Central Limit Theorem to Exponential Distribution Comparison

Sebastian Cruz

December 18, 2015

Overview

The Central Limit Theorem (CLT) states that the distribution of averages of a iid variable will become more like that of a standard normal as the sample size increases. The following will show that this is true for samples from an Exponential Distribution are averaged and compared to a normal distribution.

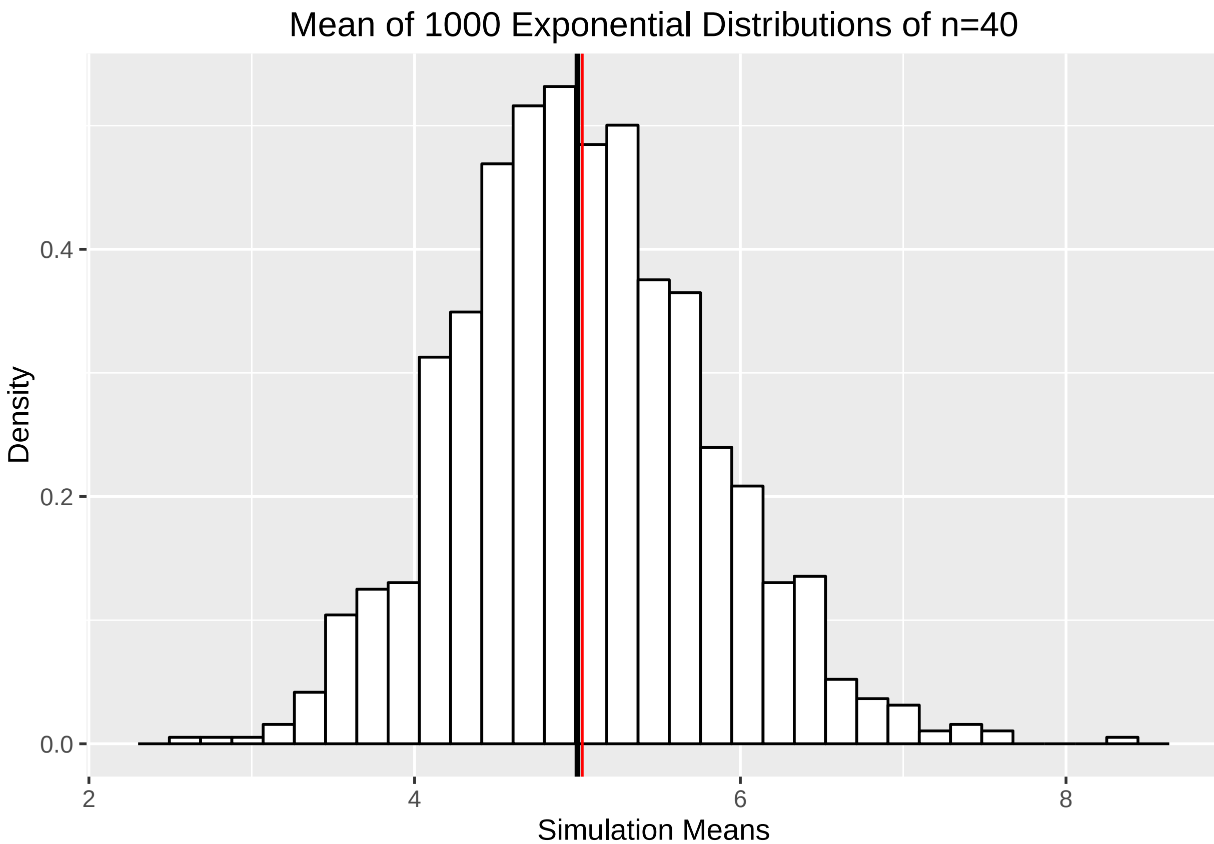
Simulations

Exponential distributions mean of 1 / lambda (aka the rate) and follow an exponential curve. For this exercise lambda is set at 0.2 so the theoretical mean is 1 / 0.2 or 5.

To verify the CTL, the R code below will take a sample of 40 observations from an exponential distribution (lambda = 0.2) and find the mean; repeating that process 1000 times. The resultant dataset of 1000 means will then be plotted to examine the distribution and other characteristics. The vertical solid black line denotes the theoretical mean of 5 (1/lambda or 1/0.2), and the vertical solid red line denotes the simulation mean.

```r
# Required Libraries
library(ggplot2)

# CLT Simulation Loop from the Mean of a Sample Exponential Distribution
set.seed(999) #For reporduceability
expMeans = NULL
for (i in 1:1000) {
  expMeans = c(expMeans, mean(rexp(40, rate=0.2)))
}
expMeansPlot = data.frame(expMeans) #ggplot requires data frames
colnames(expMeansPlot) = c("means")

# Plot the Means Distribution
ggplot(expMeansPlot, aes(x=means)) +
  geom_histogram(colour="black", fill="white", aes(y=..density..)) +
  #stat_function(fun=dnorm, args=list(mean=mean(expMeansPlot$means),sd=sd(expMeansPlot$means)),size = 1
  geom_vline(xintercept = 5, size = 1, colour = "black") + #Theoretical mean line
  geom_vline(xintercept = mean(expMeansPlot$means), size = 0.5, colour = "red") + #Simulation mean line
  ggtitle("Mean of 1000 Exponential Distributions of n=40") +
  xlab("Simulation Means") +
  ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Simulated Mean vs Theoretical Mean

The simulation output has a mean of 5.0290278 which is very close to the theoretical mean of 5 for an exponential distribution, denoted by the vertical solid black line.

