# Exploring the Central Limit Theorem using Samples from the Exponential Distribution

**Submission by Connor Lenio. Email: cojamalo@gmail.com. Completion Date: Apr 13, 2017.**

### Overview

The following report explores the Central Limit Theorem using an exponential distribution. One thousand samples of size forty are taken from the exponential distribution and the mean and variance of these samples are calculated. The mean sample estimates are compared to their theoretical equivalents and the sample distribution is tested for normality.

### Load packages

```
library(pander); library(ggplot2); library(dplyr)
```

### Simulations

Fist, a seed is set so this exact iteration of "random" sampling can be replicated in other R sessions:

```
set.seed(123)
```

### Simulate 1000 samples of 40 values from the exponential distribution with lambda = 0.2

The following code creates a data frame, `sim_samples` that contains the data for each sample of 40 values. The data frame is the mutated to calculate the mean, standard deviation, and variance for each of the 1000 samples.

```
#Initalize sim_samples
sim_samples <- NULL
#Use a for loop to take 1000 samples of 40 from the exponential distribution
for (i in 1 : 1000) {
    sample <- data.frame(sim = i, data = c(rexp(40, 0.2)))
    sim_samples <- rbind(sim_samples, sample)
}
#Mutate the simulated data to to calculate the mean, standard deviation, and variance for
#each of the 1000 samples
sim_samples <- sim_samples %>% group_by(sim) %>% summarize(x_bar_samp = mean(data), s_samp = sd(data),
                                                           var_samp = s_samp^2)
#Display a portion of the data frame produced by the code - sim_samples
pandoc.table(head(sim_samples), caption = "Table 1-1 - The summary statistics for the first
             6 simulations out of 1000 simulations of 40 samples from the exponential distribution ",
             justify = "center")
```

Table 1: Table 1-1 - The summary statistics for the first 6 simulations out of 1000 simulations of 40 samples from the exponential distribution

| sim | x_bar_samp | s_samp | var_samp |
|---|---|---|---|
| 1 | 4.811212 | 4.173642 | 17.41929 |

| sim | x_bar_samp | s_samp | var_samp |
|---|---|---|---|
| 2 | 5.360077 | 6.207536 | 38.53351 |
| 3 | 4.592871 | 4.149878 | 17.22149 |
| 4 | 4.900051 | 4.372193 | 19.11607 |
| 5 | 5.516619 | 5.124875 | 26.26434 |
| 6 | 5.612835 | 4.961209 | 24.61359 |

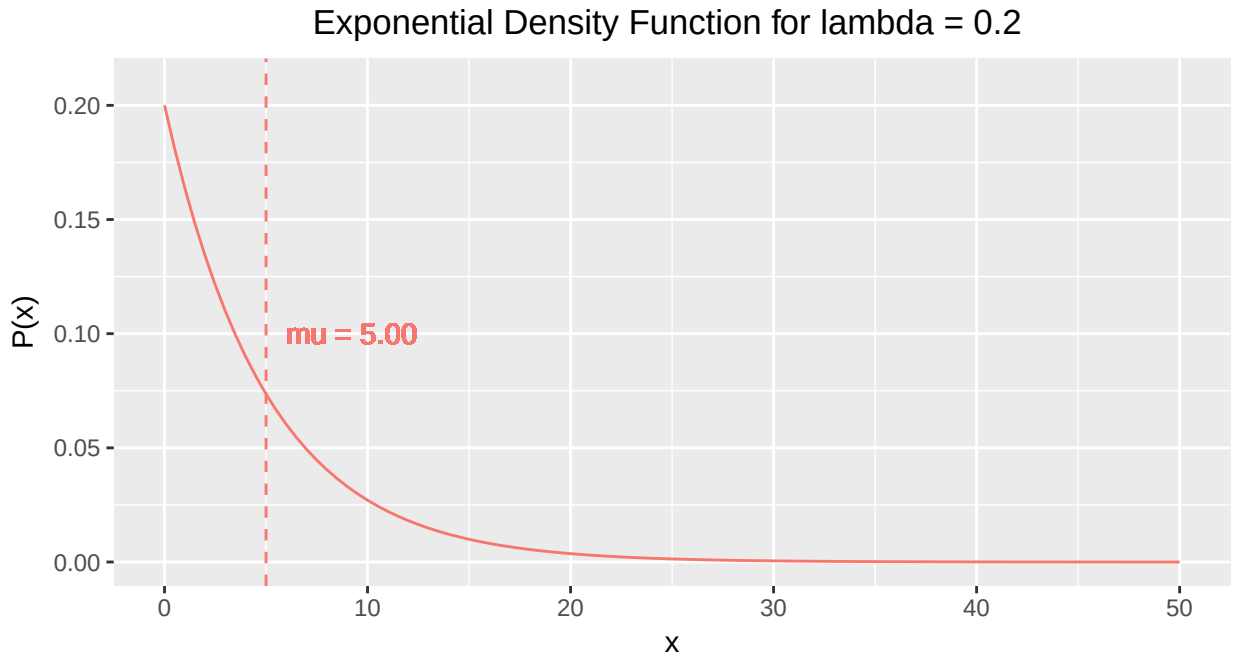
With the data simulated, the report continues with caparisons between the samples and the theoretical exponential distribution.

