Validation of the CLT Using the Exponential Distribution

Feroz Mohamed Hatha

22/09/2020

```
knitr::opts_chunk$set(echo = TRUE)
```

Overview

In this part of the project, we will look at the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter. We know that the mean as well as the standard deviation of the exponential distribution is `1/lambda`. We set `lambda = 0.2` for all of the simulations. We investigate the average of 40 exponentials and form a distribution of the averages by drawing 40 exponentials a total of 1000 times.

Simulations

We draw a sample of 40 exponentials (from an exponential distribution with `lambda = 0.2`) a total of 1000 times and store it in a matrix called `exp_data`. Each row corresponds to 40 samples from the exponential distribution. Our matrix thus has 1000 rows and 40 columns. The code for generating the matrix is as follows:

```
set.seed(1)
exp_data <- matrix(rexp(40000, rate = 0.2), 1000, 40)
```

We calculate the means of each sample of size 40 using the function `apply()` and store it in a 1000 long vector called `means` as follows:

```
means <- apply(exp_data, 1, mean)
```

The sample mean of the distribution of averages is calculated using the function `mean()` as follows:

```
sm_mean <- mean(means); sm_mean
```

```
## [1] 4.990025
```

The sample variance of the distribution of averages is calculated using the function `var()` as follows:

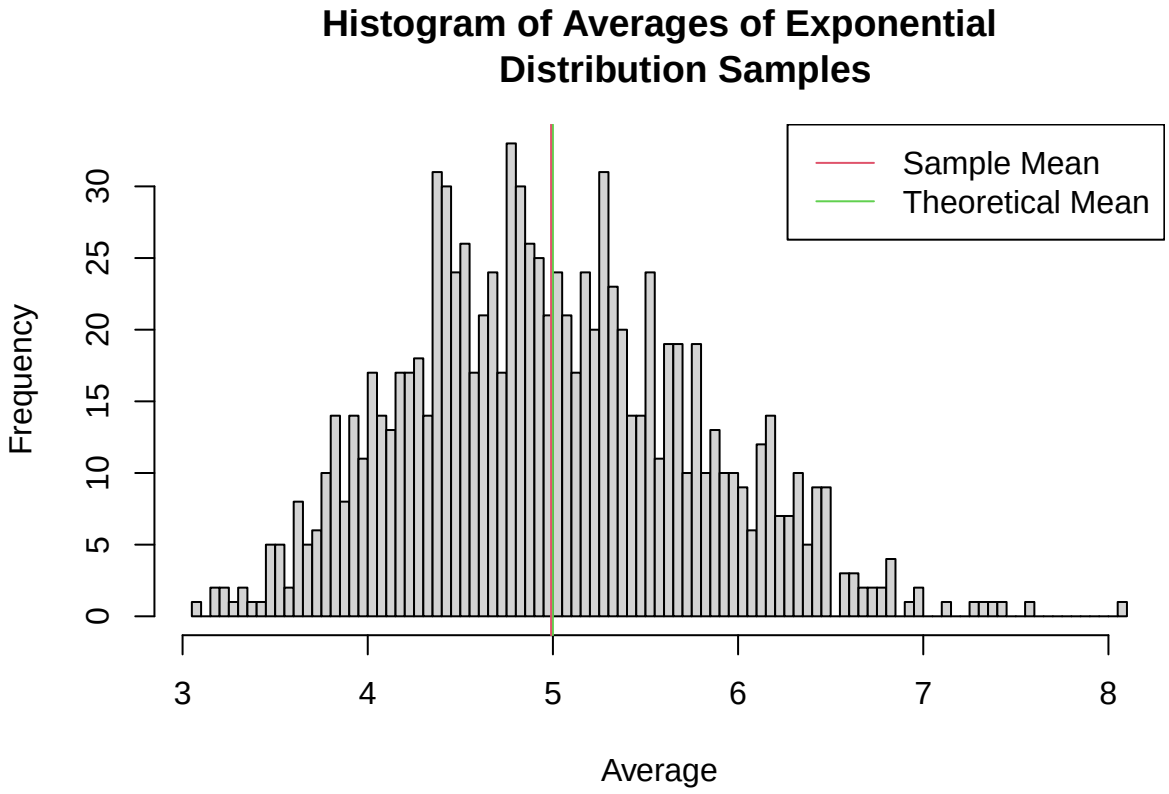
```
sm_var <- var(means); sm_var
```

```
## [1] 0.6177072
```

