JHU Data Science Spcecialization

Statistical Inference Course Project Part 1

Simulation exercise

Author: Rafael Marino

Date: April 21st, 2015

Overview

In this project the exponential distribution will be explored in R and utilized as an example of the Central Limit Theorem (CLT) using simulations. The exponential distribution (a.k.a. negative exponential distribution) is the probability distribution "that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate" (*from Wikipedia*). It 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. For all of the simulations lambda will be set at 0.2. In particular, the distribution of averages of 40 exponentials will be investigated and for this, it is required to do 1000 simulations.

The properties of the distribution of the mean of 40 exponentials will be illustrated via simulation and associated explanatory text. To do that, there are 3 important steps in this report:

1. Comparison between the sample mean and the theoretical mean of the distribution.
2. Analysis of the variability of the sample (via varaince) compared to the theoretical variance of the distribution.
3. Showing that the distribution is approximately normal.

1. Sample Mean vs Theoretical Mean

The first step to answer this question should be to make clear the difference between both concepts. The *sample mean* refers to the average of all possible samples, or at least a very large number (the number of simulations will be set at 1000 in this study), of a given size (n will be set at 40 in this study) drawn from the population. On the other hand, the *theoretical mean* does not deal with samples, instead, it's associated with the population, therefore it is considered a "true" value, but because in applied data analysis we rarely deal with whole populations, it is sometimes called *theoretical* mean.

Once that's clear, we can start the simulation. The code for the simulation can be seen on the Appendix, on the "Simulation & Graphs" section, under **"Code chunk #1"**. The first thing to do is to set the seed, so that the experiment will be reproducible; it is set at 1400. Then the number of simulations, *nosim* is set at 1000, and the sample size, *n* at 40.

Creating random exponential observations is done using the *rexp()* function in R. We use this function to create a matrix of dimension nosim*n (i.e. 1000 rows by 40 columns), the rate parameter, or lambda, is set at 0.2, and the number of rows of the random matrix is set at 1000, so that every row contains 40 observations. This way, every row constitutes a sample of size 40 drawn from the population.

Then using the *apply()* function, we calculate the mean of every row (40 obs ea), and we come up with a vector of 1000 sample means. This vector represents the distribution we're interested in analyzing.Then we calculate its mean (4.934) and compare it with the theoretical mean of the exponential distribution, which is 1/lambda, or 1/0.2 which is 5. As you can see, both numbers are pretty close with a difference of 0.065. This can be seen graphically using Figure 1 in the Simulation & Graphs section. The red line (sample mean), is quite close to the blue line (theoretical mean).

2. Sample Varaince vs Theoretical Variance

The process to compare the sample variance vs the theoretical variance is quite similar. The code for this simulation can be found on the Appendix, on the "Simulation" section, under the **"Code chunk #2"**. We can use the same matrix of random exponentials, of dimension 1000*40, but now instead of taking the mean of each row, we will take the variance of each row. This creates a vector of 1000 (40-sample) variances. This vector represents the approximation of the distribution of sample variances we want to study. We take the mean of this vector and we end up with the sample variance, 24.371. This number is compared with the theoretical variance, 1/lambdaˆ2, or 1/0.2ˆ2, or 1/0.04 or 25. With a difference between them of 0.6285.

It is very easy to see this graphically. In the Appendix go to **Figure 2** of the Simulations and Graphs Section and see how the distribution of sample variances is centered at 24.3 (red line) and the theoretical mean is exactly 25 (blue line). This result is consistent with the lectures seen in class about the sample variance. Namely, that the distribution of the sample variance is centered at what it's estimating.

3. Showing that the distribution is approximately normal

This question entails comparing the distribution of the sample mean with that of the population, meaning the exponential distribution. This can be done using the Central Limit Theorem (CLT). Formally, the CLT states that *"for any sample X1...Xn that is IID (idependently and identically distributed), its mean (μ), has a distribution which is approximately Normal, with mean μ and variance σ ˆ2/n. This is remarkable since nothing is assumed about the distribution of Xi, except the existence of mean and variance"* ("All of Statistics", Wasserman, 2004).

The larger the sample size, the more normal the approximation of the sample mean distribution to the Normal distribution. In our study, the sample size is 40, and as you can see in **Figure 3** (Simualtions and Graphics Section), the red density (which approximates the sample mean distribution), is quite similar to the blue density (which is a normal distribution with μ=5 and σ= 5/sqrt(40)).

Appendix

1. Simulations & Graphics

Code chunk #1

```
set.seed(1400) #Set seed, to make the simulation results reproducible
nosim <- 1000 #Number of simulations
n <- 40 #Setting the sample size as 4
expData <- matrix(rexp(nosim*n, rate=0.2), nrow=nosim)  #1000*40 matrix
thousandMeans <- apply(expData, MARGIN=1,mean) #Margin=1 applies on rows
mean(thousandMeans) #Calculates the sample mean
```

```
## [1] 4.934351
```

Code chunk #2

```
thousandVars <- apply(expData, MARGIN=1,var)
#Margin=1 causes to take the Variance of every row (40 obs each)
mean(thousandVars) #Calculates the mean of the sample variance distribution.
```

```
## [1] 24.37145
```

Figure 1. Sample Mean(red) vs Theoretical Mean(blue)

