Statistical Inference Project - Part 1

Bhaskar S

14th May 2016

Investigate the Exponential Distribution in R and Compare with Central Limit Theorem

Synopsis

In this project, we investigate the exponential distribution in R and compare it with the **Central Limit Theorem**. The exponential distribution can be simulated in R using the **rexp(n, lambda)** function, where n is the size and lambda is the rate parameter. The mean of exponential distribution is **1/lambda** and the standard deviation is also **1/lambda**.

Simulation Process

We initialize **lambda** with the value `0.2` for all of the simulations. The sample size **n** of the exponentials to investigate is `40`. The number of simulations will be `1000`.

For this analysis, we will be using the R package `ggplot2` for plotting graphs. The following code segment loads the desired libraries:

```
library(ggplot2)
```

We initialize few variables that will be referenced in this investigation. The following code segment performs the desired initializations:

```
n <- 40
lambda <- 0.2
nosim <- 1000
```

The following code segment initializes the random number generator state for reproducible results:

```
set.seed(1000)
```

The following code generates a sample exponential distribution using the R function `rexp` using the variables **n** and **lambda** and stores the result in the variable **sample**:

```
sample <- rexp(n, lambda)
```

The following code finds the sample `mean` and the sample `standard deviation` (square root of sample `variance`) for the generated **sample** and stores the results in the variables **sample.mean** and **sample.s** respectively:

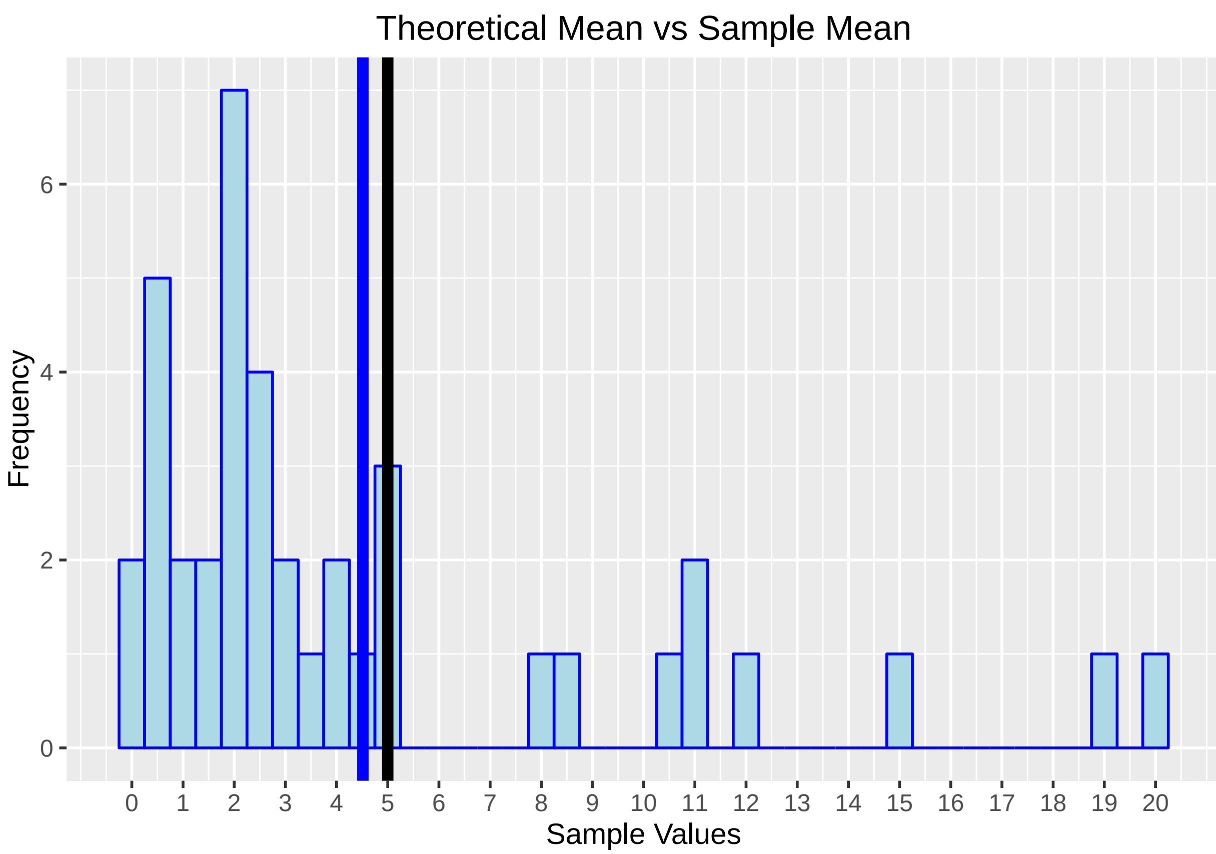
```
sample.mean <- mean(sample)
sample.s <- sd(sample)
```

