Statistical Inference course project part 1: Simulation Exercise

Mahmoud Nabegh

December 12, 2018

Contents

Overview

This the first part of the John Hopkins Coursera Statistical Inference course project. In this part, We investigate the exponential distribution in R and compare it with the Central Limit Theorem. We simulate data generated from an exponential distribution with lambda = 0.2, by generating 40 exponentials and calculating there mean, we perform the simulation a 1000 time and study the distribution of the means, which according to the Central Limit Theorem should be normal and centered around the theoretical mean.

Initial simulation setting

Here we define the parameters used in the simulation

```
lambda <- 0.2
sample_size <- 40
no_of_simulations <- 1000
set.seed(712)
```

Simulations

Here carry out the simulations, we run a 1000 simulation, where we generate 40 random numbers from an exponential distribution and then calculate its mean

```
exp_simulations <- t(replicate(no_of_simulations, rexp(40, rate = lambda)))
samples_exp_means <- apply(exp_simulations, 1, mean)
```

Theoretical distribution

Here we calculate the values for the theoretical distribution of the mean according to the Central Limit Theorem.

```
exponential_mean <- 1/lambda
exponential_standard_deviation <- 1/lambda
exponential_variance <- exponential_standard_deviation ^2

theo_dist_mean <- exponential_mean
theo_dist_sd <- exponential_standard_deviation / sqrt(sample_size)
theo_dist_var <- exponential_variance / sample_size
```

Sample Mean versus Theoretical Mean

```
sample_mean <- mean(samples_exp_means)
```

The sample mean is 4.9764409, while the theoretical mean as calculated is 5. We can see that the difference between them is very small.

Sample Variance versus Theoretical Variance

```
sample_var <- var(samples_exp_means)
sample_sd <- sd(samples_exp_means)
```

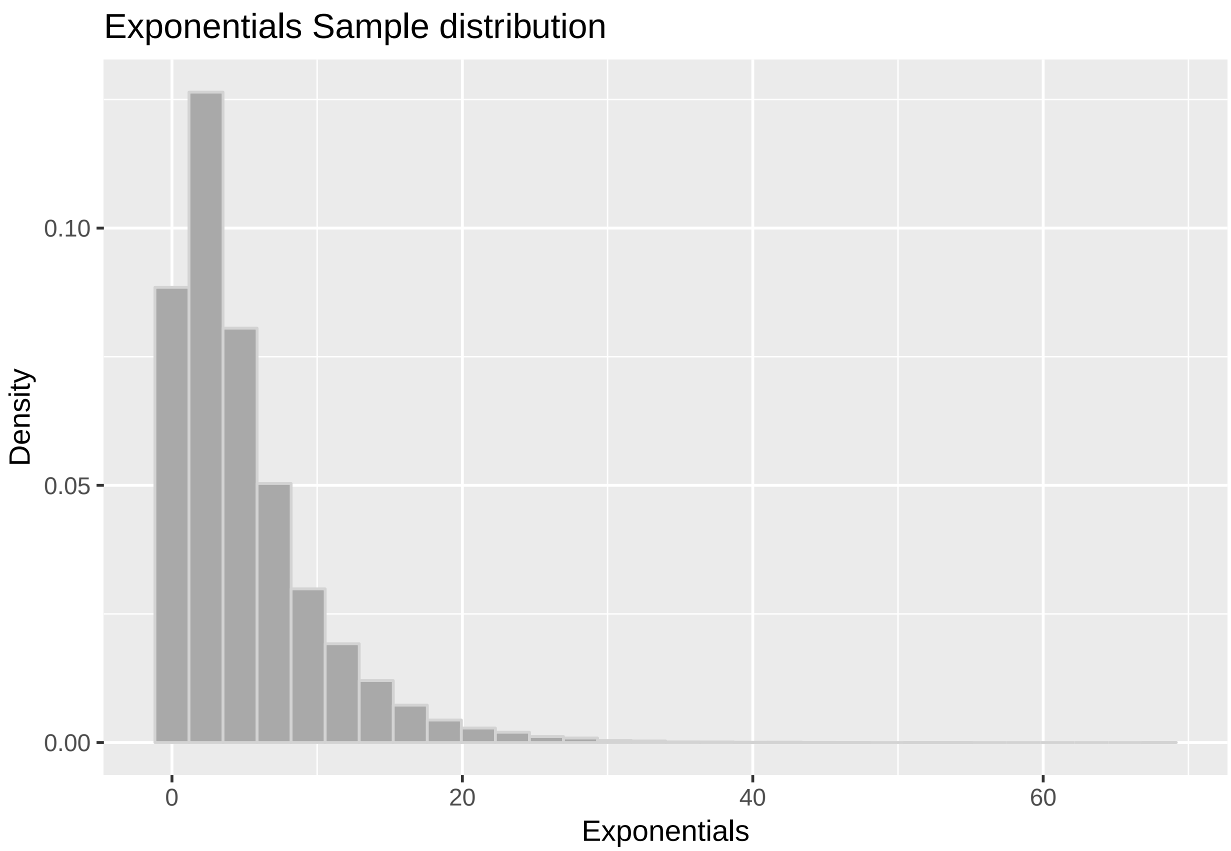
The sample variance is 0.6446548, while the theoretical variance as calculated is 0.625. Also, the sample standard deviation is 0.802904, while the theoretical standard deviation is 0.7905694. We can see that the differences between sample values and theoretical ones are very small.

