# Demonstration of the Central Limit Theorem with Simulations

*Julie Grantier*

*November 18, 2015*

---

## Overview

In this paper, we are demonstrating the Central Limit Theorem by taking samples from a theoretical exponential distribution. Though the distribution itself is exponential in shape, we will show that means of samples from the distribution approximate a normal distribution with a mean of the population mean and a variance equal to the population variance divided by the sample size.

## Population Distributions

In research we are rarely able to access the entire population to complete a study. Instead we have to take samples from the population of interest and infer information about the population from these samples. We use statistical methods to estimate population parameters from sample statistics. To determine good, unbiased estimators, statisticians use theoretical populations where can create actual population parameters and compare them to sample estimators. Programs like R allow us to easily sample from theoretical, mathematical populations.

For this short paper, we will be using a known exponential distribution to look at properties of sample means and variances. This will demonstrate properties of the Central Limit Theorem. From Hays (1988), p. 232:

> If a population has a finite variance $\sigma^2$ and a mean $\mu$, then the distribution of sample means from samples of $N$ independent observations approaches a normal distribution with variance $\sigma^2/N$ and mean $\mu$ as the sample size $N$ increases. When $N$ is very large, the sampling distribution of $\bar{x}$ is approximately normal.

## Single Sample

The general exponential distribution has a probability density function of:

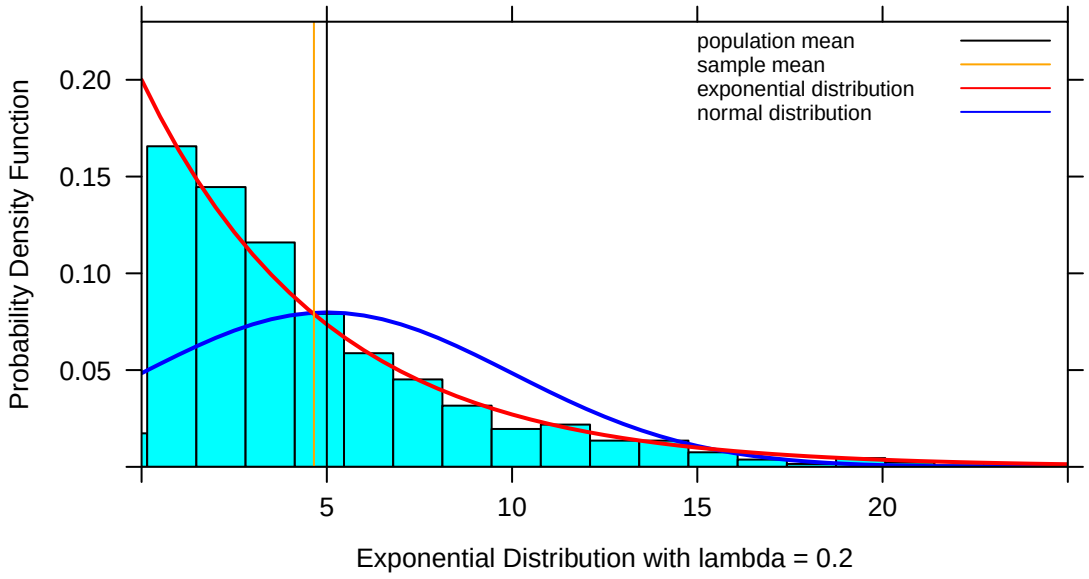
$f(x) = \lambda e^{-\lambda x}$ with a mean of $\frac{1}{\lambda}$ and a variance of $\frac{1}{\lambda^2}$

In this example, we will set $\lambda = 0.2$ for all of the simulations and drawn samples. First a single sample of 1000 observations (or 1000 samples of size 1) was drawn from the exponential distribution and graphed (code).

Figure 1 shows a histogram of this single sample from the distribution ($N = 1000$). The theoretical mean of $\frac{1}{0.2} = 5$ is marked in black and the actual mean of the sample, 4.65, is marked in orange. The theoretical standard deviation is also $\frac{1}{0.2} = 5$, and the actual standard deviation of this sample is 4.5. These estimators will vary around the expected values for each different sample taken.

Two theoretical distributions are graphed on Figure 1 below. The red line is the exponential distribution with mean and standard deviation of $\frac{1}{0.2} = 5$. As expected the histogram of the single sample closely follows this theoretical distribution. The blue line is a normal distribution with the same mean and standard deviation. This has been included to show that individual samples do not look close to the normal distribution.

**Figure 1 Histogram Single Sample (N=1000)**




## Sampling Distribution

For the sampling distribution, we took 1000 samples, each of size 40, from the exponential distribution ($\lambda = 0.2$). The mean of each sample was calculated and saved in a vector, and then the 1000 means were plotted in the histogram in Figure 2 below (code).

### Means

Figure 2 again has vertical lines marking the populaton mean 5 and the mean of the sample means 4.967. As expected from the CLT the mean of the sampling distribution is very close to its expected value, the population mean.

