Statistical Inference Project Part 1

Brandon Araujo

Overview:

The exponential distribution in R is investigated and compared with the Central Limit Theorem (CLT).
R can simulate exponential distributions with the following function: rexp(n, lambda). The mean of this distribution is $\mu = \frac{1}{\lambda}$ where lambda λ is the rate parameter. The standard deviation is $\sigma = \frac{1}{\lambda}$.

CLT explains that a sample consisting of at least 30 independent observations and fairly normally distributed data, the distribution can be notated as: $\bar{x}_{n}$ ~ $N(\mu, \frac{\sigma }{\sqrt{n}})$. This project will demonstrate that the sampling distribution of an exponential distribution with $n = 40$ and $\lambda = 0.2$ is approximately $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$ distributed.

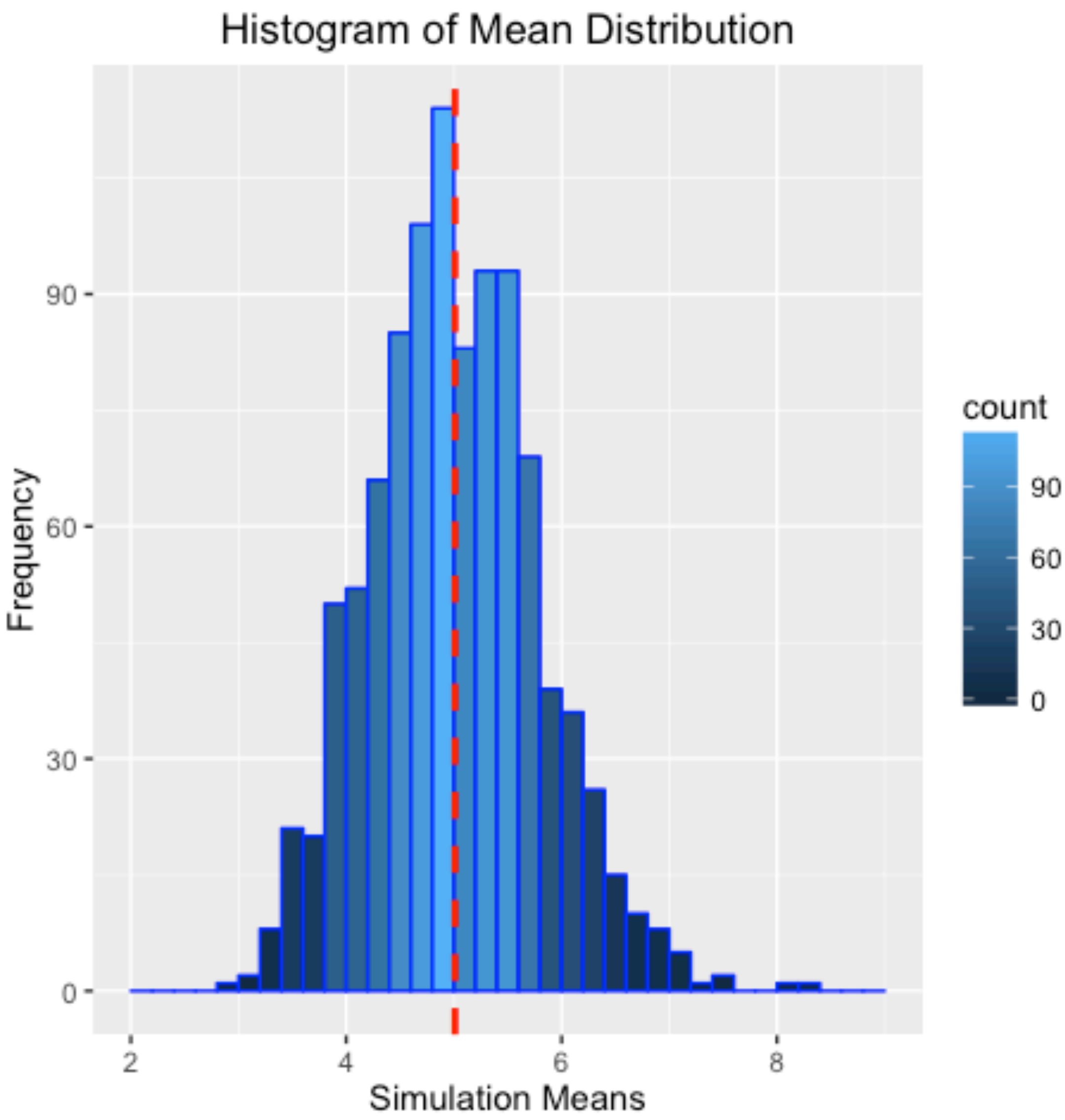
Simulations:

