# Testing Central Limit Theorem on the Exponential Distribution

*Tamer Köksal*

*February 13, 2015*

## The Central Limit Theorem (CLT)

According to the Central Limit Theorem (CLT), no matter what the distribution of the underlying population (or the sample data) is, the distribution of sample means is approximately normal where the mean of sample means (the mean of the distribution) converges to the true population mean $\mu$. The variance of the distribution of sample means converges to $\frac{\sigma^2}{n}$, where $\sigma^2$ is the population variance. This statement might be confusing at first, so it needs some explanation.

When you do survey research, often what you get is a single sample out of the larger population and you try to estimate about the population based on the sample at hand. Therefore, the natural question is how typical is the sample that you get, that is how representative of the population is that sample. The answer to this question is nothing but the CLT. Although we are not in a position to draw as many samples as possible to investigate how typical the samples are, the CLT tells us that as $n$, the sample size, gets large, sample means tend to accumulate around the true population mean. That is, imagine that using simulations we draw many samples (for the purpose of this assignment 1000 samples), and we calculate their means and plot the histogram (the distribution) of these sample means. According to the CLT this distribution of sample means is approximately normal, 95% of these sample means fall within approximately 2 standard deviations above and below the population mean.

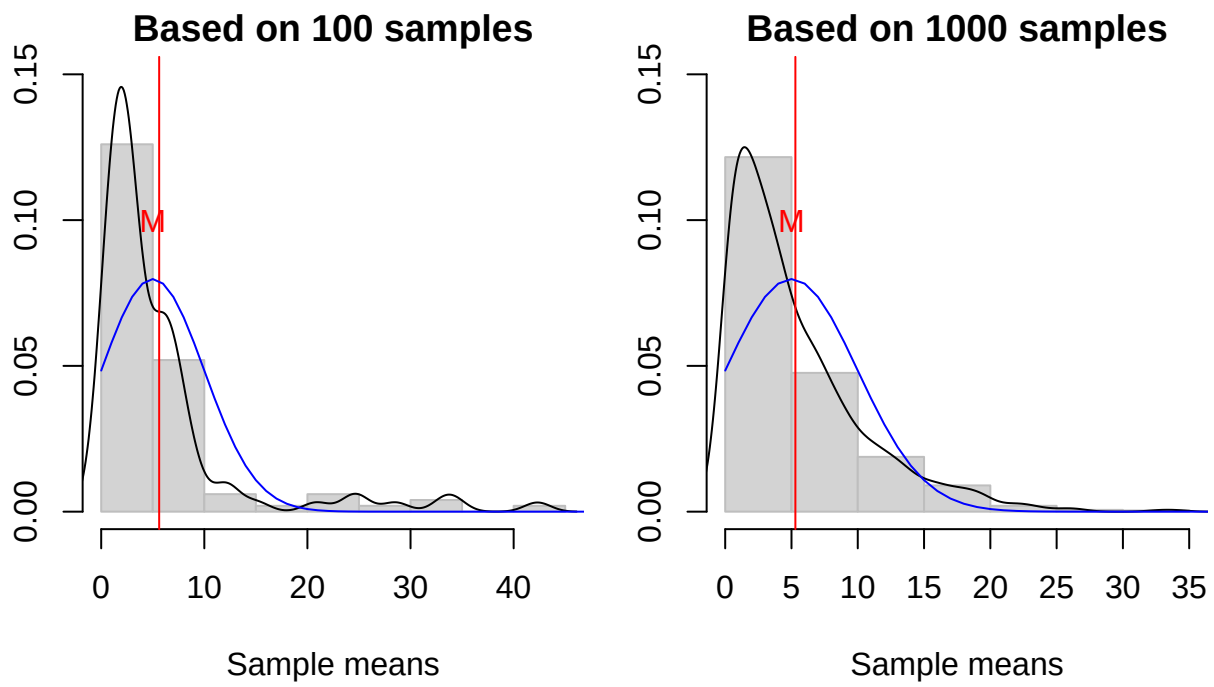
## Test the validity of the CLT on the exponential distribution

Since according to above introduction the validity of the CLT does not depend on the distribution of the sample or the respective population, in this part we are going to show using simulations that the CLT indeed works for exponentially distributed samples or populations as well. Exponential distributions are highly skewed distributions. To demonstrate the distribution of their shape we first perform single draws ($n = 1$) from the family of exponential distributions using the R function `rexp(n, rate)` and simulate these (repeat draws) 100 and 1000 times, respectively. The function `plot_sample_means()` is used to perform and visualize these simulations (Please refer to the Appendix for the code and explanation of this function). Two distributions are produced side by side, where the left one corresponds to the means of 100 samples, and the right one to the means of 1000 samples. Each plot consists of a histogram and the corresponding density curve (the black curve) of the distribution of samples means as well as the corresponding theoretical normal curve (the blue curve) for the purpose of comparison of the exponential distribution with the CLT.

