the sample variance vs the theoretical variance in two ways. First, we demonstrate that the sample variances are tightly clustered around the theoretical variance. Second, we show that the difference between the mean of the variances and the theoretical variance is very small.

```r
# Calculate the theoretical variance
theoreticalSD <- 1 / lambda
```

```r
theoreticalVariance <- theoreticalSD^2

# Show a histogram of the simulation variances
# Add a vertical blue line at the theoretical variance
hist(
    simSigmas^2,
    breaks = 20,
    xlab = "Variance",
    ylab = "Frequency",
    main = "Simulation Variance Frequencies")
abline(v = theoreticalVariance, col = "lightblue", lwd = 3)
```




```r
# Show the difference between the average sample variance and the theoretical variance
mean(simSigmas^2) - theoreticalVariance
```

```
## [1] -0.6219853
```

## Distribution

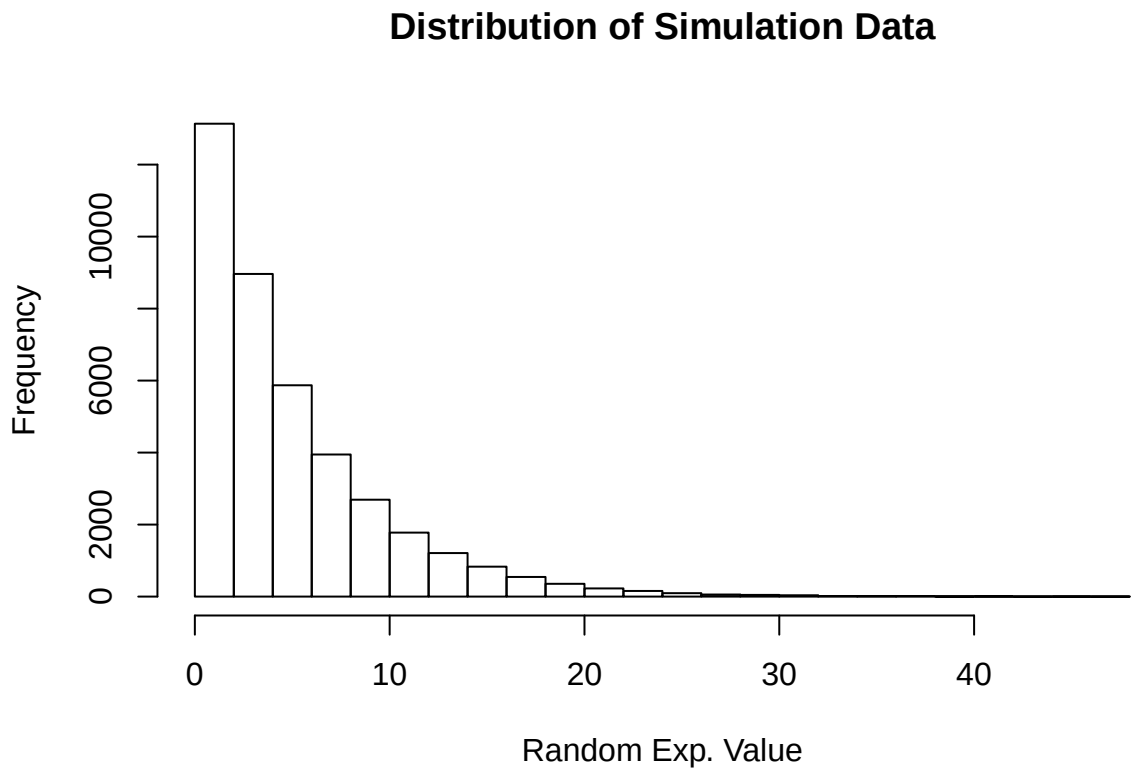
In this section, we demonstrate that the distribution of sample averages is approximately normal.

First, we display a histogram of the raw simulation data to show that the exponential distribution itself is not approximately normal.

Second, we show that the distribution of the sample means *is* approximately normal by plotting their density overlayed by the standard normal distribution. Since the two densities are very similar, we can conclude that

the distribution is approx. normal.

```
# Show a histogram of the raw simulation data
# This should demonstrate that the exponential distribution is itself not normal
hist(
    sims,
    breaks=20,
    xlab="Random Exp. Value",
    ylab="Frequency",
    main="Distribution of Simulation Data")
```




```
# Function to transform the sample mean to its CLT statistic
# It does so by first subtracting the theoretical mean from the sample mean, and then
# divides by the standard error
standardError <- theoreticalSD / sqrt(numberPerSim)
calcCLTStatistic <- function(m) (m - theoreticalMean) / standardError

# Create a data frame of the transformed sample means
dat <- data.frame(x = vapply(simMeans, calcCLTStatistic, FUN.VALUE=numeric(1)))

# Plot the density of the sample averages, overlayed by the standard normal dist
g <- ggplot(dat, aes(x = x)) +
    geom_histogram(
        aes(y = ..density..),
        fill="lightblue",
        color="black",
        binwidth=0.3) +
```

```
    stat_function(fun = dnorm) +
    labs(
        x = "x",
        y = "Density",
        title = "Density of Sample Means Overlayed By Standard Normal Distribution"
    )
print(g)
```