Distributions

We will show the distrubtion of exponentials and the distribution of the average of the exponentials.

Distribution of large collection of random exponentials

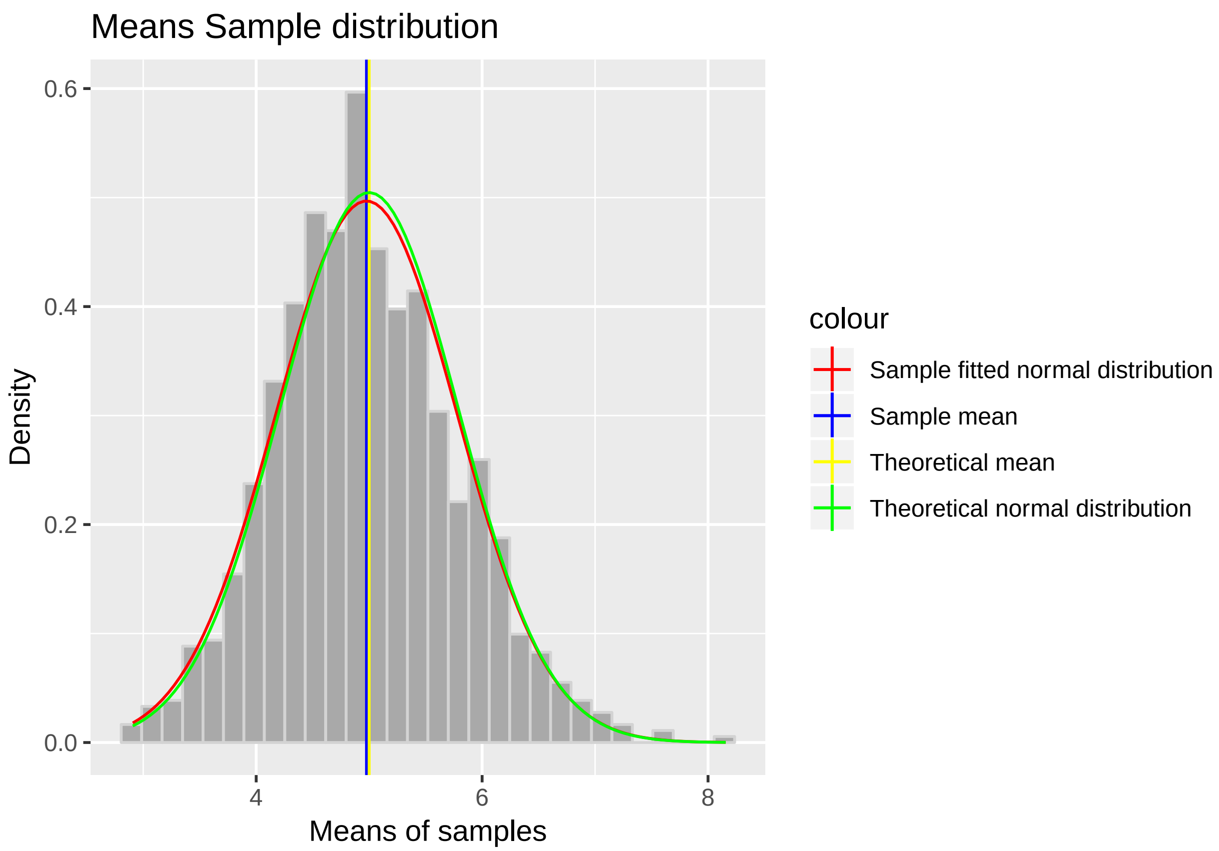
```
random_exp <- data.frame(exps = as.vector(exp_simulations))
ggplot(random_exp, aes(exps)) +
    geom_histogram(aes(y= ..density..), color = "lightgrey", fill = "darkgrey", bins = 30) +
    labs(title = "Exponentials Sample distribution", x = "Exponentials", y = "Density")
```



We can see that the distribution of large number of exponentials is exponential.

Distribution of the average of exponentials

```
samples_exp_means <- data.frame(means = samples_exp_means)
p <- ggplot(samples_exp_means, aes(means))
p + geom_histogram(aes(y = ..density..), color = "lightgrey", fill = "darkgrey", bins = 30) +
    geom_vline(aes(xintercept = mean(samples_exp_means$means), color = "Sample mean")) +
    geom_vline(aes(xintercept = theo_dist_mean, color = "Theoretical mean")) +
    stat_function(fun = dnorm, args = list(mean = sample_mean, sd = sample_sd)
                  , aes(color = "Sample fitted normal distribution")) +
    stat_function(fun = dnorm, args = list(mean = theo_dist_mean, sd = theo_dist_sd),
                  aes(colour = "Theoretical normal distribution")) +
    labs(title = "Means Sample distribution", x = "Means of samples", y = "Density") +
    scale_color_manual(values = c("red", "blue", "Yellow", "green"))
```



In this plot, we need to compare the random data to the normal distribution that it theoretically follows as stated by the central theorem. We begin by explaining the plot, the histogram represents the distribution of the sample. The two smooth red and green curves representing the two normal distributions constructed using the actual mean and standard deviaiton of the sample, and the theoretical mean and standard deviation as cacluated using the lambda value, while the vertical lines represent the means. From the plot, we can see that the histogram fits nicely distribution to the curves of expected distribution with the most frequent bin being the central one representing the mean, with only few exceptions of bins slightly above or below the curve, this is to be expected as to abide with the law of large numbers to get a perfect normal distribution we need to run infinite number of simulations.