### Sample Mean versus Theoretical Mean

#### 1. Plot the theoretical (population) distribution and mean

The assignment gives the necessary information to calculate mu, the theoretical mean of the exponential distribution with lambda = 0.2, as mean = 1/lambda. With mu calculated, the density function can be graphed with the mean marked by a vertical line:

```
#Calculate mu from lambda
lambda <- 0.2
mu <- 1/lambda
#Plot the exponential density function with mu
ggplot(sim_samples, aes(x=sim)) +
    stat_function(fun=dexp, args=list(rate = lambda), geom = "path", color = "#F8766D") +
    geom_vline(xintercept = mu, color = "#F8766D", lty = 2) +
    geom_text(aes(x=6, label="mu = 5.00", y=0.10), colour="#F8766D", hjust = 0) +
    xlim(0,50) +
    ylim(0,0.21) +
    labs(title = "Exponential Density Function for lambda = 0.2 ", x = "x", y = "P(x)") +
    theme(plot.title = element_text(hjust = 0.5))
```



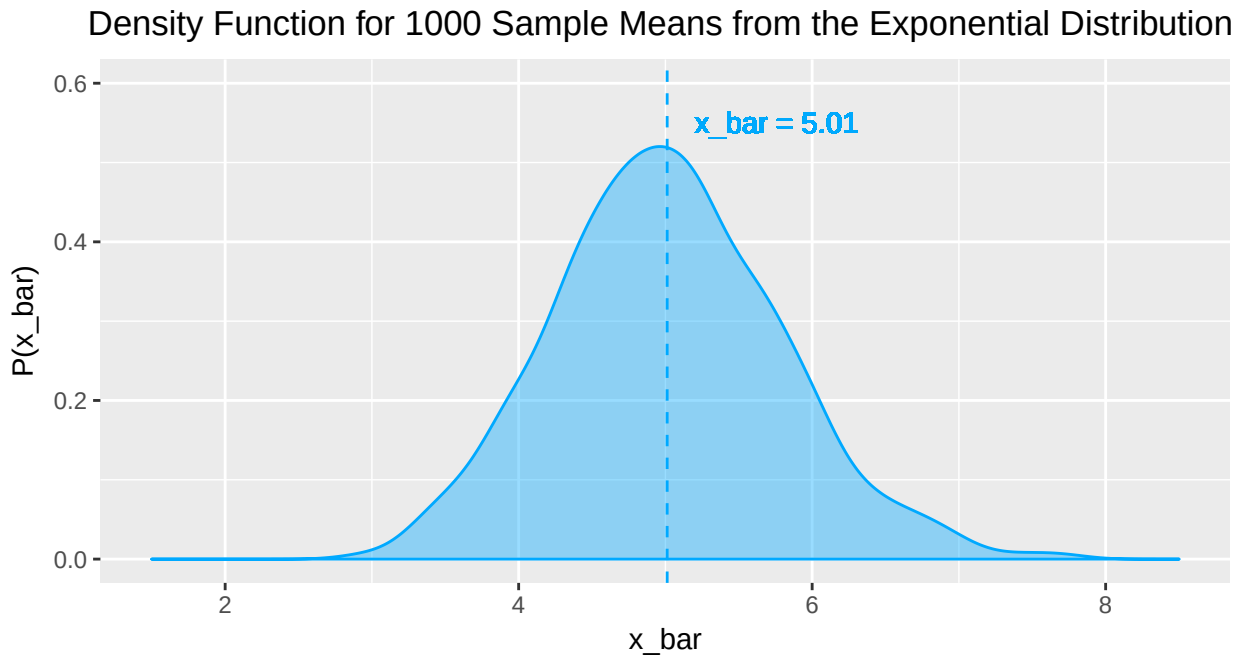

The Central Limit Theorem indicates that the distribution of the means of multiple samples of independent

random variables will be approximately normally distributed, regardless of the underlying distribution of the random variables. Thus, even though the underlying distribution is not normal, but exponential, one would expect the distribution of the sample means to be normal. The next step checks this expectation.