### Variances

The variance of the sample means is 0.638, which is not close at all to the population variance 25. The variance of a sampling distribution predicted by the Central Limit Theorem $\sigma^2/N$. So in our case the variance should be $\sigma^2/N = 25/40 = 0.625$. Our value of 0.638 is close to this prediction.

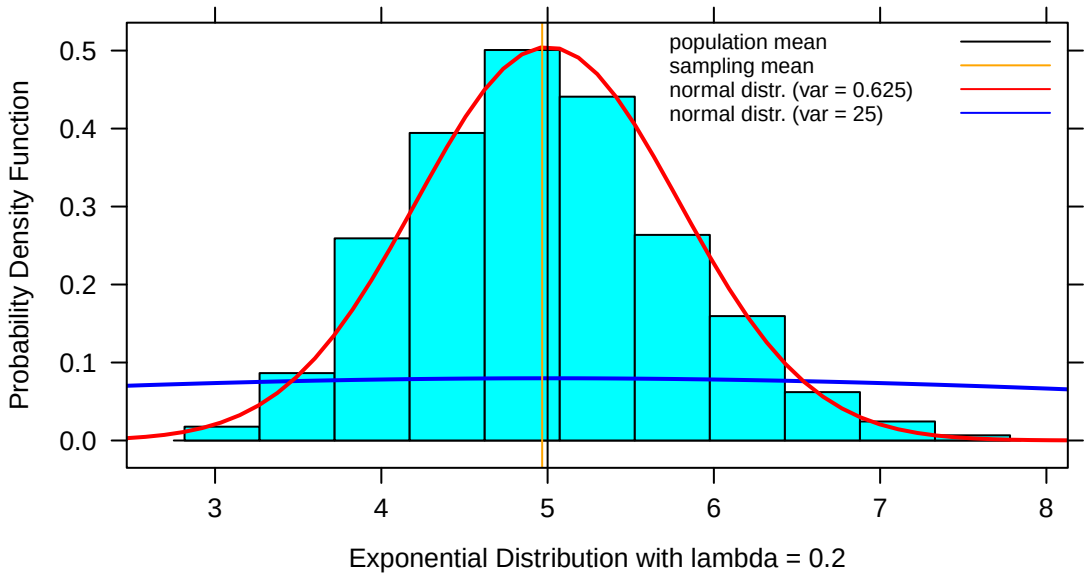
To demonstrate this, two normal distributions have been graphed on Figure 2. The first is in blue and represents the same normal curve as on Figure 1 with a mean of 5 and a variance of 25, the population values. (Note that the axes are different on the two figures.) It is clear that this curve has a variance that is much too large to represent the sampling distribution. As samples get larger, the sample means get closer and closer to the population mean. Because these means cluster closer together, it is sensible that the variance decreases as sample size increases. The second normal distribution in red also has a mean of 5, but its variance has been adjusted by sample size as calculated above. That this normal curve seems to fit the histogram so well supports the CLT's adjustment of the variance.

### Normality

Figure 2 shows that the sampling distribution closely matches the normal curve expected from the CLT. Extensive testing of the normality of these sample means is beyond the scope of this paper, but a couple of descriptive demonstrations are useful.

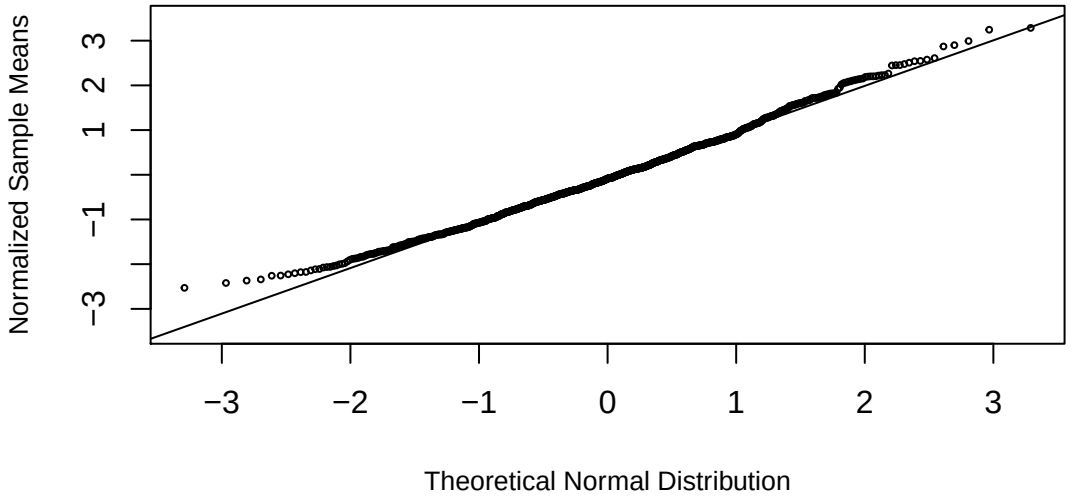
Figure 3 shows a Quantile-Quantile plot which plots the theoretical normal distribution versus the z-scores of the sample means (code). If the sample means were completely normal, all points would lie on the diagonal line. From the plot we can see that this is mostly true with some small deviation in the left tail. This deviation in the left tail can also be seen by comparing the histogram in Figure 2 to the normal curve.

