Statistical Inference Course Project | Part 1

Siyang Ni

1/4/2021

1. Overview

This is the final project for John Hopkins University's Statistical Inference Course, which consists of two parts:

1. A simulation exercise.
2. A basic inferential data analysis.

This is the first part of the project, in which a simulation will be conducted to explore inference and do some simple inferential data analysis.

```
# Setting the environment
library(tidyverse)
```

2. Simulation

Per required, we will generate **1000** simulations of the average of **exponentials**. The mean and the standard deviation of the exponential distribution is **1/lambda**. In our simulations, we set **lambda = 0.2**.

```
# Set seed at 1
set.seed(1)

# Set relevant parameters
lambda <- 0.2
n <- 40
sim <- 1000

# Run simulations
```

```r
simulation_exp <- replicate(sim, rexp(n, lambda))

# Take the sample mean
sim_avg <- apply(simulation_exp, 2, mean)
```

3. Sample Mean versus Theoretical Mean

The code below shows:

1. The value of the mean of the 1,000 averages of 40 exponentials (simulated mean)
2. The theoretical mean of the exponential distribution
3. The difference between the simulated mean and the theoretical mean

We can see that the difference between the two means is very small.

```r
# Simulated Mean
sim_mean <- mean(sim_avg)
sim_mean
```

```
## [1] 4.990025
```

```r
# Theoretical Mean
theo_mean <- 1/lambda
theo_mean
```

```
## [1] 5
```

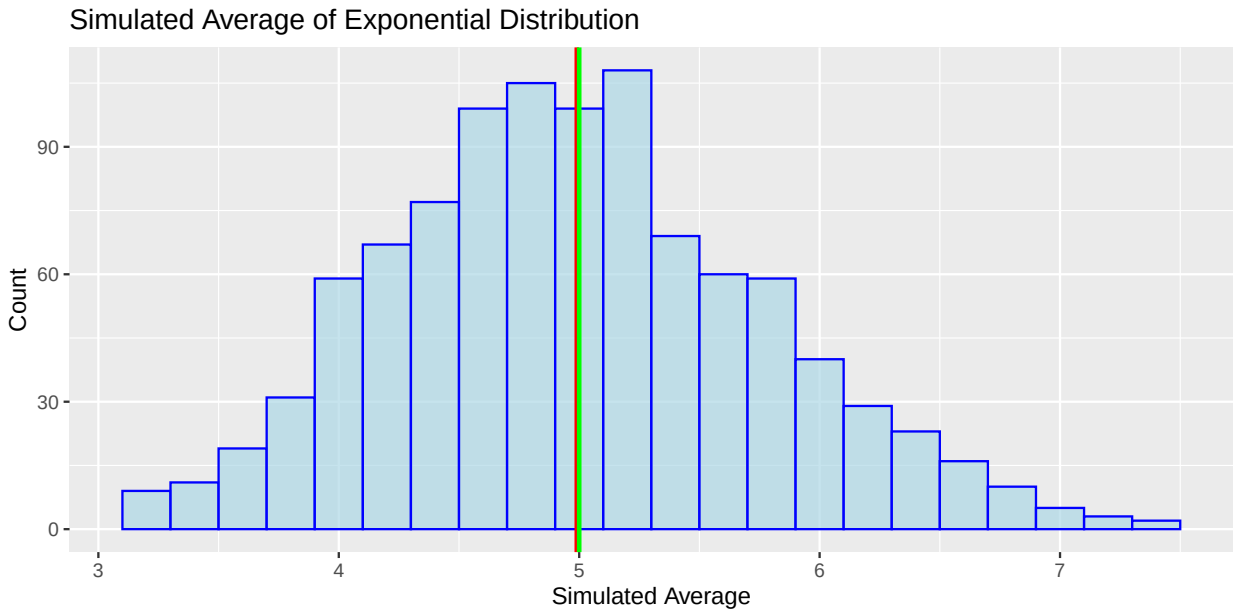
```r
# Difference
theo_mean - sim_mean
```

```
## [1] 0.009974799
```

The histogram just confirms what we have shown above. The simulated mean is marked by the red vertical line, the theoretical mean is marked by the green vertical line. The two lines are very close to each other.

```r
# Visualization
sim_avg1 <- data.frame(sim_avg)

g = ggplot(data = sim_avg1, aes(x = sim_avg))
g = g + geom_histogram(binwidth = 0.2, fill = "lightblue", color = "blue",
    alpha = 0.7)
g = g + geom_vline(xintercept = sim_mean, col = "red", size = 1)
g = g + geom_vline(xintercept = theo_mean, col = "green", size = 1)
g = g + labs(title = "Simulated Average of Exponential Distribution",
    x = "Simulated Average", y = "Count")
g
```