### 2. Plot the distribution of sample means

The mean of the sample means is calculated as the sample estimate, `x_bar`. The distribution of the sample means is then plotted with `x_bar` marked.

```
#Calculate x_bar from the sample means
x_bar <- mean(sim_samples$x_bar_samp)
#Plot the distribution of means using ggplot
ggplot(sim_samples, aes(x = x_bar_samp)) +
    geom_density(geom = "area", fill = "#00A9FF", alpha = 0.4,color = "#00A9FF") +
    geom_vline(xintercept = x_bar, color = "#00A9FF", lty = 2) +
    geom_text(aes(x=5.2, label=paste("x_bar =" , round(x_bar, digits = 2)), y=0.55), colour="#00A9FF",
              hjust = 0, text=element_text(size=12)) +
    ylim(0,0.6) +
    xlim (1.5,8.5) +
    labs(title = "Density Function for 1000 Sample Means from the Exponential Distribution ",
         x = "x_bar", y = "P(x_bar)") +
    theme(plot.title = element_text(hjust = 0.5))
```




As expected, this plot is normally distributed and is centered near to the theoretical mean of the distribution that the samples were taken from. The estimate of the sample mean for this distribution is about 5.01. This is almost exactly the same as the theoretical mean, mu, which is calculated as exactly 5.00.

### Sample Variance versus Theoretical Variance

### 1. Calculate the theoretical (population) variance

The assignment gives the necessary information to calculate sigma^2, the theoretical variance of the exponential distribution of lamda 0.2, as sigma^2 = 1/lambda.:

```
sigma <- 1/lambda
sigma_2 <- sigma^2
sigma_2
```

