Statistical Inference Course Project - Part 1

Juan Agustín Morello

20/8/2020

Overview

This is the first part (simulation exercise) of the Statistical Inference Course Project from Coursera.

In this project I'll look at the exponential distribution in R. I'll investigate the distribution of averages of 40 exponentials setting lambda = 0.2 for all the simulations and doing a thousand simulations. At the end, I'll compare it with the Central Limit Theorem.

The first thing I'll do is setting the seed to make the project reproducible.

```
set.seed(1111)
```

Simulations

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter.

The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. I´ll be using lambda = 0.2 for all the thousand simulations.

```
lambda <- .2
n_exponentials <- 40
simulations <- 1000

sample_means = rep(NA, simulations)
for (i in 1 : simulations) {
  sample_means[i] = mean(rexp(n_exponentials, lambda))
}
```

Sample Mean versus Theoretical Mean

Both mean values are very close to each other

```
print(paste("The theoretical mean of the distribution is 1/lambda:", 1/lambda))
```

```
## [1] "The theoretical mean of the distribution is 1/lambda: 5"
```

```r
print(paste("The mean of the simulated sample is:", round(mean(sample_means),2)))
```

```
## [1] "The mean of the simulated sample is: 4.99"
```

Sample Variance versus Theoretical Variance

Both variance values are very close to each other.

```r
print(paste("The theoretical variance of the distribution is [(1/lambda)^2]/n_exponentials:",
            ((1/lambda)^2)/n_exponentials))
```

```
## [1] "The theoretical variance of the distribution is [(1/lambda)^2]/n_exponentials: 0.625"
```

```r
print(paste("The variance of the simulated sample is:", round(var(sample_means),3)))
```

```
## [1] "The variance of the simulated sample is: 0.611"
```

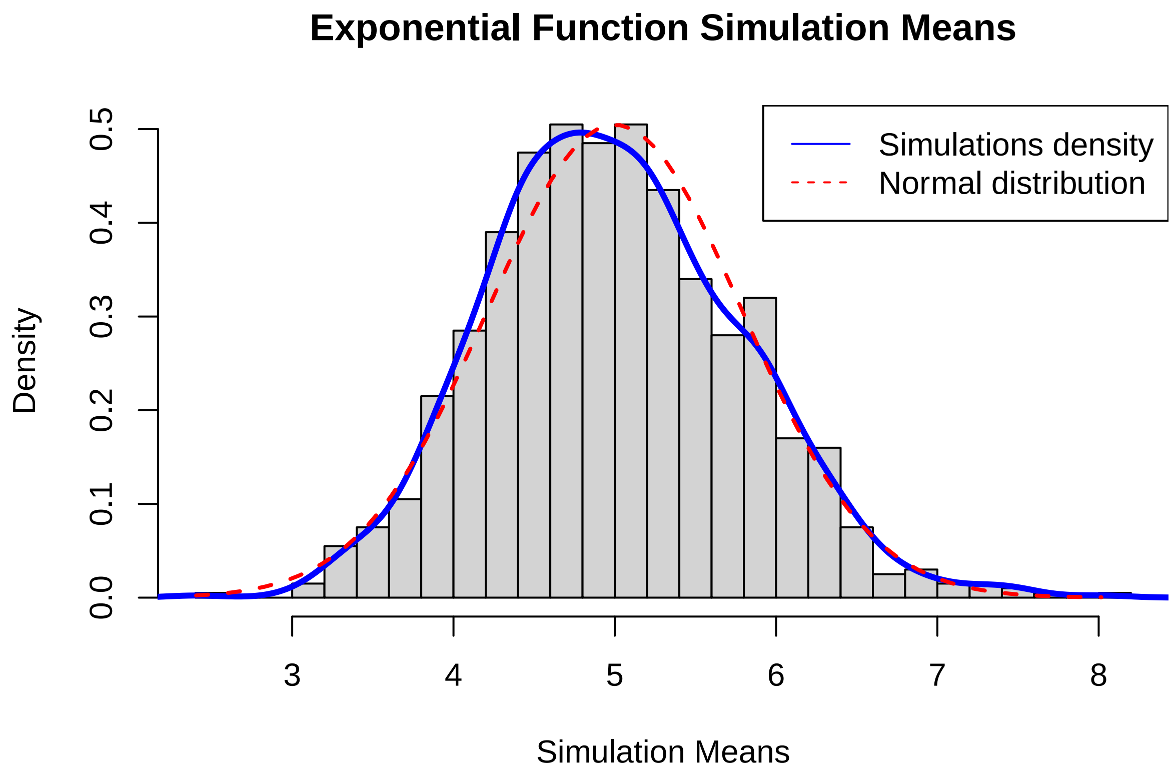
Distribution

Finally, we'll investigate whether the exponential distribution is approximately normal. The Central Limit Theorem stats that the distribution of averages of IID variables becomes that of a standard normal as the sample size increases. So the means of the sample simulations should resemble a normal distribution.

```r
# Plotting the sample means distribution
hist(sample_means, prob=TRUE, main="Exponential Function Simulation Means",
     breaks=n_exponentials, xlab = "Simulation Means")
lines(density(sample_means), lwd=3, col="blue")

# Plotting Normal distribution
x <- seq(min(sample_means), max(sample_means), length=2*n_exponentials)
y <- dnorm(x, mean=1/lambda, sd=((1/lambda))/n_exponentials^.5)
lines(x, y, pch=22, col="red", lwd=2, lty = 2)

legend("topright", legend=c("Simulations density", "Normal distribution"),
       col=c("blue", "red"), lty=c(1,2))
```



The blue lie shows the curve formed by the sample mean distribution. The red dashed line shows the normal curve formed by the theoretical mean and standard deviation. As the graph shows, the distribution of means of our simulated sampled distribution appear to follow a normal distribution, due to the Central Limit Theorem. If we increased our number of samples (currently 1000), the distribution would be even closer to the standard normal distribution.