The exponential distribution can be simulated in R with rexp(n, lambda), where lambda is the rate parameter and n is the number of observations. For the purpose of all the simulations in this project, value of lambda is set to 0.2.

First we load the ggplot2 plotting library.

```
library(ggplot2)

#Create variables

Sims <- 1000

n <- 40

lambda <- 0.2

#Set random seed

set.seed(12)

#Create a matrix rows corresponding to 1000 simulations and columns corresponding to the 40
random simulations.

simMatrix <- matrix(rexp(n = Sims * n, rate = lambda), Sims, n)

#Vector containing the value of each simulations mean

simMean <- rowMeans(simMatrix)
```

```
#Create full data frame

simData <- data.frame(cbind(simMatrix, simMean))

#Create Visualization

ggplot(data = simData, aes(simData$simMean)) + geom_histogram(breaks = seq(2, 9, by = 0.2), col =
"blue", aes(fill = ..count..)) + labs(title = "Histogram of Mean Distribution", x = "Simulation
Means", y = "Frequency") + geom_vline(aes(xintercept=mean(simData$simMean)), color="red",
linetype="dashed", size=1)
```



Sample Mean Versus Theoretical Mean:

The actual mean of the simulated mean sample data is 5.01, calculated by:

`actualMean <- mean(simMean)` And the theoretical mean is 5, calculated by:

```
theoreticalMean <- (1 / lambda)
```

The two means are nearly equivalent. Thus demonstrating our initial intentions of this project.

Sample Variance Versus Theoretical Variance:

The actual variance of the simulated mean sample data is 0.615, calculated by:

```
actualVariance <- var(simMean)
```

And the theoretical variance is 0.625, calculated by:

```
theoreticalVariance <- ((1 / lambda) ^ 2) / n
```

Thus, the actual variance of the simulated mean sample data is very close to the theoretical

variance of original data distribution.