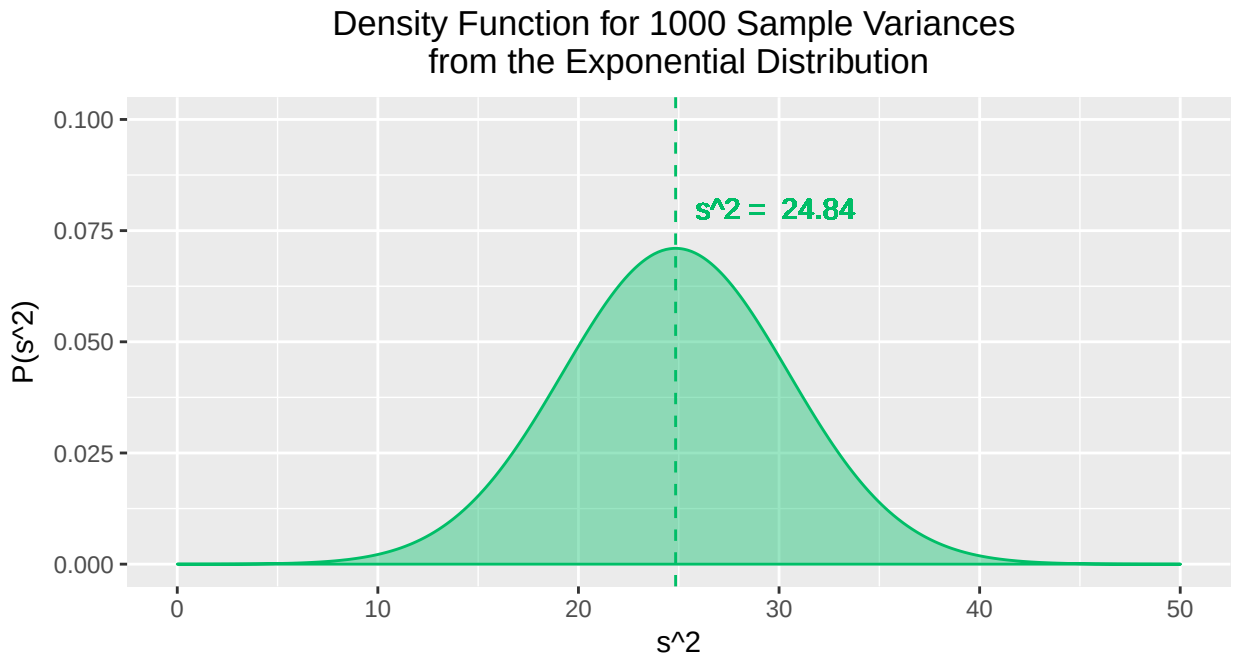
```
## [1] 25
```

The theoretical variance is exactly 5^2, or 25.

### 2. Plot the distribution of sample variance

The mean of the sample variances is calculated as the sample estimate, s^2‘. The distribution of the sample variances is then plotted with s^2 marked.

```
#Calculate the center of the distribution of sample variances
mean_var_samp <- mean(sim_samples$var_samp)
#Plot the distribution of sample variances
ggplot(sim_samples, aes(x = mean_var_samp)) +
    geom_density(geom = "area", fill = "#00BE67", alpha = 0.4,color = "#00BE67") +
    geom_vline(xintercept = mean_var_samp, color = "#00BE67", lty = 2) +
    geom_text(aes(x=mean_var_samp+1, label=paste("s^2 = " , round(mean_var_samp, digits = 2)), y=0.08),
              colour="#00BE67", hjust = 0, text=element_text(size=12)) +
    ylim(0,0.1) +
    xlim (0,50) +
    labs(title = "Density Function for 1000 Sample Variances \n from the Exponential Distribution ",
         x = "s^2", y = "P(s^2)") +
    theme(plot.title = element_text(hjust = 0.5))
```