```r
# Mean of Simulation Output
mean(expMeans)
```

```
## [1] 5.029028
```

```r
# Theoretical Mean for an Exponential Distribution
1/0.2
```

```
## [1] 5
```

Sample Variance vs Theoretical Variance

Exponential distributions have a varience of (1 / lambda)^2 / n or 0.625. The variance of the simulated output is 0.605616.

```r
# Variance of the Simulation Output
var(expMeans)
```

```
## [1] 0.605616
```

```r
# Variance of the Theoretical Exponential Distribution
(1/0.2)^2/40
```

```
## [1] 0.625
```

Distribution

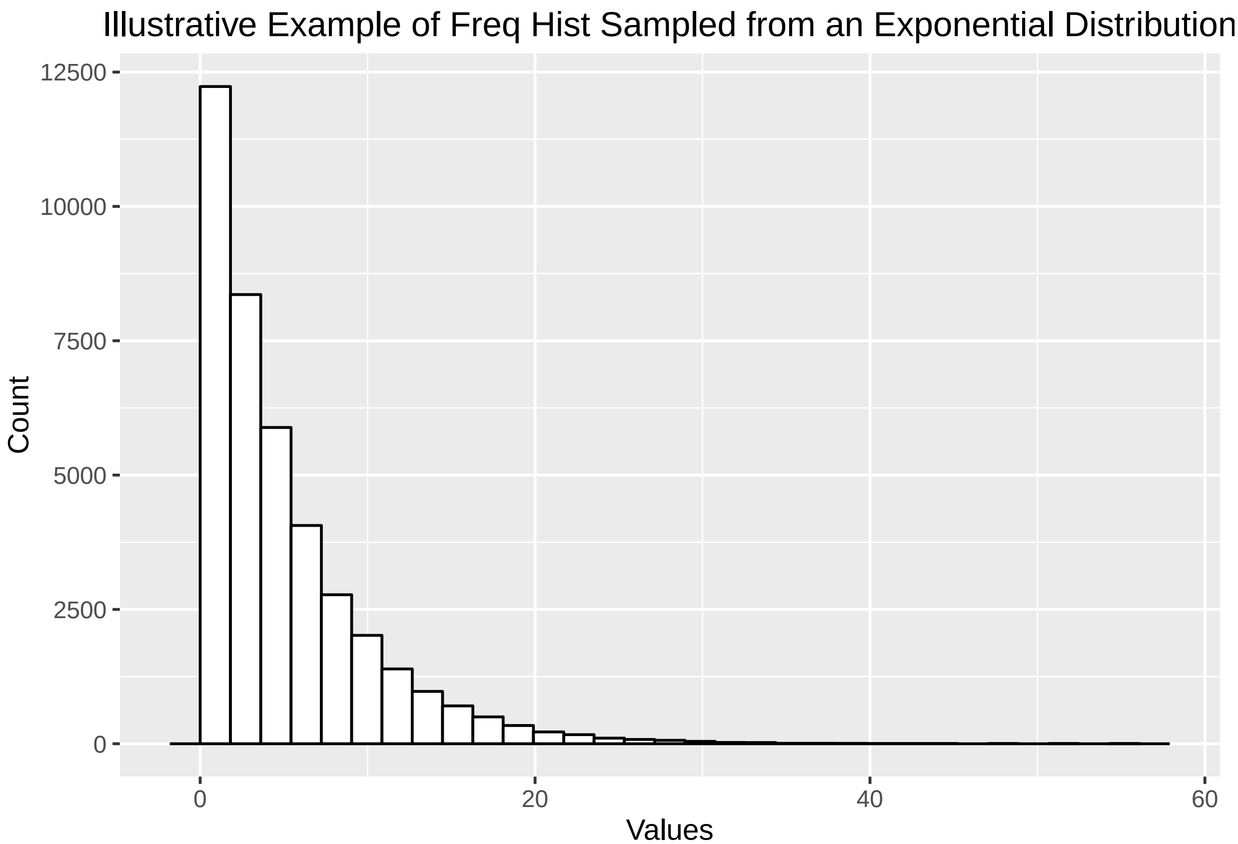
The below illustrates plot demonstrate two tenants of the Central Limit Theorem. That the sampled means from an exponential distribution will become more normal than that original distribution it was taken from and that the more samples taken will result in a more normal distribution.

The first plot is a frequency historgram of 40K samples from an Exponential Distribution with a lambda, or rate, of 0.2. The frequency histogram clearly shows a non-normal, exponential distribution. This demonstates my first point.

