Statistical Inference Course Project Part 1

jacethedatascientist

August 12, 2019

Statistical Inference Course Project Part 1

Jace Galleon

I. Overview

This is the first part of the Statistical Inference Course Project from Coursera.

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, λ) where λ is the rate parameter.

The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. Set $\lambda = 0.2$ for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

II. Objectives

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.
    - focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

As a motivating example, compare the distribution of 1000 random uniforms.

III. Analysis

A. Preparation

You can use the **rexp(x,y)** function in R where **x** could be the sample size n and **y** could be *lambda* (λ) which is the rate parameter. Both the mean and standard deviation of exponential distribution is $1/\lambda$. Set the value of $\lambda = 0.2$ for all the simulations. In this simulation, you will investigate the distribution of the average of 40 exponentials.

Note that you'll gonna be doing a simulation to get a thousand average for 40 exponentials.

```r
set.seed(0000)
exp <- 40
lambda <- 0.2
n <- 1000
sim_sample <- replicate(n, rexp(exp, lambda))
simulation <- apply(sim_sample, 2, mean)
```

B. Comparison using Mean

```r
print(paste("Simulated Mean: ", round(mean(simulation) ,4)))
```

```
## [1] "Simulated Mean:  4.9897"
```

```r
print(paste("Theoretical Mean: ", round(1/lambda ,4)))
```

```
## [1] "Theoretical Mean:  5"
```

There is a *0.21%* difference between the mean of the simulated and the theoretical mean which is 4.9897 and 5, respectively.

C. Comparison using Variance and Standard Deviation

```r
print(paste("Simulated Standard Deviation: ", round(sd(simulation) ,4)))
```

```
## [1] "Simulated Standard Deviation:  0.7862"
```

```r
print(paste("Theoretical Standard Deviation: ", round((1/lambda)/sqrt(exp) ,4)))
```

```
## [1] "Theoretical Standard Deviation:  0.7906"
```

```r
print(paste("Simulated Standard Variance: ", round((sd(simulation))^2 ,4)))
```

```
## [1] "Simulated Standard Variance:  0.6182"
```

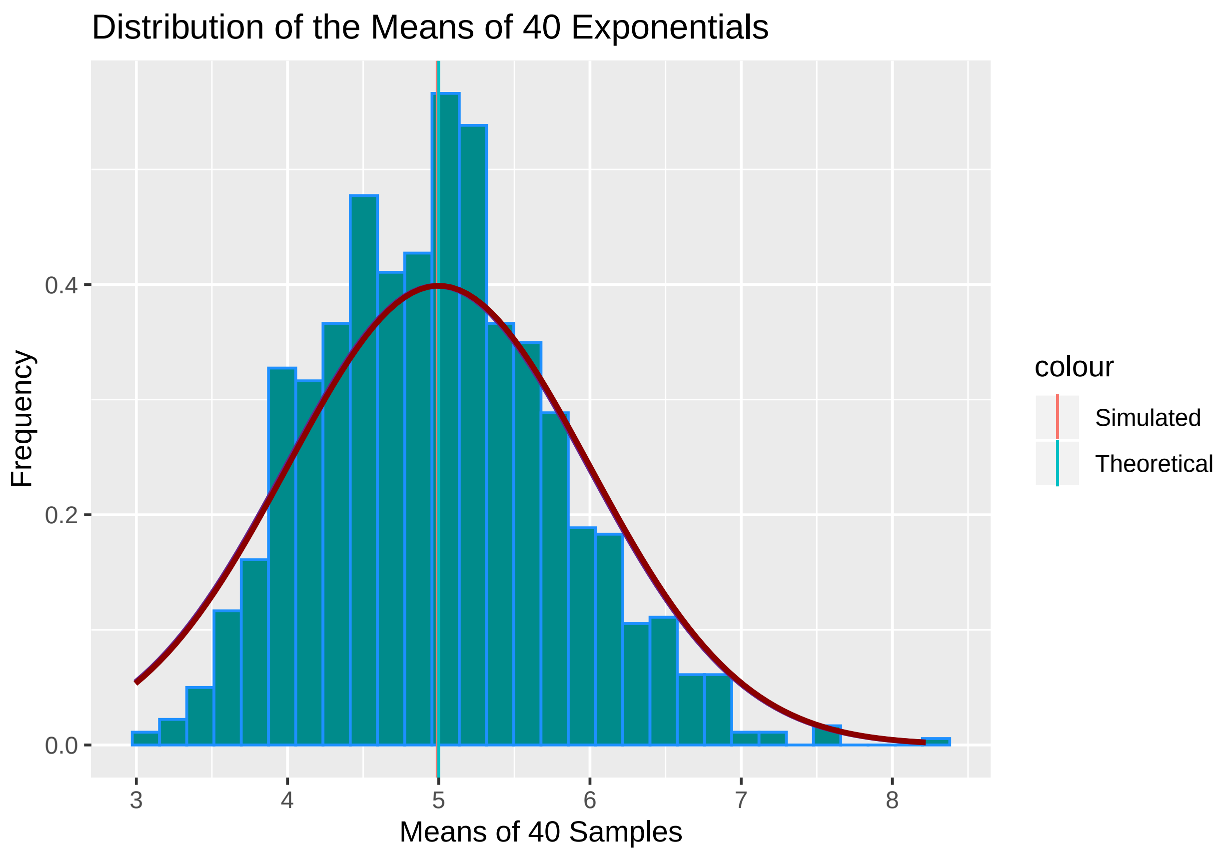
```r
print(paste("Theoretical Standard Deviation: ", round(((1/lambda)/sqrt(exp))^2 ,4)))
```

```
## [1] "Theoretical Standard Deviation:  0.625"
```

There is a *0.56%* difference between the mean of the simulated and the theoretical standard deviation which is 0.7862 and 0.7906, respectively. While the simulated and theoretical variance is 0.6182 and 0.625, respectively, with a difference of *1.1%*.

D. Distribution Analysis

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot shows the **histogram** containing the distribution of of the simulated values. The mean of the simulated versus the theoretical values or shown as the vertical lines.

The **red** line shows the normal curve formed by the Theoretical Mean while the **purple** is for the Simulated.

This formally ends the Part 1 of the Course Project. Thank You!