Here are the parameters (arguments) that we provide the function with:

Explanation regarding the parameters can be found in the Appendix.

```
plot_sample_means(rexp, n=1, nosim100=100, nosim1000=1000, lambda = .2,
        ylim = .15, a=0, b=100, title="Sample means from exponential distribution")
```




As you can see from the plots the exponential distribution is highly skewed and not normally distributed both for 100 and 1000 simulations. However, note that although its distribution is away from being normal, the mean of sample means, $M$ (the red vertical line), is the same as the true population mean $\mu = \frac{1}{\lambda} = \frac{1}{.2} = 5$.

---

As for the purpose of this assignment let's answer the questions below based on the simulations plotted in the below figure, where $n = 40$ and $\lambda = .2$.

**1. Show the sample mean and compare it to the theoretical mean of the distribution.**

According to the CLT as also stated above the mean of sample means $M$ is supposed to be equal to the theoretical mean, that is $\mu = 5$. By calling the `plot_sample_means` function as below, we retrieve the means of sample means for both 100 and 1000, respectively.

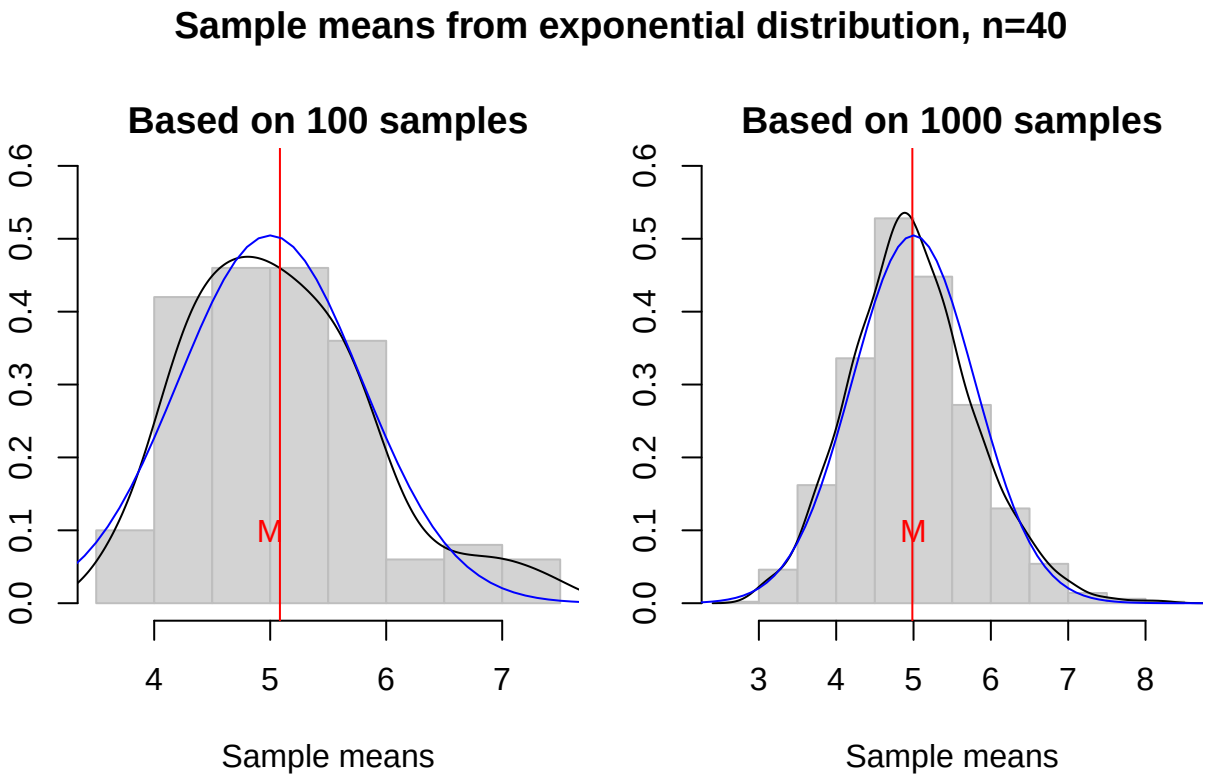
```
# Get the means and variances of the distribution of sample means for both 100 and 1000
# simulations and assign them to the object `simulationdata`
simulationdata <- plot_sample_means(rexp, n=40, nosim100=100, nosim1000=1000, lambda = .2,
       ylim = .6, a=0, b=10, title="Sample means from exponential distribution")
```

```
simulationdata[["means"]] # Print the means
```

```
## [1] 5.083814 4.984977
```

As you can see, both of the means of sample means are very close to 5, the theoretical mean. Also notice that, the simulation with 1000 sample draws has a more precise estimate (4.985) for the theoretical mean. (See the red vertical lines in the below figure)

```
plot_sample_means(rexp, n=40, nosim100=100, nosim1000=1000, lambda = .2,
         ylim = .6, a=0, b=10, title="Sample means from exponential distribution")
```




