Statistical Inference Coursera – The Data Science Specialization

Course Project Report by Konstantinos Papastamos

Part 1

All the code written for the project can be found on the following Page*

*I didn't use github for obvious reasons (it's public)

The Exercises

Exercise 1

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution
3. Show that the distribution is approximately normal

Difference in means

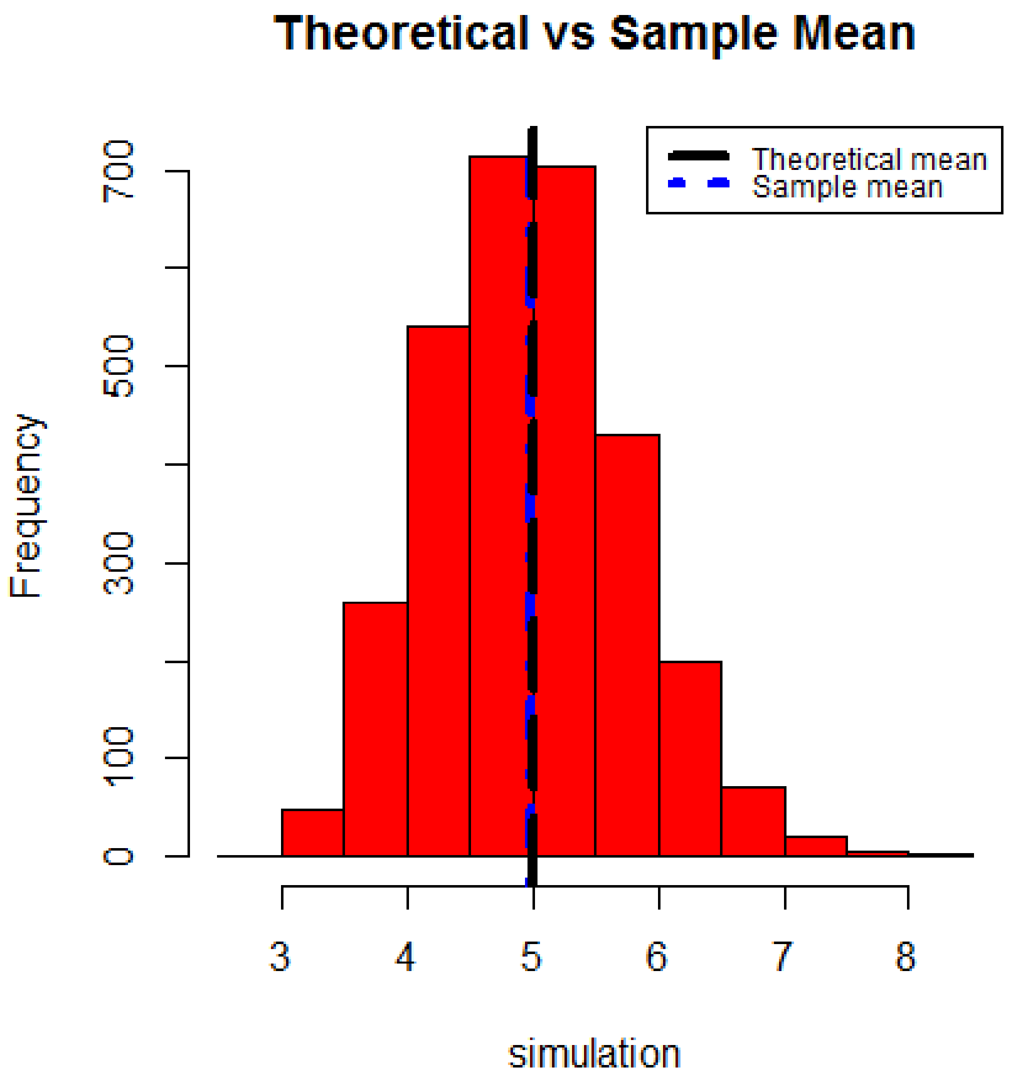
The theoretical mean of the exponential distribution with lambda = 0.2 (as requested) is equal to 1/lambda so 1/0.2 = 5 . In the simulation I did three thousand simulations of 40 exponentials and compared the sample mean to the theoretical mean of the distribution.

set.seed(15)

simulation = replicate(3000,mean(rexp(40,1/5)))

means = mean(simulation)

The sample mean is approximately 5 (4.983771 in my case). In the plot presented above we can see that the difference between the theoretical (Black line) and the sample (Blue line) mean is barely visible.



Difference in Variance

The theoretical variance of the exponential distribution with lambda = 0.2 is equal to ((1/lambda)^2)/n = 0.625 and the theoretical standard deviation is equal to sqrt(theoretical variance) = sqrt(0.625) = 0.7905694

Using the simulation that I used to find the difference in means, I will compare the sample variance and standard deviation of three thousand simulations of 40 exponentials to the theoretical variance and standard deviation of the exponential distribution variance = var(simulation)

stdev = sd(simulation)

The sample variance and standard deviation are approximately equal to the theoretical variance and standard deviation

