An Application of the Central Limit Theorem to the Exponential Distribution

Kevin E. D'Elia

March 5, 2016

Overview (Synopsis of Analysis)

The Central Limit Theorem (CLT), and the concept of the sampling distribution, are critical for understanding why statistical inference works. The CLT says that if you take many repeated samples from a population, then calculate the averages (or sum) of each one, the collection of those averages will be normally distributed. The purpose of this project is to examine the mean and variance of a sample from the exponential distribution, then produce a sampling distribution of the mean and variance, and finally demonstrate how the sampling distribution is approximately normal.

Simulations

As with any good simulation, the first step is to set the seed for the random number generators. This is done so that the results generated by the random number engine will be consistent across all invocations of the code, thus ensuring that the results are reproducible by other interested parties.

```r
set.seed(1234)
```

The R code for the simulations will be included in-line with the report for easy verification, along with explanations about what is being accomplished by each R markdown code chunk.

There will be two simulations run. The first will draw 40 values from the exponential distribution. The second will perform the first simulation, except that it will be executed 10,000 times and, for each execution, the mean and variance will calculated for the sample and stored in a vector. The idea behind this last point is you keep taking a sample of size **n** from the population, calculate some statistic on it, such as the mean, store that value and repeat the process, using a large number of repetitions and increasing sample sizes. This results in what is known as a **sampling distribution**. Note that this report will not alter the sample size.

The exponential distribution can be simulated in R with **rexp(n, lambda)** where λ is the *rate* parameter. The mean of the exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$. A requirement of this report is to set $\lambda = 0.2$ for all of the simulations. I chose 10,000 replications because this number is large enough to ensure that the mean of the sampling distribution approaches the mean of the population.

Sample Mean versus Theoretical Mean:

The first step in the analysis is to draw 40 random exponentials, like so:

```r
lambda <- 0.2
n <- 40
sample1 <- rexp(n, lambda)
```

For comparison, and as a step in demonstrating the approach, a second sample is drawn, using a new seed to ensure distinct randomness (otherwise, the previous seed will be used and the distribution will be identical):

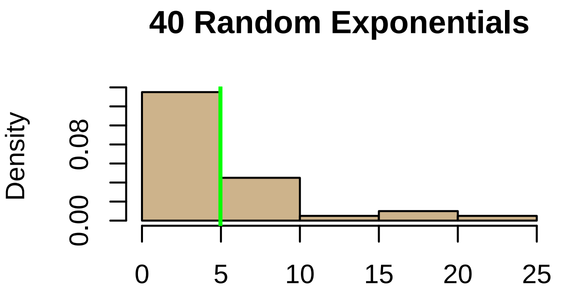
```
set.seed(9898)
sample2 <- rexp(n, lambda)
```

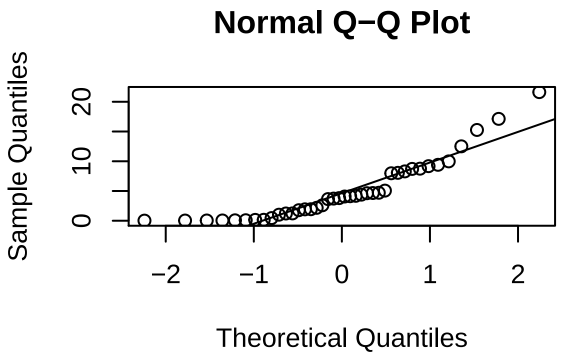
The mean, variance, and standard deviation is computed for each sample and this is what the values look like:

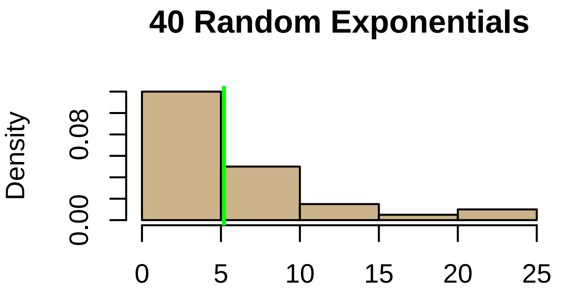
Sample	Mean	Var	StdDev
Sample 1	4.97	25.99	5.10
Sample 2	5.19	31.60	5.62