The theoretical `mean` and the theoretical `standard deviation` (square root of theoretical `variance`) of an exponential distribution are the *same* and is **1 / lambda**. The following code finds the theoretical `mean` and the theoretical `standard deviation` (square root of theoretical `variance`) and stores the results in the variables **theoretical.mean** and **theoretical.sigma** respectively:

```
theoretical.mean <- 1 / lambda
theoretical.sigma <- 1 / lambda
```

The following code plots the **sample** exponential distribution as a *histogram* and displays the theoretical `mean` (black vertical line) and the sample `mean` (blue vertical line):

```
ggplot(data.frame(data = sample), aes(x = data)) +
    labs(title = 'Theoretical Mean vs Sample Mean') +
    labs(x = 'Sample Values', y = 'Frequency') +
    geom_histogram(colour = 'blue', fill = 'light blue', binwidth = 0.5) +
    scale_x_continuous(breaks=seq(0, 20, 1)) +
    geom_vline(xintercept = theoretical.mean, color = 'black', size = 2) +
    geom_vline(xintercept = sample.mean, color = 'blue', size = 2)
```



From the above plot, we see that the sample `mean` is close enough to the theoretical population `mean` but not an accurate estimator.

The same is true for the `standard deviation` (square root of `variance`). The following code displays the sample and theoretical `standard deviation` (a measure of `variance`):

```
cat('Sample Standard Deviation (s):', sample.s)
```

```
## Sample Standard Deviation (s): 5.035085
```

```
cat('Theoretical (population) Standard Deviation (sigma):', theoretical.sigma)
```

```
## Theoretical (population) Standard Deviation (sigma): 5
```

As we take more and more samples of the exponential distribution using simulation (creating what we call a **Sampling Distribution**), the `mean` as well as the variance (`standard deviation`) of the **Sampling Distribution** more accurately estimate the theoretical population `mean` and variance (`standard deviation`). Also, the **Sampling Distribution** looks more like a **Normal Distribution**. This is the essense of the **Central Limit Theorem**.

We will now perform the `1000` simulations to generate the **sampling distribution**.

The following code performs `1000` simulations to generate different exponential distribution samples and stores the result as a `1000 x 40` matrix in the variable **simulations**:

```
simulations <- matrix(rexp(n * nosim, lambda), nosim)
```

The following code uses the R `apply` function on each of the `1000` exponential distribution samples (each row of the matrix) and computes the sample `mean` for each matrix row and stores the result in the variable **simulation.means**:

```
simulation.means <- apply(simulations, 1, mean) ### 1 here indicates ROW
```

The `standard deviation` (sigma) of the population and the `standard deviation` (s) of the sampling distribution for a sample size (n) are related as follows:

```
s = sigma / sqrt(n)
```

From the above equation, to find the estimated population `standard deviation` (sigma) from the estimated sample `mean` (s), we use the following equation:

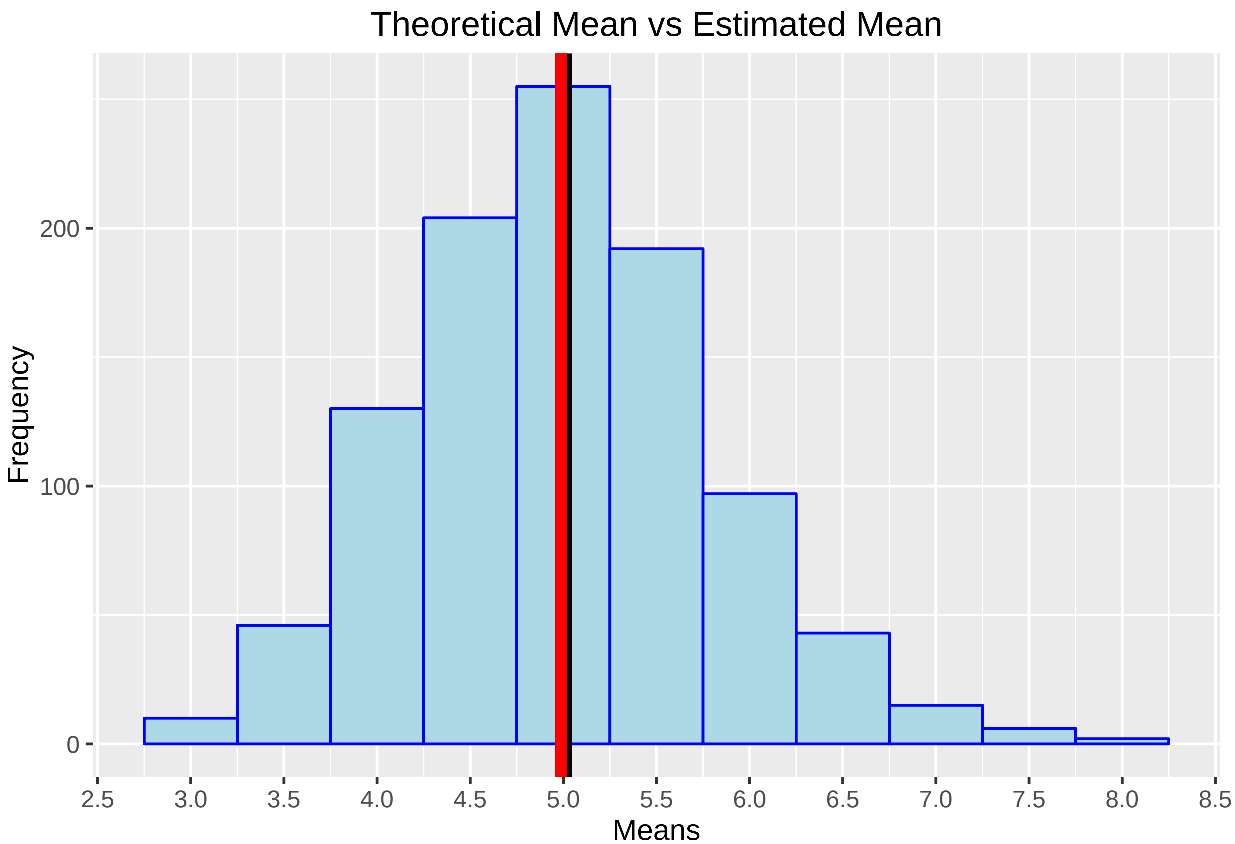
```
sigma = s * sqrt(n)
```

The following code computes the estimated sampling `mean` and the estimated sampling `standard deviation` (square root of sample `variance`) from the sampling distribution **simulations** and stores the results in the variables **estimated.mean** and **estimated.sigma** respectively:

```
estimated.mean <- mean(simulation.means)
estimated.sigma <- sd(simulation.means) * sqrt(n)
```

The following code plots the means of the sampling distribution **simulation.means** as a *histogram* and displays the theoretical `mean` (black vertical line) and the estimated `mean` (red vertical line):

```
ggplot(data.frame(means = simulation.means), aes(x = means)) +
    labs(title = 'Theoretical Mean vs Estimated Mean') +
    labs(x = 'Means', y = 'Frequency') +
    geom_histogram(colour = 'blue', fill = 'light blue', binwidth = 0.5) +
    scale_x_continuous(breaks=seq(0, 10, 0.5)) +
    geom_vline(xintercept = theoretical.mean, color = 'black', size = 3) +
    geom_vline(xintercept = estimated.mean, color = 'red', size = 2)
```



From the above plot, we see that the estimated `mean` accurately estimates the theoretical population `mean`, validating the essence of the **Central Limit Theorem**.

