Simulation Exercise - Central Limit Theorem

Jason Battles

September 21, 2016

Overview

This study investigates a sample exponential distribution in R and compares it with the Central Limit Theorem.

The **Central Limit Theorem (CLT)** is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. *(source: Investopedia http://www.investopedia.com/terms/c/central_limit_theorem.asp*

To explore the validity of this theorem, we create an exponential distribution and compare it to the Central Limit Theorem.

Simulation Parameters

The exponential distribution is simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda, standard deviation is also 1/lambda, and lambda = 0.2 for all of the simulations. The study investigates distribution of averages of 40 exponentials and performs a thousand simulations.

To begin, load any required environments and libraries. Also, create the simulation variables as described above.

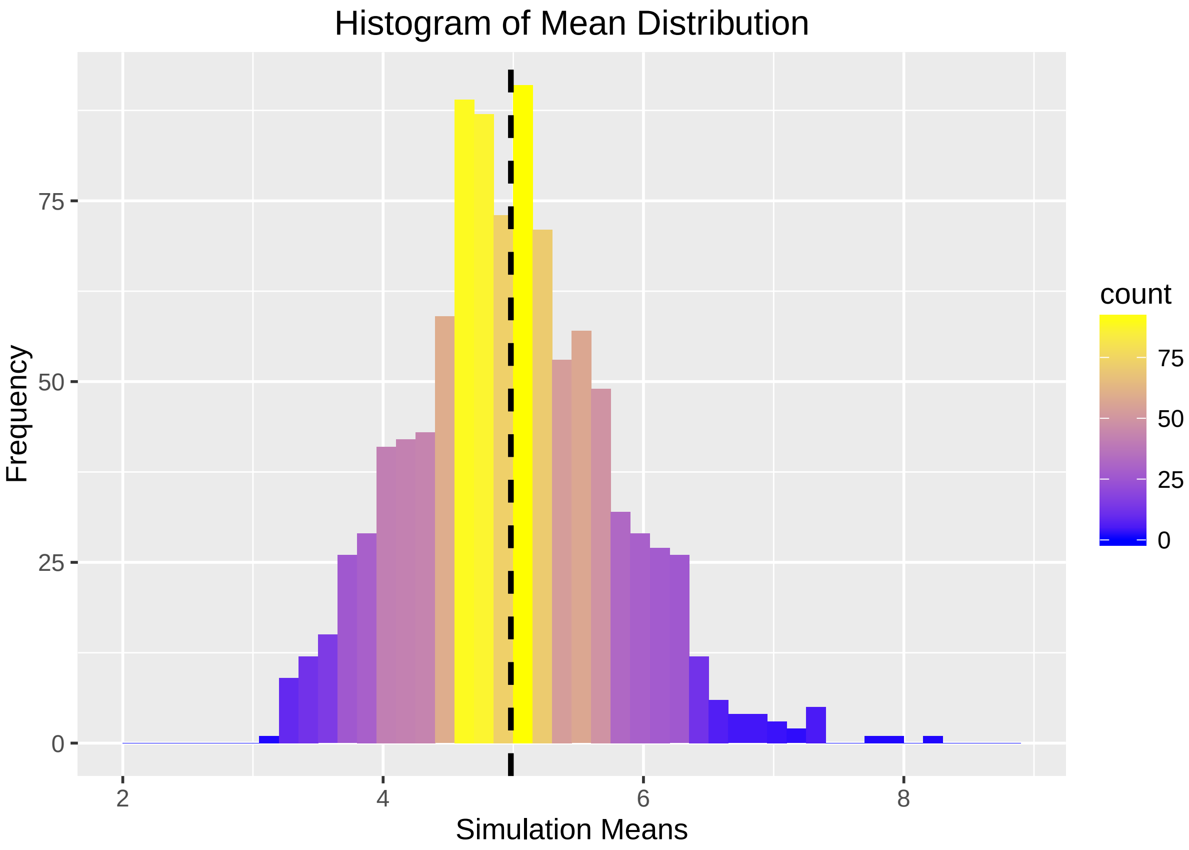
```r
library(ggplot2)
set.seed(21)     # Set seed so analysis can be reproduced
lambda <- 0.2
numExp <- 40     # Number of Exponentials
numSim <- 1000  # Number of Simulations to be performed
```

Simulations

Now with the variables and environment now configured, we can move forward with the simulation. To perform the simulation, the following computational steps are performed.