4. Sample Variance versus Theoretical Variance

The code below shows:

1. The value of the variance of the 1,000 averages of 40 exponentials (simulated variance)
2. The theoretical variance of the exponential distribution
3. The difference between the simulated variance and the theoretical variance

We can see that the difference between the two variances is very small.

```r
# Simulated Variance
sim_var <- sd(sim_avg)^2
sim_var
```

```
## [1] 0.6111165
```

```
# Theoretical Variance
theo_var <- (1/lambda/sqrt(n))^2
theo_var
```

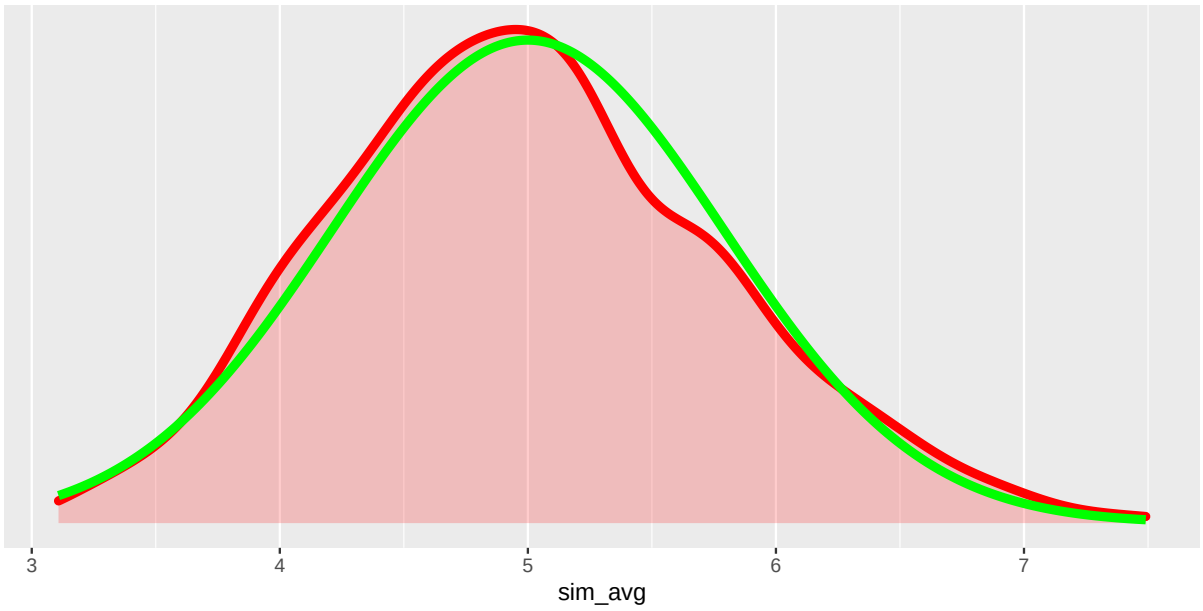
```
## [1] 0.625
```

```
# Difference
theo_var - sim_var
```

```
## [1] 0.01388353
```

The following graph illustrates what we discussed above. The red curve is the simulated distribution, the green curve is the theoretical distribution according to the Central Limit Theorem. We can see that the two distributions have very similar variances.

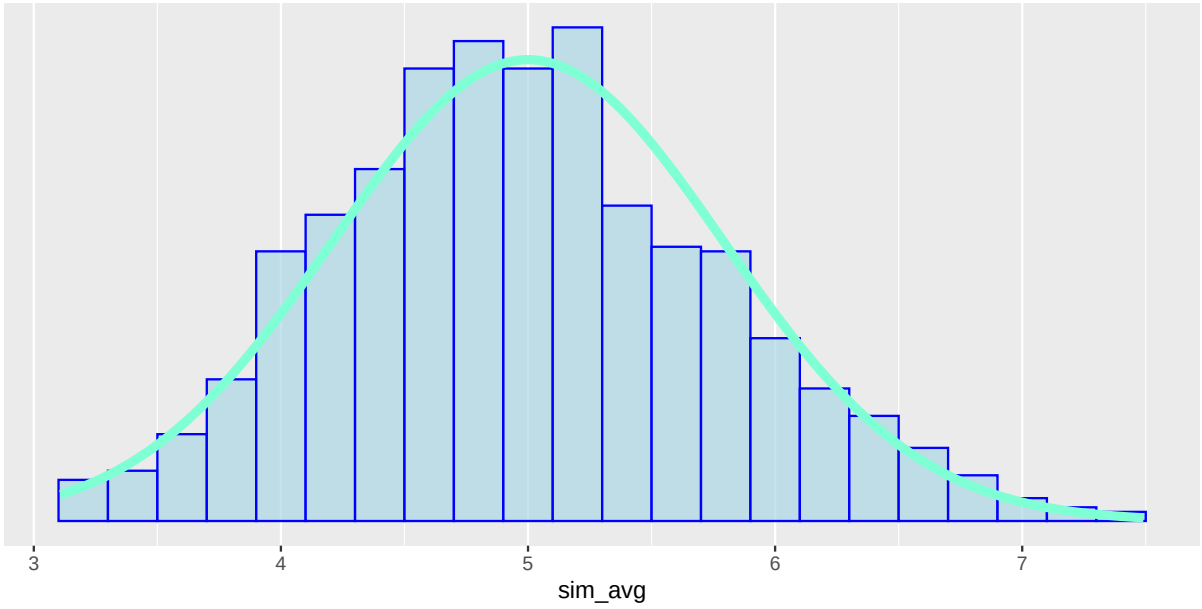
```
g1 = ggplot(data = sim_avg1, aes(x = sim_avg))
g1 = g1 + geom_density(color = "red", size = 2, fill = "red",
    alpha = 0.2)
g1 = g1 + stat_function(color = "green", size = 2, fun = dnorm,
    n = 1000, args = list(mean = 1/lambda, sd = sqrt(theo_var))) +
    ylab("") + scale_y_continuous(breaks = NULL)
g1
```