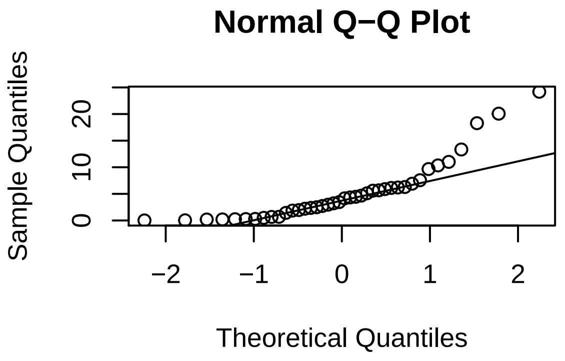
We can quickly see from the table that there is a difference in values between the mean of two samples. Graphically, this can be shown using a combination of histograms and Q-Q plots:

```
hist(sample1, prob = TRUE, col = "navajowhite3", border = "black", xlab = "", main = title)
abline(v = mean(sample1), col = "green", lwd = 2)
qqnorm(sample1);qqline(sample1)
hist(sample2, prob = TRUE, col = "navajowhite3", border = "black", xlab = "", main = title)
abline(v = mean(sample2), col = "green", lwd = 2)
qqnorm(sample2);qqline(sample2)
```









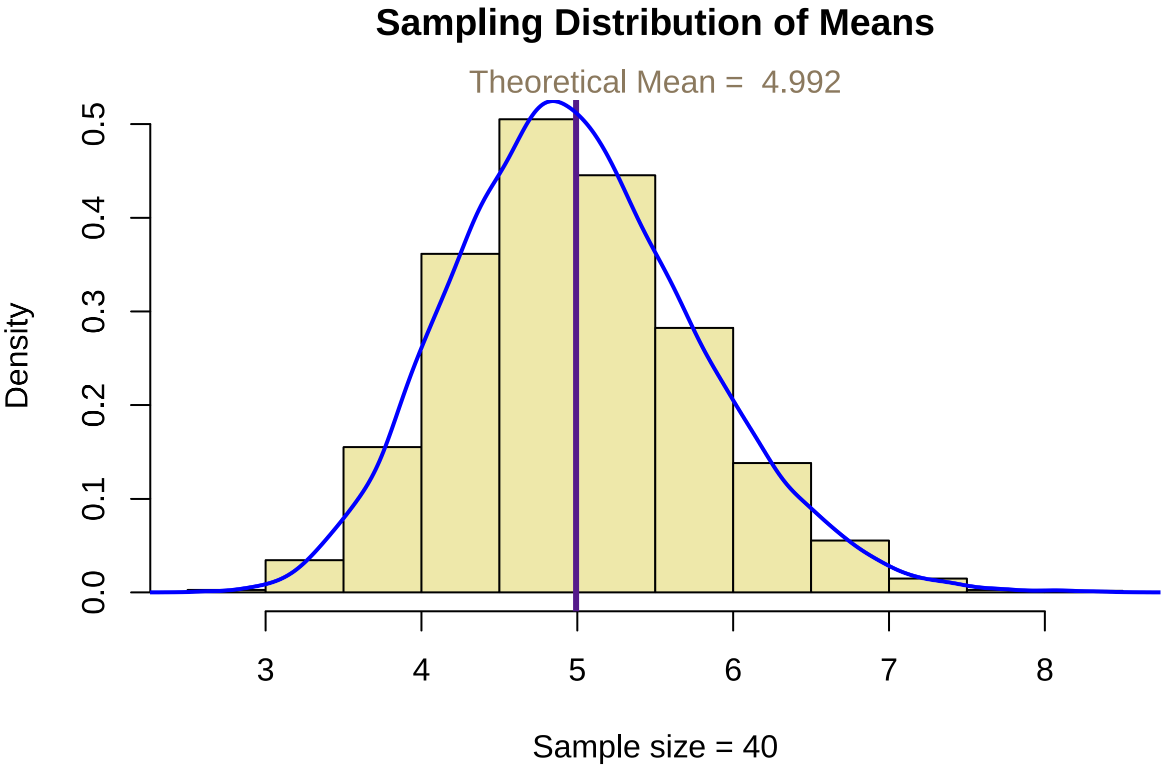
Several comments are in order regarding the figures:

1. The histograms are not normally distributed
2. The Q-Q plots indicate that neither of the samples form a normal distribution. If the values in either sample were normally distributed, the points on the plot will fall (more or less) along a straight line. As it happens, the data are strongly skewed away from the expected straight line.

3. The sample mean in both cases is centered around 5 and we will expect this to approximate the theoretical mean, or, in other words, we use the sample mean as an **estimator** of the **estimand**, that is, the population mean.

Now, what happens when this drawing of a sample of size **n** and subsequent calculation of a sample statistic is done repeatedly? The sample size remains the same, but the calculation is now done 10,000 times. The following graphic shows the distribution of the sample means, or what is also known as the **sampling distribution**.

```
xbar <- NULL
ul <- 10000
for (i in 1:ul) {xbar <- c(xbar, mean(rexp(n, lambda) ) ) }
hist(xbar, prob = TRUE, xlab = "Sample size = 40", main = main.title, col = "palegoldenrod")
abline(v = mean(xbar), col = "purple4", lwd = 3)
mtext(paste("Theoretical Mean = ", round(mean(xbar), 3)), col = "navajowhite4")
lines(density(xbar), col="blue", lwd=2)
```



From this graph, it can be seen that the distribution appears normal in shape. The balance point, or the *mean of the sample means*, appears to be reasonably close to the mean of the distribution from which the samples were drawn. So, it can be concluded that the sample mean is a good estimator of the population mean.