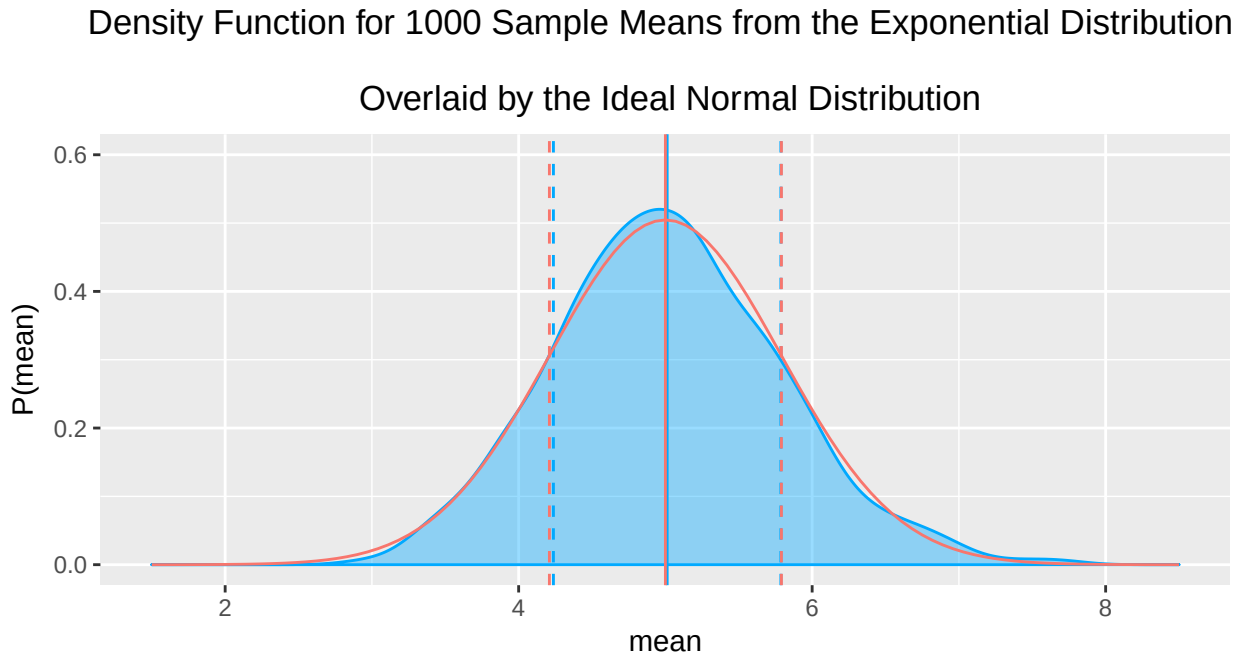

The sample variance estimate is about 24.84. This number is almost identical to the theoretical variance of 25. Since variance is the square of the standard deviation, any minor differences in standard deviation are magnified when computing variance. Thus, the small difference between 24.84 and 25 is acceptable.

**Distribution**

There are many ways to determine if a distribution is normal. In this section, two graphical methods are used to show the distribution of sample means is normally distributed.

In the first test for normality, the sample distribution is plotted with the ideal normally distributed sample population using what is known from the Central Limit Theorem. One can calculate the ideal normal distribution using mu and the equation for the standard error, SE = sigma/sqrt(n):

```
#Calculate the standard error, SE, as well as the standard error of the sample means, s using n, the
#size of each sample
n <- 40
SE <- sigma/sqrt(n)
s <- sd(sim_samples$x_bar_samp)
#Plot the sample distribution (blue) with the ideal normal distribution (pink) overlaid for comparison.
#Also, include the ideal and sample means and standard deviation as vertical lines
ggplot(sim_samples, aes(x = x_bar_samp)) +
    geom_density(geom = "area", fill = "#00A9FF", alpha = 0.4,color = "#00A9FF") +
    geom_vline(xintercept = x_bar, color = "#00A9FF") +
    geom_vline(xintercept = x_bar-s, color = "#00A9FF", lty = 2) +
    geom_vline(xintercept = x_bar+s, color = "#00A9FF", lty = 2) +
    stat_function(fun=dnorm, args=list(mean=mu, sd=SE), geom = "path", fill = "#F8766D",
                  color = "#F8766D") +
    geom_vline(xintercept = mu, color = "#F8766D") +
    geom_vline(xintercept = mu-SE, color = "#F8766D", lty = 2) +
    geom_vline(xintercept = mu+SE, color = "#F8766D", lty = 2) +
    ylim(0,0.6) +
    xlim (1.5,8.5) +
    labs(title = "Density Function for 1000 Sample Means from the Exponential Distribution
         \n Overlaid by the Ideal Normal Distribution",
         x = "mean", y = "P(mean)") +
    theme(plot.title = element_text(hjust = 0.5))
```