5. Distribution

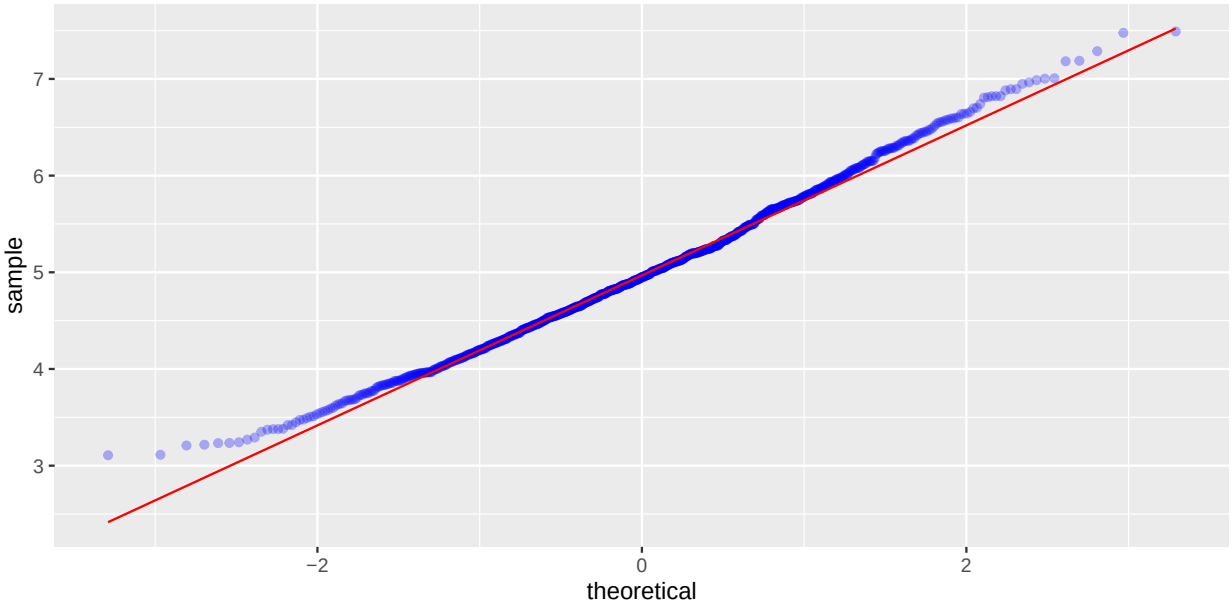
We fit the theoretical normal curve to the histogram of the simulated distribution. From the following graph we can see that the simulated distribution is roughly normal.

```
g2 = ggplot(data = sim_avg1, aes(x = sim_avg))
g2 = g2 + geom_histogram(aes(y = ..density..), binwidth = 0.2,
    fill = "lightblue", color = "blue", alpha = 0.7)
g2 = g2 + stat_function(color = "aquamarine", size = 2, fun = dnorm,
    n = 1000, args = list(mean = 1/lambda, sd = sqrt(theo_var))) +
    ylab("") + scale_y_continuous(breaks = NULL)
g2
```



The Quantile-Quantile plot confirms what we see from the above histogram. Therefore, we can be confident that the simulated distribution is roughly a normal distribution.

```
g3 = ggplot(data = sim_avg1, aes(sample = sim_avg))
g3 = g3 + stat_qq(color = "blue", alpha = 0.3) + stat_qq_line(color = "red")
g3
```


sample
7
6
5
4
3
−2
0
2
theoretical