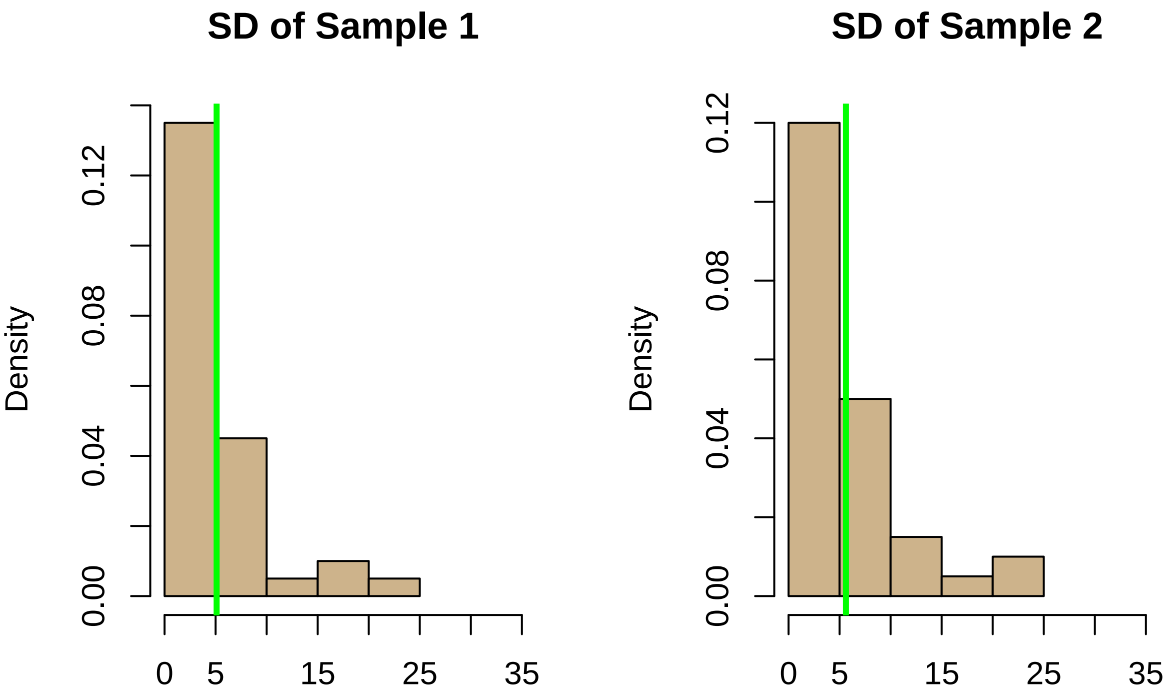
Sample Variance versus Theoretical Variance:

How well does the sample mean $\bar{x}$ estimate the hypothesized mean μ? To answer that, the sample variance and standard deviation for each sample is taken, and then the sampling distribution for the standard deviation is generated using 10000 experiments. For each of the two samples generated before, the variance and standard deviation are as follows:

Sample	Variance	StdDev
Sample 1	25.99	5.10
Sample 2	31.60	5.62

Looking at this data graphically gives the following result:

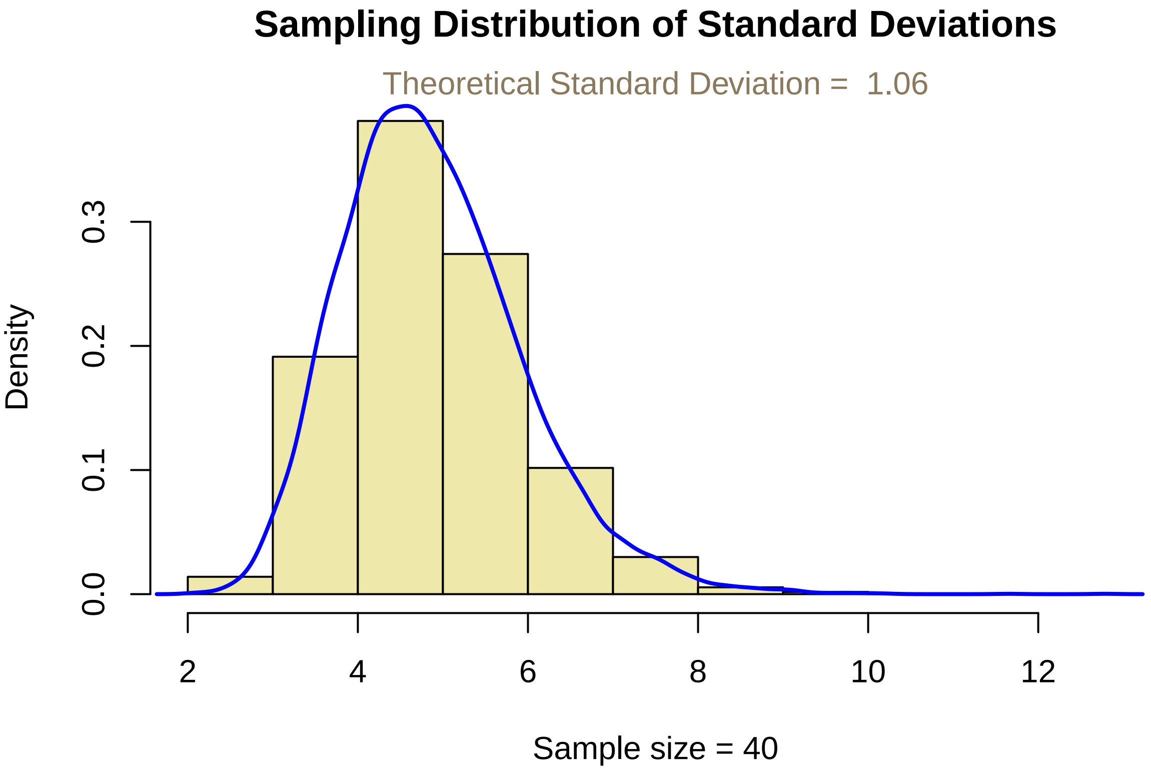
```
hist(sample1, prob = TRUE, col = "navajowhite3", xlim = c(0, 35), xlab = "", main = title1)
abline(v = sd(sample1), col = "green", lwd = 3)
hist(sample2, prob = TRUE, col = "navajowhite3", xlim = c(0, 35), xlab = "", main = title2)
abline(v = sd(sample2), col = "green", lwd = 3)
```



So most of the data lies within 1 standard deviation, or 5 units, from the mean, as expected.

The same technique used to generate a sampling distribution of the mean is used to generate a sampling distribution of the standard deviation. The resulting graph is shown below:

```
stddevs <- NULL
ul <- 10000
for (i in 1:ul) {stddevs <- c(stddevs, sd(rexp(n, lambda) ) ) }
hist(stddevs, prob = TRUE, xlab = x_label, main = main_title, col = "palegoldenrod")
mtext(paste("Theoretical Standard Deviation = ", round(sd(stddevs), 3)), col = "navajowhite4")
lines(density(stddevs), col="blue", lwd=2)
```



The resulting distribution appears relatively normal as predicted by the Central Limit Theorem. Furthermore, as the number of experiments increased, the standard deviation decreased accordingly, indicating that the variance among the sample means was not very wide. A lower standard deviation implies a more normal, bell-shaped distribution, along with a higher degree of confidence that the estimated mean is accurate with respect to the population mean.

A further check is to calculate the **standard error of the mean**, also called the **standard deviation of the mean**, which estimates the standard deviation of a sampling distribution. The SEM informs about the change in calculated means across a large number of experiments measuring the same quantity. It estimates the variability between samples whereas the standard deviation measures the variability within a single sample. Thus, if the effects of random changes are significant, then the standard error of the mean will be higher, and vice versa. The formula for calculating the SEM is:

$\sigma_M = \frac{\sigma}{\sqrt{N}}$ where

σ_M is the standard error of the mean

σ is the standard deviation of the original distribution

N is the sample size

$\sqrt{N}$ is the square root of the sample size

In this study, $\sigma = 5$, N = 40, and $\sigma_M = 0.791$. If the sample size were increased, it would approach the value of the standard error.

Distribution

By The Central Limit Theorem, for large n, $\bar{X} \sim$ N(μ, σ^2/n). As one of the most important concepts in statistics, the CLT states that, for any distribution with a finite mean and standard deviation, samples taken from that population will tend towards a normal distribution around the mean of the population as sample size increases. Furthermore, as sample size increases, the variance of the sample means will decrease.

The document answers the question “Does the distribution of means of 40 exponentials behave as predicted by the Central Limit theorem?” From the analysis shown above, it can be stated that this is the case.

In this analysis, the population is all numbers in an exponential distribution. The sample size taken is 40 and the number of samples taken (which hopefully approximates the population) is 10,000. The population mean is $\frac{1}{\lambda}$ as is the standard deviation. In the case of this study, that amounts to 1/0.2 or 5. It was shown that, for both the sample and the sampling distribution, the mean was very close to 5. In other words, the mean of the sampling distribution is roughly equivalent to the mean of the population. It was also shown that the standard deviation of the samples approximated the standard deviation of the population.