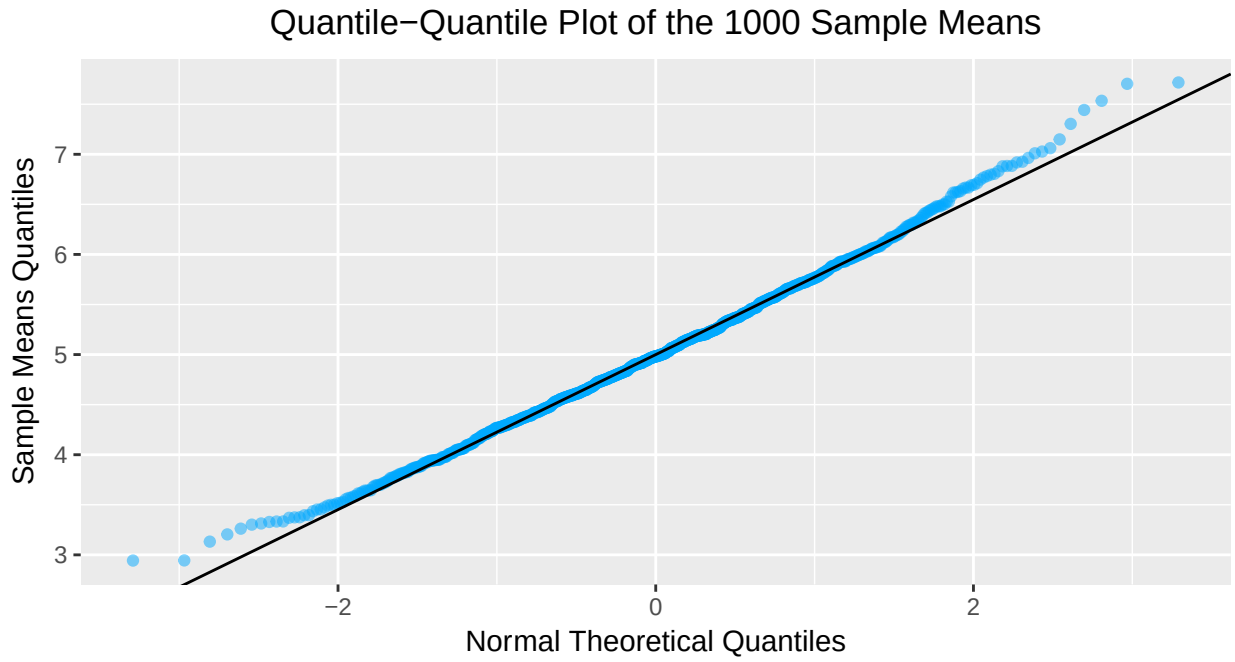



The plot shows that the blue sample distribution almost exactly matches the pink, ideal normal distribution,

giving strong evidence for the sample distribution's normality.

Further, a quantile-quantile plot is used to check the sample mean distribution for normality:

```
#Find the slope and intercept of the line that passes through the 1st and 3rd quartile of
#the normal q-q plot
y     <- quantile(sim_samples$x_bar_samp, c(0.25, 0.75)) #Find the 1st and 3rd quartiles
x     <- qnorm( c(0.25, 0.75))          #Find the matching normal values on the x-axis
slope <- diff(y) / diff(x)              #Compute the line slope
int   <- y[1] - slope * x[1]            #Compute the line intercept
#Plot the qunatile-quantile plot
ggplot(sim_samples, aes(sample = x_bar_samp)) +
    geom_qq(color = "#00A9FF", alpha = 0.5) +
    geom_abline(intercept=int, slope=slope) +
    labs(title = "Quantile-Quantile Plot of the 1000 Sample Means", x = "Normal Theoretical Quantiles",
         y = "Sample Means Quantiles") +
    theme(plot.title = element_text(hjust = 0.5))
```




The vast majority of points fall on the line, serving as additional strong evidence that the distribution is normal.