To demonstrate that the normality of the samples increases with size, a low resolution 100 simulations plot was created to contrast that of the 1000 run simulation. Both plots of the simulation data have been overlayed with that of a normal distribution denoted by the solid black curve. The low resultion simulation data has a shape that roughly follows that of the normal distribution. The 1000 simulation plot adheres very closely to that normal distribution line as predicted by the Central Limith Theorem.

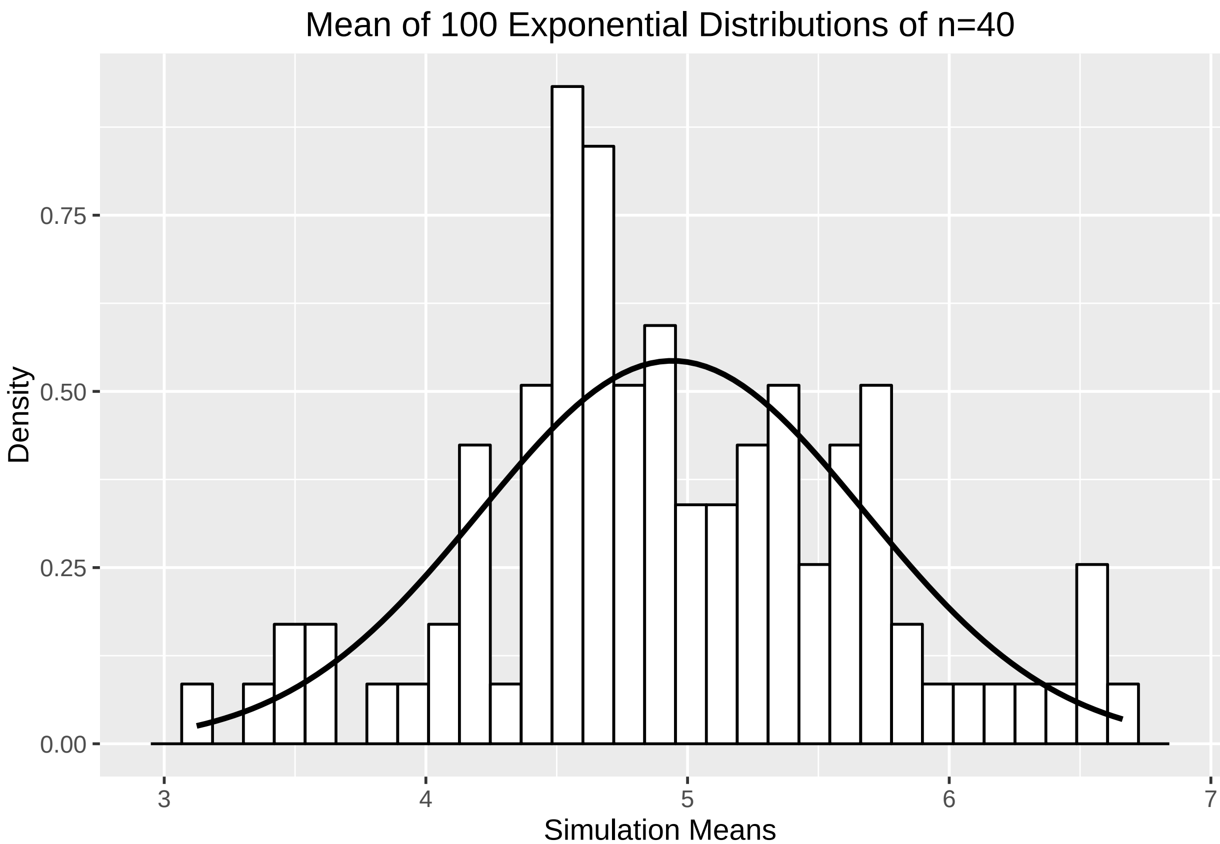
```r
# Illustrative Example of a Large Sample Exponential Distribution
expDist = NULL
set.seed(999) #For reporduceability
expDist = data.frame(rexp(40000, rate=0.2))
colnames(expDist) = c("values")

# Plot the Frequency Histogram of the Sampled Exponential Distribution
ggplot(expDist, aes(x=values)) +
  geom_histogram(colour="black", fill="white") +
  ggtitle("Illustrative Example of Freq Hist Sampled from an Exponential Distribution") +
  xlab("Values") +
  ylab("Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