Sample Mean versus Theoretical Mean

We saw in the previous section the value of the sample mean which was contained in `sm_mean`. Also, we know that for an exponential distribution with the rate parameter `lambda = 0.2`, the mean is given by `1/lambda = 5`. We therefore see that sample mean and the theoretical mean are approximately equal. A histogram of the averages of exponentials can be created using the function `hist()`. The code for the same and the histogram produced as the output are as follows:

```
hist(means, breaks = 100, main = "Histogram of Averages of Exponential
     Distribution Samples", xlab = "Average", ylab = "Frequency")
abline(v = 5, col = 3, lwd = 1)
abline(v = sm_mean, col = 2, lwd = 1)
legend("topright", c("Sample Mean", "Theoretical Mean"), bty = "o",
       lty = c(1,1), lwd = c(1, 1), col = c(col = 2, col = 3))
```



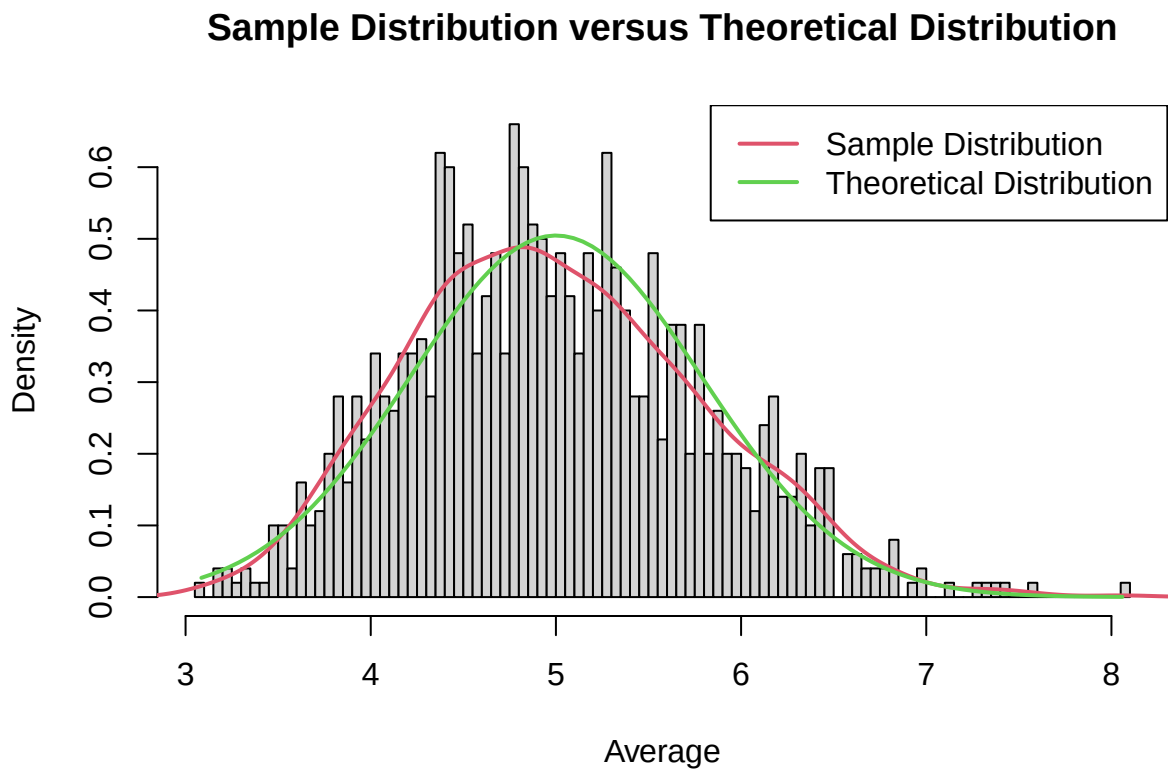
Sample Variance versus Theoretical Variance

We saw in one of the earlier sections the value of the sample variance which was contained in `sm_var`. Also, we know that for an exponential distribution with the rate parameter `lambda = 0.2`, the standard deviation `sigma` is given by `1/lambda = 5`. Now, we know that the theoretical variance of the distribution of averages of samples from a population with standard deviation `sigma` is given by `(sigma^2)/n`, where `n` is the sample size. Therefore, for this particular exponential distribution, the variance of averages of samples (of size 40) drawn from it is given by `25/40 = 0.625`. We thus see that this value of the theoretical variance is very close to the sample variance contained in `sm_var`.