1. Create a Simulation Matrix with a thousand rows (1 for each simulation) and forty columns (1 for each exponential)
2. Generate vector with the mean of each row of the Simulation Matrix.
3. Generate a data frame combining Simulation Matrix with means from each observation row.
4. Visualize the simulated data with a histogram plot. *The R code for this graph is included in Appendix.*

```r
simMatrix <- matrix(rexp(n = numSim * numExp, rate = lambda), numSim, numExp)
simMean <- rowMeans(simMatrix)
simData <- data.frame(cbind(simMatrix, simMean))
```



Sample Mean versus Theoretical Mean

Calculate the actual mean of the simulated mean samples (4.91144) and the theoretical mean (5). The difference is very small (0.0186).

```
actMean <- mean(simMean)
theoMean <- (1 / lambda)
theoMean - actMean
```

```
## [1] 0.01855962
```

Sample Variance versus Theoretical Variance

Calculate the actual variance of the simulated mean samples (0.6047) and the theoretical variance (0.6250). Again, the difference is quite small (0.0204).

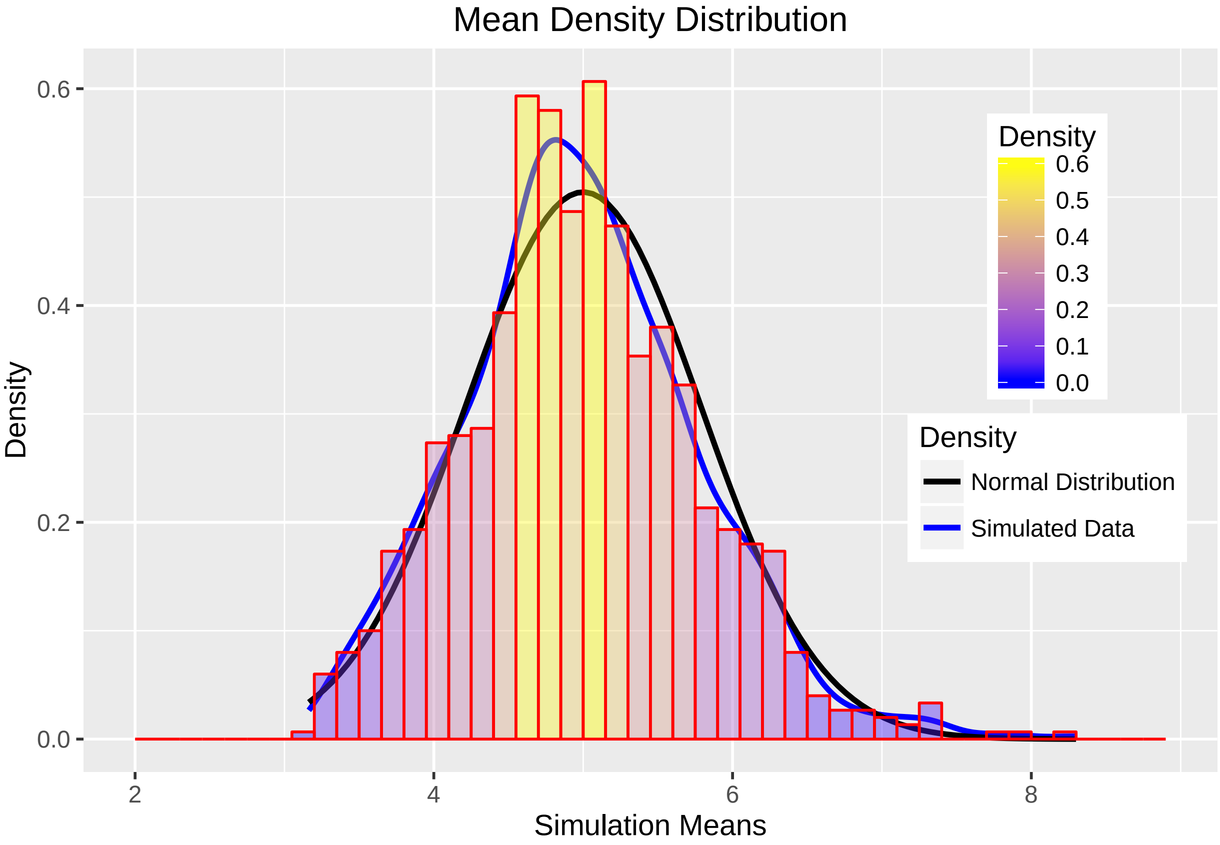
```
actVar <- var(simMean)
theoVar <- ((1 / lambda)^2/numExp)
theoVar - actVar
```

```
## [1] 0.02035499
```