The same is true for the `standard deviation` (square root of `variance`). The following code displays the estimated and theoretical `standard deviation` (a measure of `variance`):

```
cat('Estimated Standard Deviation (sigma):', estimated.sigma)
```

```
## Estimated Standard Deviation (sigma): 5.135701
```

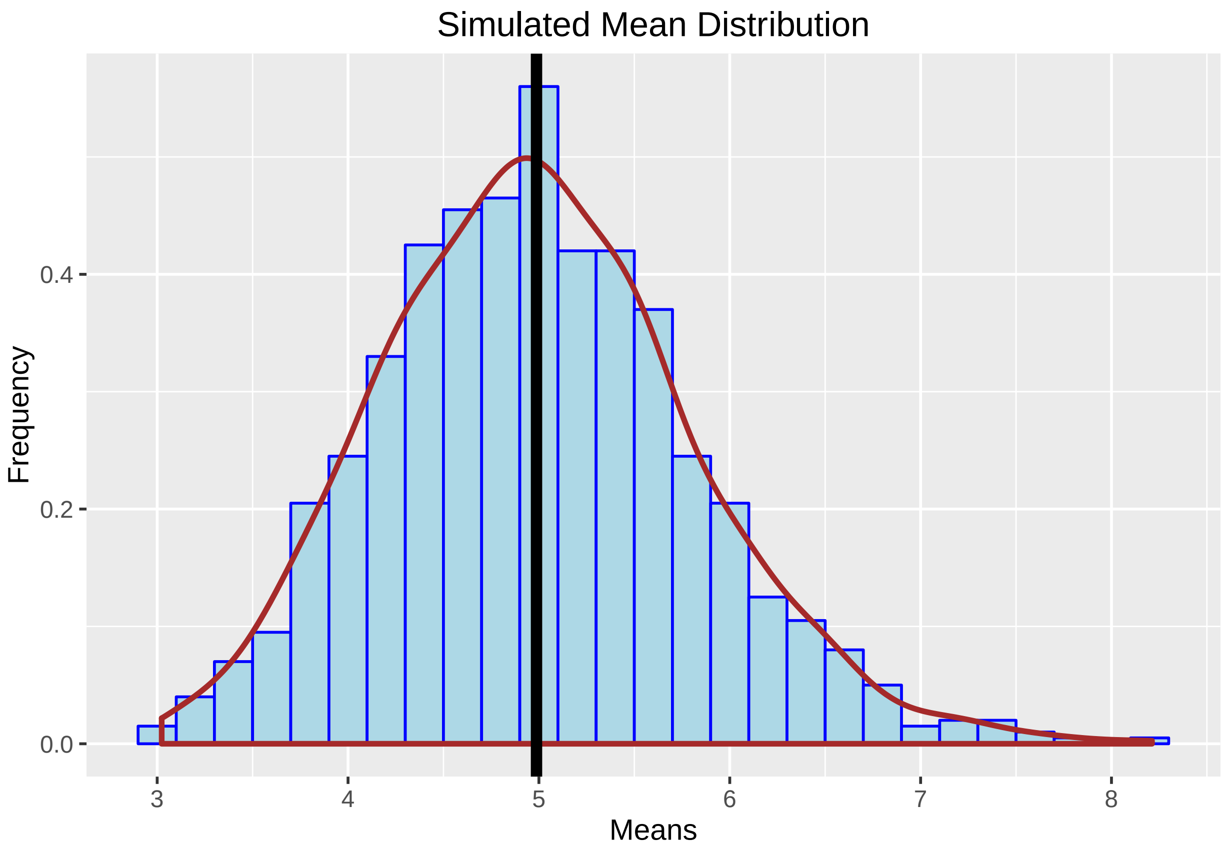
```
cat('Theoretical (population) Standard Deviation (sigma):', theoretical.sigma)
```

```
## Theoretical (population) Standard Deviation (sigma): 5
```

Now, let us plot the **Sampling Distribution** to show that it looks more like a **Normal Distribution**.

The following code plots the means of the sampling distribution **simulation.means** as a *histogram* and overlays a density curve:

```
ggplot(data.frame(means = simulation.means), aes(x = means)) +
    labs(title = 'Simulated Mean Distribution') +
    labs(x = 'Means', y = 'Frequency') +
    geom_histogram(aes(y = ..density..), colour = 'blue',
                   fill = 'light blue', binwidth = 0.2) +
    geom_density(color = 'brown', size = 1) +
    geom_vline(xintercept = estimated.mean, color = 'black', size = 2)
```



Summary

According to the **Central Limit Theorem**, a **Sampling Distribution** that is created through a large number of simulations of exponential distribution samples, has the following properties:

- The `mean` of the **Sampling Distribution** accurately estimates the population `mean`
- The `variance` of the **Sampling Distribution** accurately estimates the population `variance`
- Plotting a **Sampling Distribution** follows a **Normal Distribution** curve