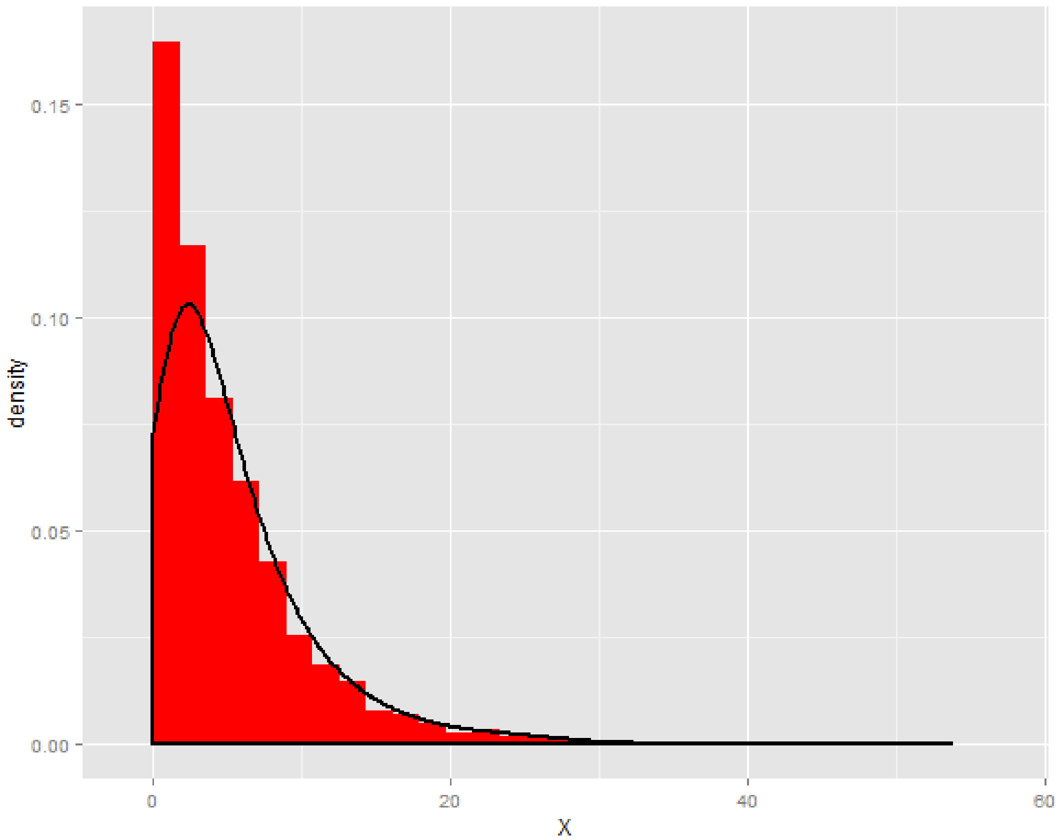
(0.6149436 and 0.7841834 respectively in my case)

Normality Test

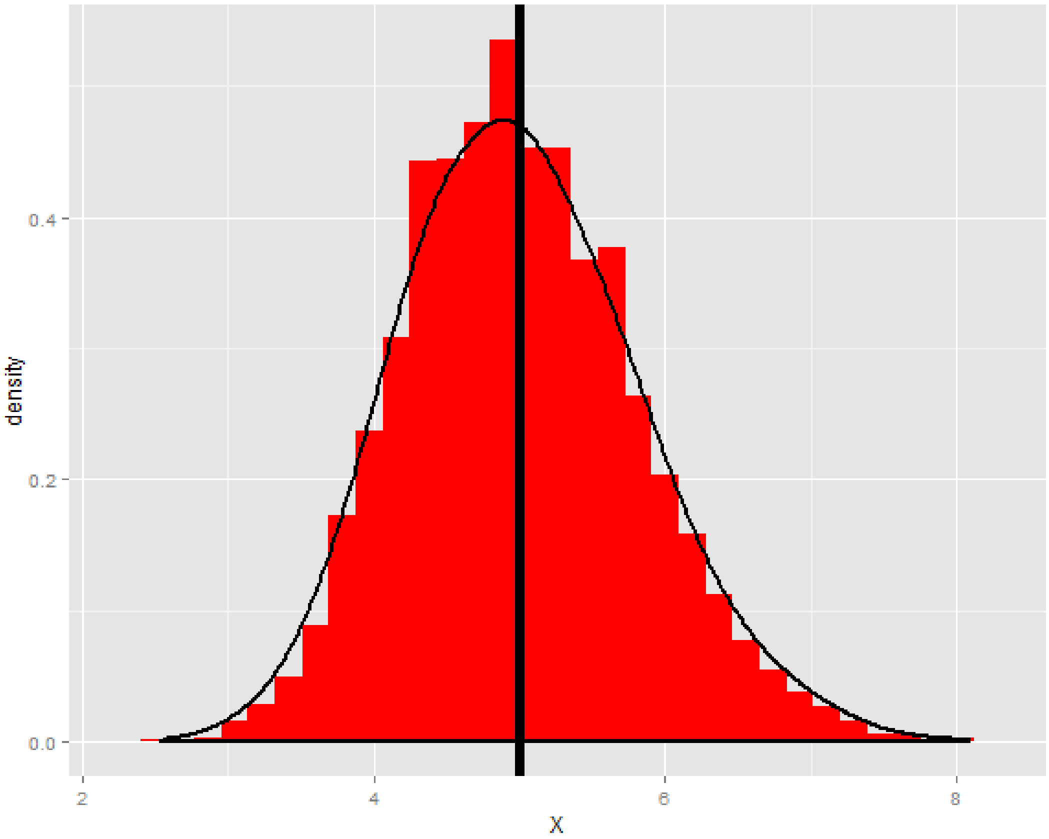
Here I am going to compare the distribution of 4000 random exponentials with the distribution of 4000 averages of 40 random exponentials.

sim1 = rexp(4000,1/5)
sim2 = replicate(4000,mean(rexp(40,1/5)))

Distribution of 4000 random exponentials



Distribution of 4000 averages of 40 random exponentials



In the second plot we can see that the distribution looks far more Gaussian than the original Exponential distribution! The black line in the middle represents the theoretical mean of the exponential distribution, which in our case equals to 5. We can see that the distribution is approximately normal.