Distribution Comparison

In an attempt to demonstrate the **Central Limit Theorem**, we compare a standard normal distribution with the simulated means.

We see that the simulated means from 1000 observations, closely aligns with a normal distribution. *The actual R code for this graph is included in the Appendix.*



The visual comparison between the black line (Normal Distribution) and the blue line (Simulated Means) creates some comfort level that the Central Limit Theorem is true. The Simulated Means may indeed be adequately approximated with a Normal Distribution.

Appendix

R Code for Visualizing the Simulated Means

```
ggplot(data = simData, aes(simData$simMean)) +
    geom_histogram(breaks = seq(2, 9, by = 0.15), aes(fill = ..count..)) +
    labs(title = "Histogram of Mean Distribution", x = "Simulation Means", y = "Frequency") +
    geom_vline(aes(xintercept=mean(simData$simMean)), color="black",
               linetype="dashed", size=1) +
    scale_fill_gradient(low="blue", high="yellow")
```

R Code for Comparing Normal Distribution with Simulated Means

```
qplot(simMean, geom = 'blank') +
    geom_line(aes(y=..density.., colour='Simulated Data'), stat='density', size=1) +
    stat_function(fun=dnorm, args=list(mean=(1/lambda), sd=((1/lambda)/sqrt(numExp))),
                  aes(colour='Normal Distribution'), size=1) +
    geom_histogram(aes(y=..density.., fill=..density..), alpha=0.4,
                   breaks = seq(2, 9, by = 0.15), col='red') +
    scale_fill_gradient("Density", low = "blue", high = "yellow") +
    scale_color_manual(name='Density', values=c('black', 'blue')) +
    theme(legend.position = c(0.85, 0.60)) +
    labs(title = "Mean Density Distribution", x = "Simulation Means", y = "Density")
```

A Mathematical Comparison of Confidence Intervals

For a more scientific comparison, we can mathematically compare the 95% Confidence Intervals between the Normal Distribution with the Simulated Means.

The 95% Confidence Interval of the theoretical normal distribution is [4.755, 5.245].

```
theoConfInterval <- theoMean + c(-1,1) * 1.96 * sqrt(theoVar)/sqrt(numExp)
theoConfInterval
```

```
## [1] 4.755 5.245
```

The 95% Confidence Interval of the actual simulated data is [4.9433, 5.0195].

```
actConfInterval <- actMean + c(-1,1) * 1.96 * sqrt(actVar)/numExp
actConfInterval
```

```
## [1] 4.943339 5.019542
```

We can see that the mathematical differences between the theoritical and actual confidence intervals are very small.

```
theoConfInterval - actConfInterval
```

```
## [1] -0.1883385  0.2254577
```