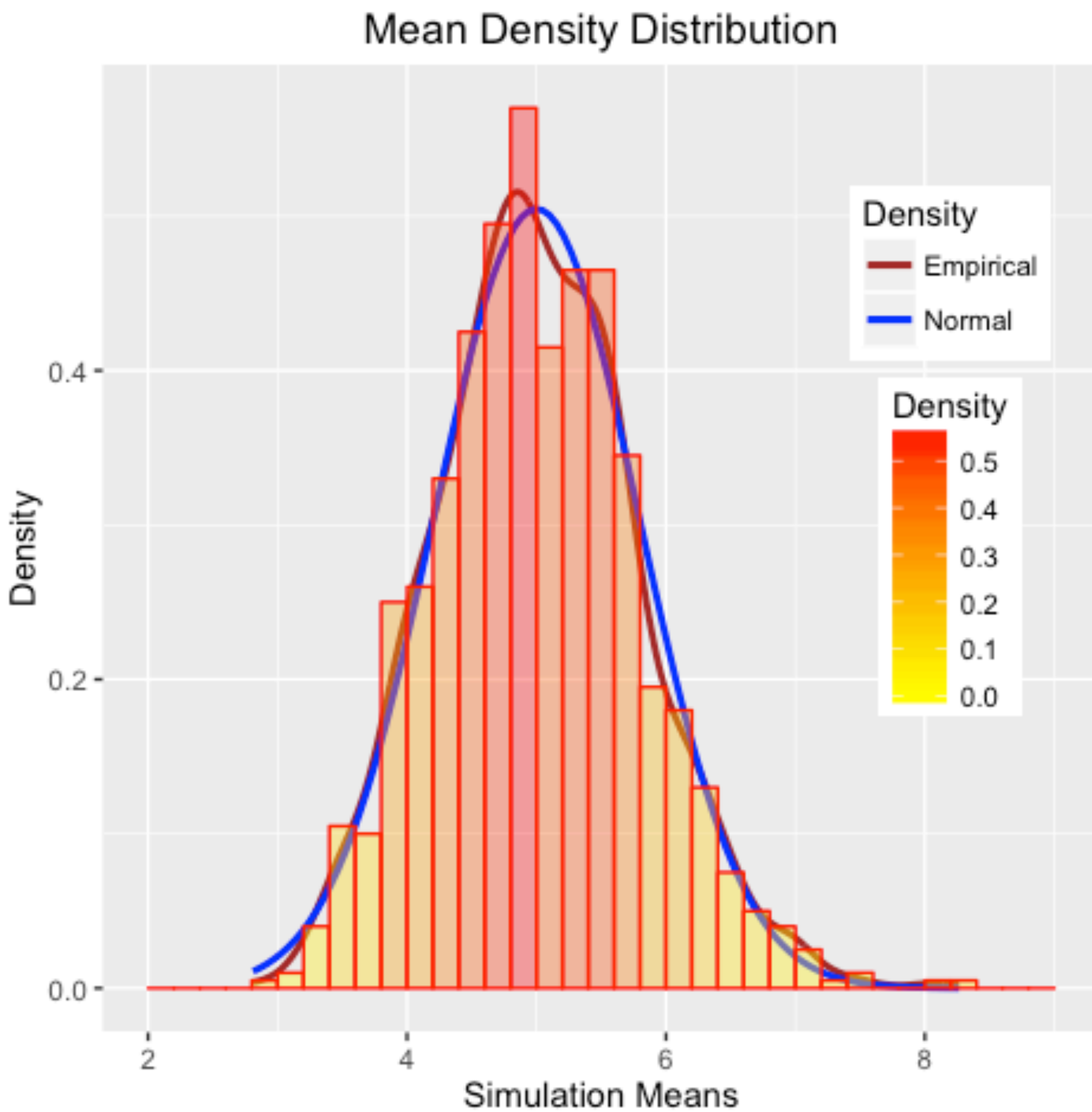
Distribution:

To prove that the simulated mean sample data approximately follows the Normal distribution, we perform the following three steps:

Step 1: Create an approximate normal distribution and see how the sample data alligns with it.

```
qplot(simMean, geom = 'blank') +
geom_line(aes(y=..density.., colour='Empirical'), stat='density', size=1) + stat_function(fun=dnorm,
args=list(mean=(1/lambda), sd=((1/lambda)/sqrt(n))),

aes(colour='Normal'), size=1) + geom_histogram(aes(y=..density.., fill=..density..), alpha=0.4,

breaks = seq(2, 9, by = 0.2), col='red') + scale_fill_gradient("Density", low = "yellow", high = "red") +
scale_color_manual(name='Density', values=c('brown', 'blue')) + theme(legend.position = c(0.85, 0.60)) +
labs(title = "Mean Density Distribution", x = "Simulation Means", y = "Density")
```

Mean Density Distribution



From above histogram, the simulated mean sample data can be adequately approximated with the normal distribution.