**2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.**

By visually examining both of the figures above, one can see that as $n$, the sample size, gets large and as the number of samples drawn increases (in our case from 100 to 1000) the distribution of the sample means gets more normal. To show it mathematically, we need to show that the variance of the distribution of sample means is approximately equal to the variance of the theoretical normal distribution $N(\mu, \sigma^2)$, where the variance $\sigma^2 = \frac{1}{\lambda^2 n}$, since the theoretical variance of the exponential distribution is $\frac{1}{\lambda^2}$. Thus, the theoretical variance is

$$\sigma^2 = \frac{1}{\lambda^2 n} = \frac{1}{(.2)^2 40} = .625$$

When we compute (by calling the `simulationdata` object created previously) the variances of the sample means, we see that they are as given below and they are approximately equal to the theoretical variance. As the number of samples drawn and sample size increase, so does the precision of the variance of the distribution of sample means.

```r
simulationdata[["variances"]] # Print the variances
```

```
## [1] 0.6391735 0.6367179
```

**3. Show that the distribution is approximately normal.**

It is clearly seen from the plots that the distributions are approximately normal. As the number of samples drawn and sample size increase the distribution gets more normal.

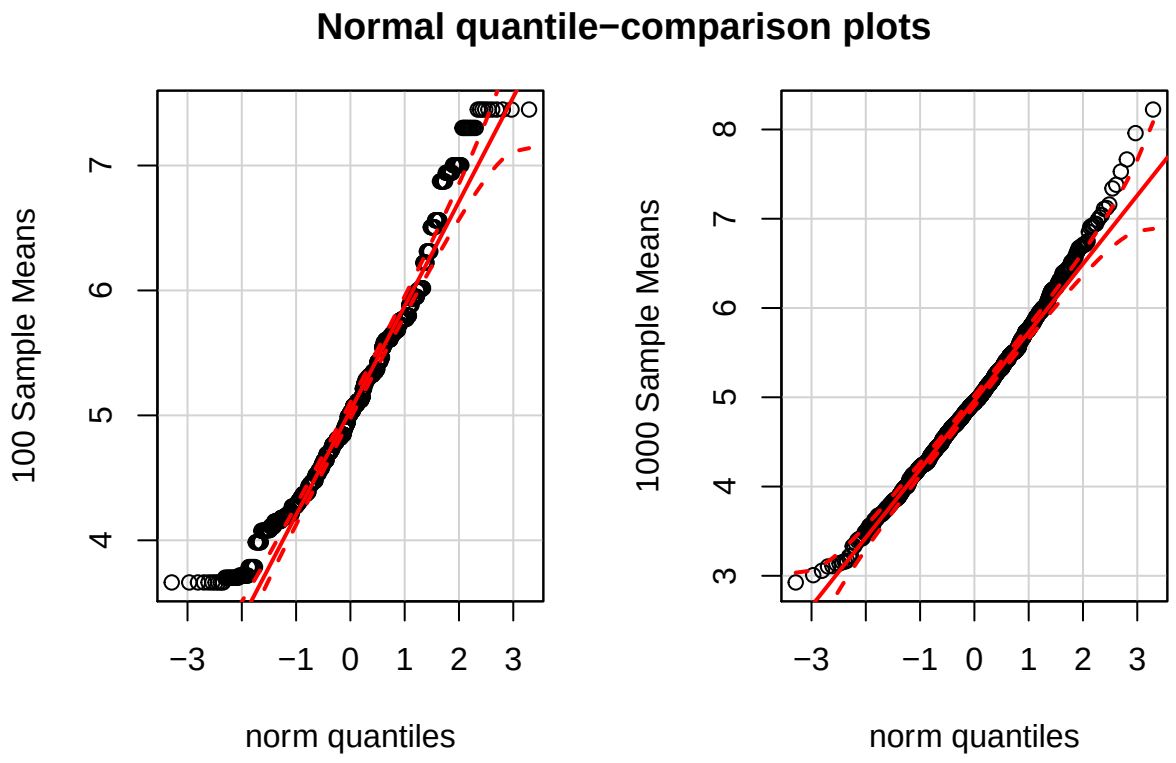
Another way to check for the normality of a distribution is to draw a quantile-comparison plot of the distribution against the theoretical normal distribution. If the points of the corresponding scatter plot fall along a straight line, then the distribution is said to be normal.

```r
library(car)
# set up a two panel plot
par(mfrow=c(1,2))

# qqplot (quantile-quantile plot) of the data of "100 Sample Means" against the normal quantiles
qqPlot(simulationdata[["df"]][,1], ylab = "100 Sample Means")

# qqplot of the data of "1000 Sample Means" against the normal quantiles
qqPlot(simulationdata[["df"]][,2], ylab = "1000 Sample Means")

title("\n\n\n\nNormal quantile-comparison plots", outer = TRUE)
```



As you can see from the quantile-comparison plots, the data with 1000 simulations more neatly follows a straight line. And thus, as the number of simulations increase the distribution of sample means gets more normal.