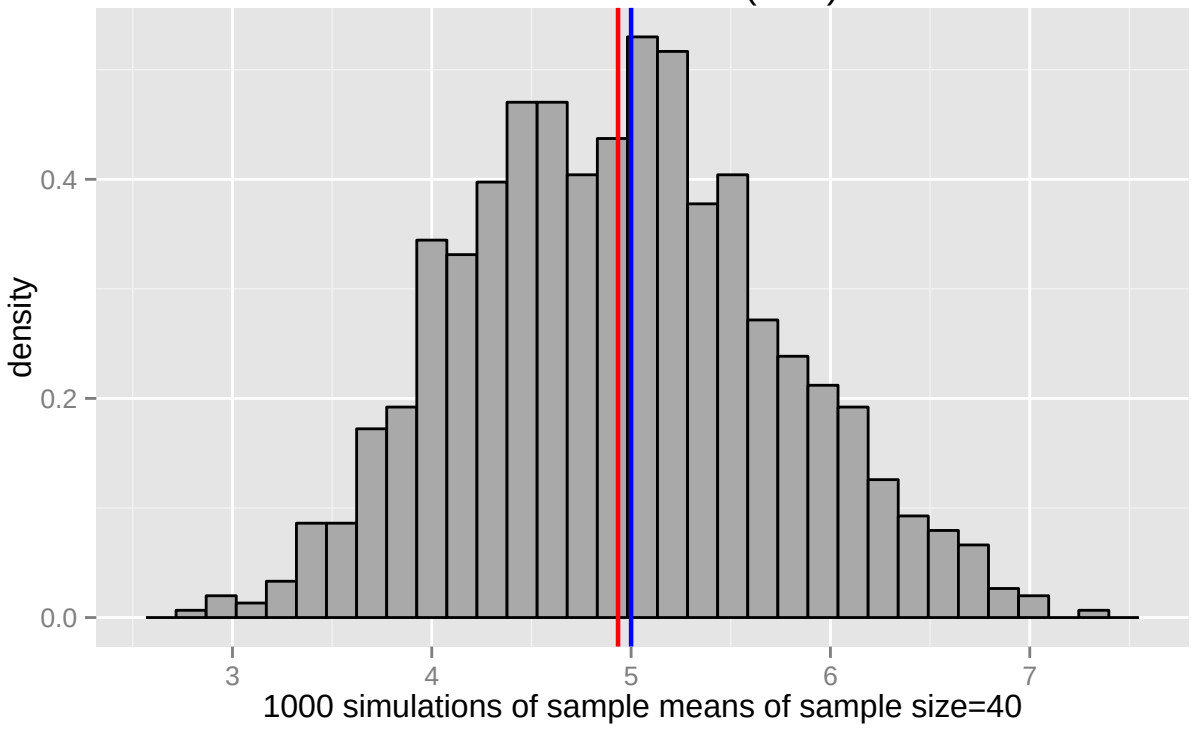

Figure 2. Sample Variance(red) vs Theoretical Variance(blue)

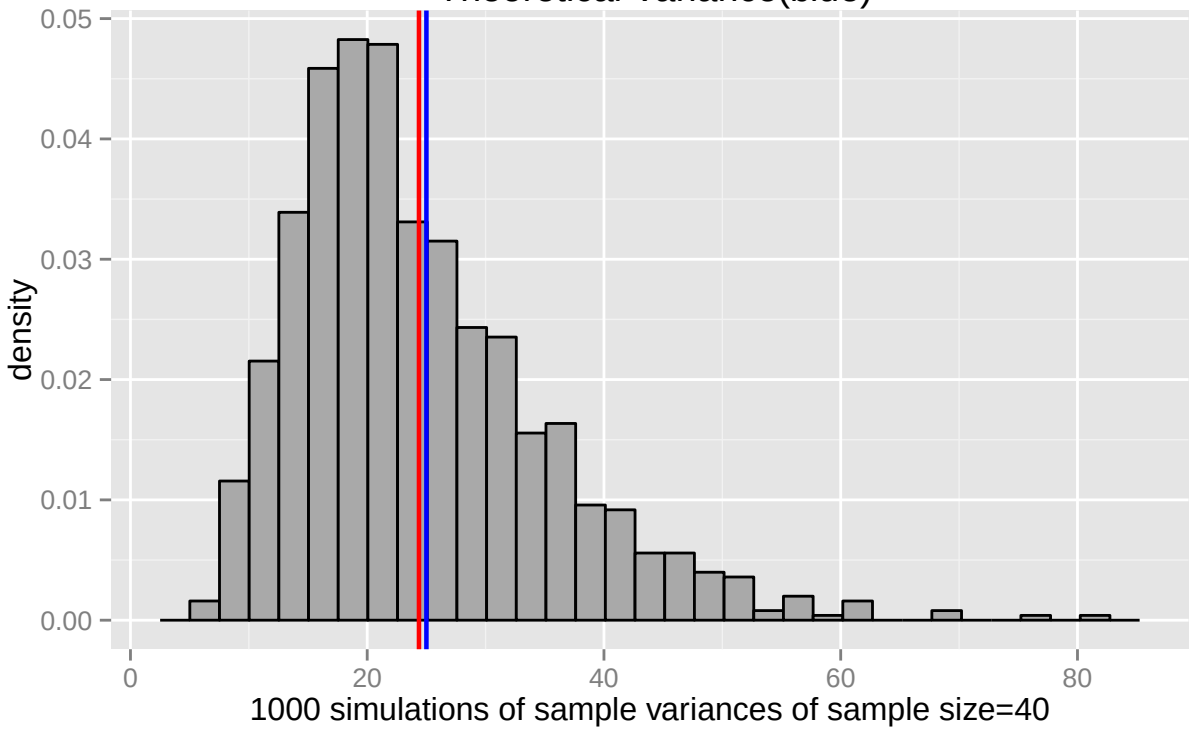

Figure 3. Sample mean distribution (n=40)
(red) vs Normal distribution(blue)