Sample Distribution versus Theoretical Distribution

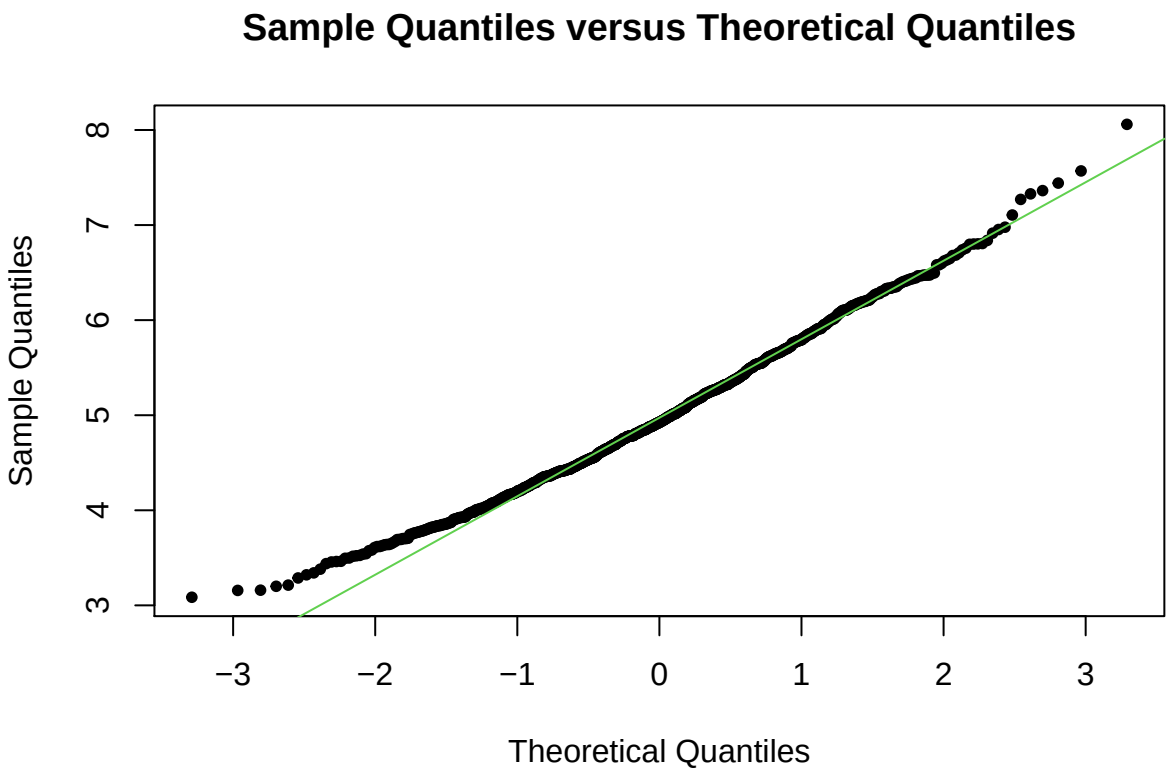
A probability distribution of the averages of exponential samples can be created using the functions `hist()`, `lines()`, and `density()`. In addition to this, we plot the theoretical normal distribution of the averages (as dictated by the CLT) with mean = `5` and variance = `0.625`. The code for generating the distributions and the output generated are as follows:

```
hist(means, breaks = 100, main = "Sample Distribution versus Theoretical Distribution",
     xlab = "Average", ylab = "Density", prob = TRUE)
lines(density(means), lty = 1, col = 2, lwd = 2)
x_vals <- seq(min(means), max(means), length = 100)
y_vals <- dnorm(x_vals, mean = 5, sd = 5/sqrt(40))
lines(x_vals, y_vals, col = 3, lty = 1, lwd = 2)
legend("topright", c("Sample Distribution", "Theoretical Distribution"),
       bty = "o", lty = c(1, 1), col = c(2, 3), lwd = c(2, 2))
```



We see that the sample distribution very closely resembles the theoretical normal distribution dictated by the CLT. Additionally, a Q-Q plot tells us how close the sample quantiles are to the theoretical quantiles. The code for generating the Q-Q plot and the output plot are as follows:

```
qqnorm(means, main = "Sample Quantiles versus Theoretical Quantiles", pch = 20)
qqline(means, col = 3)
```



Once again, we see that the sample quantiles are very close to the theoretical normal quantiles expected from the CLT. Thus, we can conclude that the distribution of means of exponential samples are approximately normal.