**Figure 2 Means of 1000 Samples of Size 40**




.

**Figure 3 Normal Q−Q Plot**




Other measures of normality include skew and kurtosis. For a normal curve, these values would be expected to be zero. Our sample means have a skew of 0.298 and a kurtosis of -0.054. These numbers are close to zero, although again statistical tests of these numbers is beyond the scope of this paper.

## Conclusion

The simulation from this paper does demonstrate three parts of the Central Limit Theorem: the sample means will be centered around the population mean, the variance of the sample means will equal the population variance divided by sample size, and the distribution of the sample means will approximate a normal curve.

# Appendix

## Code

### Setup Code:

```
knitr::opts_chunk$set(warning = FALSE, fig.width=6, fig.height=3.6)
library(stats)          #for Shapiro-Wilk normality test
library(psych)          #for skew and kurtosis
library(lattice)
library(latticeExtra)   #for graph options
trellis.par.set(superpose.line = list(col = c("black","orange","red","blue"))) #for custom legend
set.seed(747)
```

### Single Sample Creation

```
single_sample <- rexp(1000,0.2)             #take a single sample of size 1000
mean_single_sample <- mean(single_sample)
sd_single_sample <- sd(single_sample)
```

### Figure 1

```
histogram(single_sample,type = "density",nint= 24, xlim = c(0,25), ylim = c(0,0.23),
          main = list(label = "Figure 1 Histogram Single Sample (N=1000)", cex = 0.9),
          xlab = list(label = "Exponential Distribution with lambda = 0.2", cex = 0.8),
          ylab = list(label = "Probability Density Function", cex = 0.8),
          panel = function(x, ...) {
                panel.histogram(x,...)
                panel.mathdensity(dmath = dnorm,args = list(mean=(1/0.2),sd=(1/0.2)), lwd=2,
                                  col.line = "blue")
                panel.mathdensity(dmath = dexp,args = list(rate = 0.2), lwd=2,
                                  col.line = "red")
                panel.abline(v = mean(single_sample), col.line = "orange")
                panel.abline(v = 1/0.2, col.line = "black")       #population mean
                panel.key(c("population mean", "sample mean","exponential distribution",
                            "normal distribution"), cex =0.7,corner = c(1,.98), lines = TRUE,
                            points= FALSE)})   #custom legend
```

### Sampling Distribution Creation

```
num_samples <- 1000
sample_size <- 40
lambda <- 0.2

sample_means <- vector(mode = "numeric",length = num_samples)
for(i in 1:num_samples){                  #1000 times
      sample <- rexp(sample_size,0.2)     #create a random sample of 40
      sample_means[i] <-  mean(sample)    #take the mean and store in vector
}
sampling_mean <- mean(sample_means)       #mean of sample means
sampling_variance <- var(sample_means)    #variance of sample means
```

**Figure 2**

```
histogram(sample_means,type = "density",
          main = list(label = "Figure 2 Means of 1000 Samples of Size 40", cex = 0.9),
          xlab = list(label = "Exponential Distribution with lambda = 0.2", cex = 0.8),
          ylab = list(label = "Probability Density Function", cex = 0.8),
      panel = function(x, ...) {
            panel.histogram(x,...)
            panel.mathdensity(dmath = dnorm,args =
                  list(mean=1/lambda,sd=(1/lambda)),
                  lwd=2, col.line = "blue")
            panel.mathdensity(dmath = dnorm,args =
                  list(mean=1/lambda,sd=(1/lambda)/(sqrt(sample_size))),
                  lwd=2, col.line = "red")
            panel.abline(v = sampling_mean, col.line = "orange")
            panel.abline(v = 1/0.2, col.line = "black")     #population mean
            panel.key(c("population mean", "sampling mean","normal distr. (var = 0.625)",
                        "normal distr. (var = 25)"), corner = c(1,.98), cex = 0.7,
                  lines = TRUE, points= FALSE)})        #custom legend
```

**Figure 3**

```
zsample_means = (sample_means - 1/lambda)/((1/lambda)/(sqrt(sample_size)))
      #find z scores for the sample means so they can be plotted versus the normal distribution
qqnorm(zsample_means, ylim =c(-3.5,3.5),cex = 0.4,cex.main = 0.9, cex.lab = 0.8,
       main = "Figure 3 Normal Q-Q Plot",
       xlab = "Theoretical Normal Distribution",
       ylab = "Normalized Sample Means",
       plot.it = TRUE, datax = FALSE)
qqline(zsample_means, datax = FALSE, distribution = qnorm)  #add line to the qq plot

skew = skew(sample_means)
kurtosis = kurtosi(sample_means)
library(stats)
shapiro_wilk <- shapiro.test(sample_means)
```

## References

Hays, William L. 1988. *Statistics*. Fort Worth, TX: Hartcourt Brace Jovanovich.