Coursera's Statistical Inference - Simulation

by Thomas Guenther

September 2014

Introduction

For this assignment our goal is to simulate the exponential distribution with R. This type of distribution can be simulated in R via the following formula: `rexp(n, lambda)`. The mean and the standard deviation are also given with $\frac{1}{\lambda}$. For this exercise we have to set λ=0.2 and we have to compute at least 1000 distributions of averages of 40 exponentials. Due to the page limit i've setup a GitHub repository with all my files: https://github.com/dreammaster38/coursera_statistical_inverence_project_1

1. Theoretical and computed Center of distribution

The theoretical center of distribution is given by $\frac{1}{\lambda}$. So it is $\frac{1}{0.2} = 5$. On the other hand we have 1000 simulated values of our simulated distribution which is:

```
## [1] 4.989281
```

As we can see the mean of our simulated distribution is close to the theoretical mean.

2. Theoretical and computed variance

Let us first take a look at the theoretical variance. This will be calculated as follows:

$$Var(\bar{X}) = \frac{\frac{1}{\lambda^2}}{n} = \frac{1}{\lambda^2 \cdot n} = \frac{1}{0.2^2 \cdot 40} = 0.625$$

Now we compare this theoretical variance to the variance of our sample disribution. It's value is:

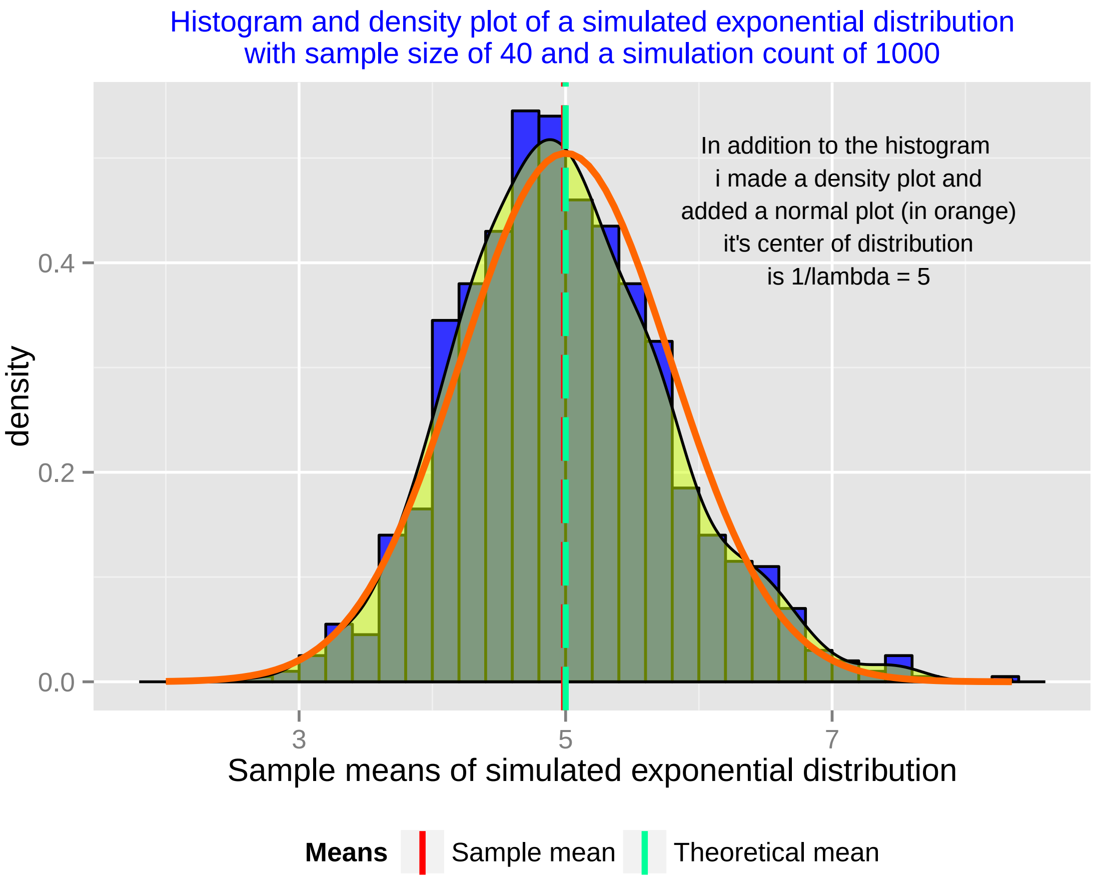
```
## [1] 0.6475366
```

As for the mean it's values are close each other too.

3. Is our sample approximately normal?

For our purposes, the CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases.

If we have samples with a high enough sample size n (in literature they say more than 30) and we create the average of them with a high count of repeats, then the distribution of this values is approximately normal also if the population distribution isn't a normal one. This is stated by the Central Limit Theorem. The following plot shows this behavior:



As you can see in the histogram we get an approximately normal distribution out of our 1000 simulated averages with sample size 40. I added a density plot as well so the result gets more clear.

4. Coverage of the confidence intervall

A confidence interval (CI) measures how reliable an estimate is and is an intervall estimation. The confidence level of 95% is one of the most frequently used confidence intervalls in practice, so will test against it. That means we will test how many of the population parameters of 1000 simulated exponantial distribution samples with size 40 falls into the CI-bounds of 95%. We use this given formula for calculation:

$$\left[\bar{X} - 1.96 \cdot \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{s}{\sqrt{n}}\right] \hat{=} mean(sample) \pm 1.96 \cdot \frac{SampleStandardDeviation}{\sqrt(n)}$$

```
## [1] 0.917
```

As can been seen $\approx$ 92% of our sample means estimate the population mean.