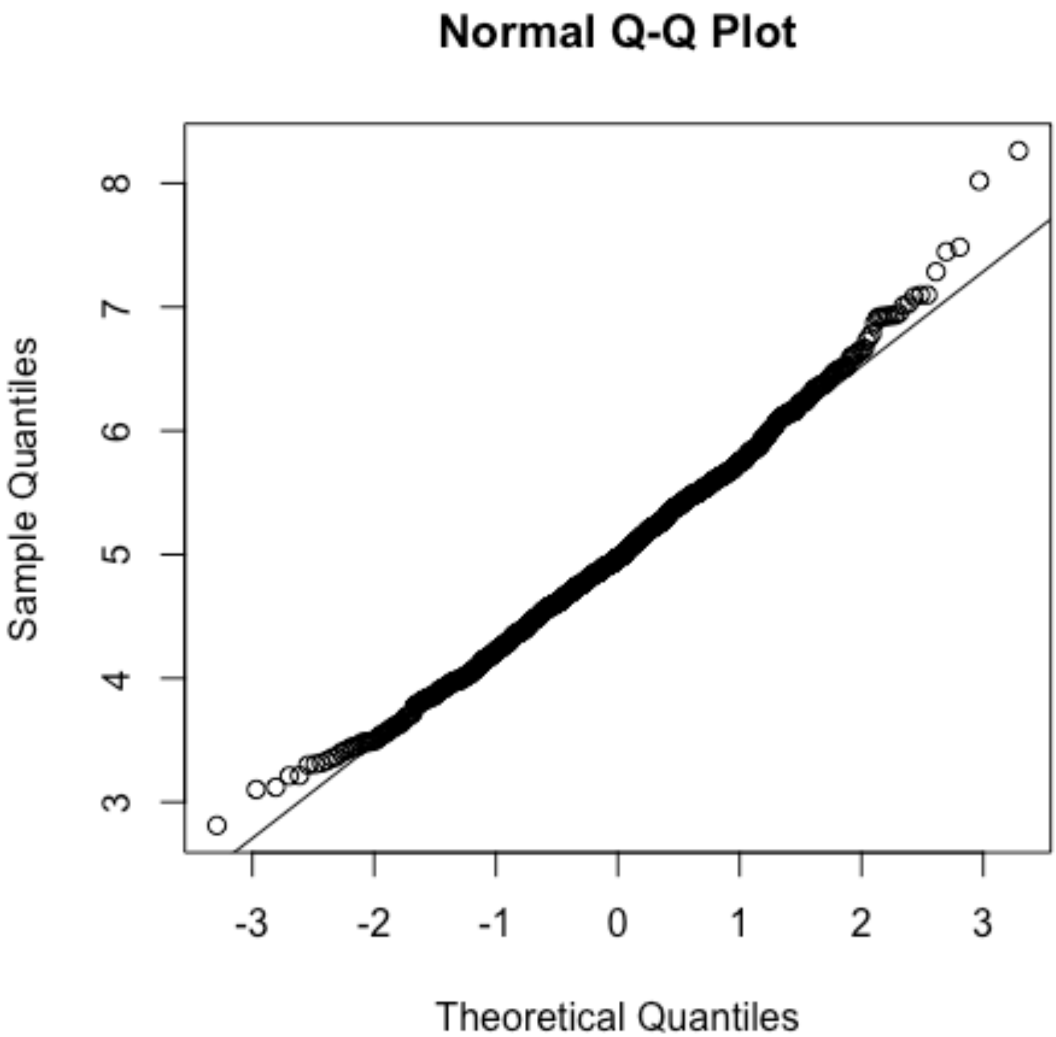
Step 2: Compare the 95% confidence intervals of the simulated mean sample data and the theoretical normally distributed data.

```
actualConfInterval <- actualMean+c(-1,1)*1.96*sqrt(actualVariance)/sqrt(sampSize)
theoreticalConfInterval <- theoreticalMean+c(-1,1)*1.96*

sqrt(theoreticalVariance)/sqrt(sampSize)
```

Actual 95% confidence interval is [4.77, 5.25] and Theoretical 95% confidence interval is [4.75,

5.25] and we see that both of them are approximately same.

Step 3: q-q Plot for Qunatiles.



The actual quantiles also closely match the theoretical quantiles, hence the above three steps prove that the distribution is approximately normal.