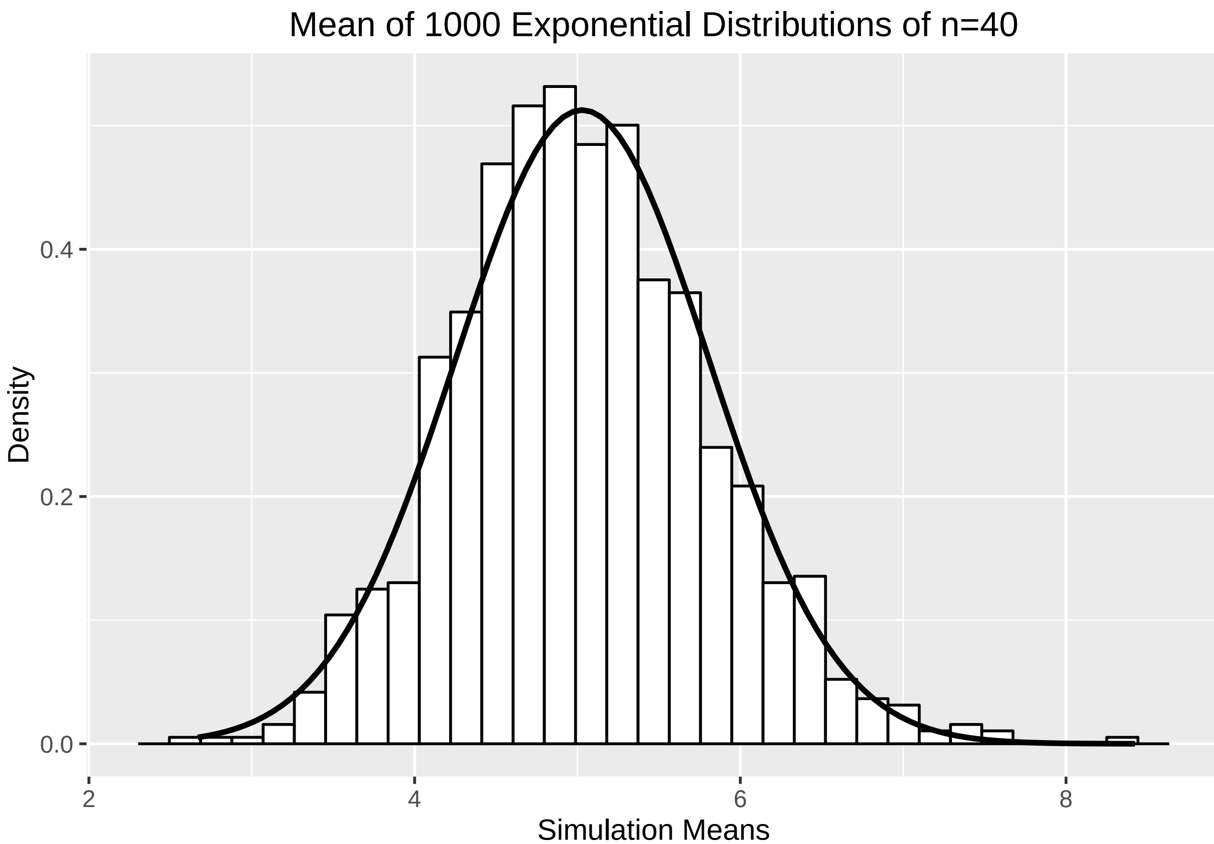

```r
# Low Resolution 100 Simulations for Contrast
set.seed(999) #For reporduceability
expMeans100 = NULL
for (i in 1:100) {
  expMeans100 = c(expMeans100, mean(rexp(40, rate=0.2)))
}
expMeansPlot100 = data.frame(expMeans100)
colnames(expMeansPlot100) = c("means")

ggplot(expMeansPlot100, aes(x=means)) +
  geom_histogram(colour="black", fill="white", aes(y=..density..)) +
  stat_function(fun=dnorm, args=list(mean=mean(expMeansPlot100$means),sd=sd(expMeansPlot100$means)),size
  ggtitle("Mean of 100 Exponential Distributions of n=40") +
  xlab("Simulation Means") +
  ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Higher Resolution 1000 Simulations
ggplot(expMeansPlot, aes(x=means)) +
  geom_histogram(colour="black", fill="white", aes(y=..density..)) +
  stat_function(fun=dnorm, args=list(mean=mean(expMeansPlot$means),sd=sd(expMeansPlot$means)),size = 1,
  ggtitle("Mean of 1000 Exponential Distributions of n=40") +
  xlab("Simulation Means") +
  ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Mean of 1000 Exponential Distributions of n=40
Density
0.4
0.2
0.0
2
4
6
8
Simulation Means