Statistical Inference Course Final Project - Part 1

André Marinho

05/08/2020

1. Overview

This is the first part of project report from Coursera Statistical Inference Course. In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem.

2. Simulations

Tasks:

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.

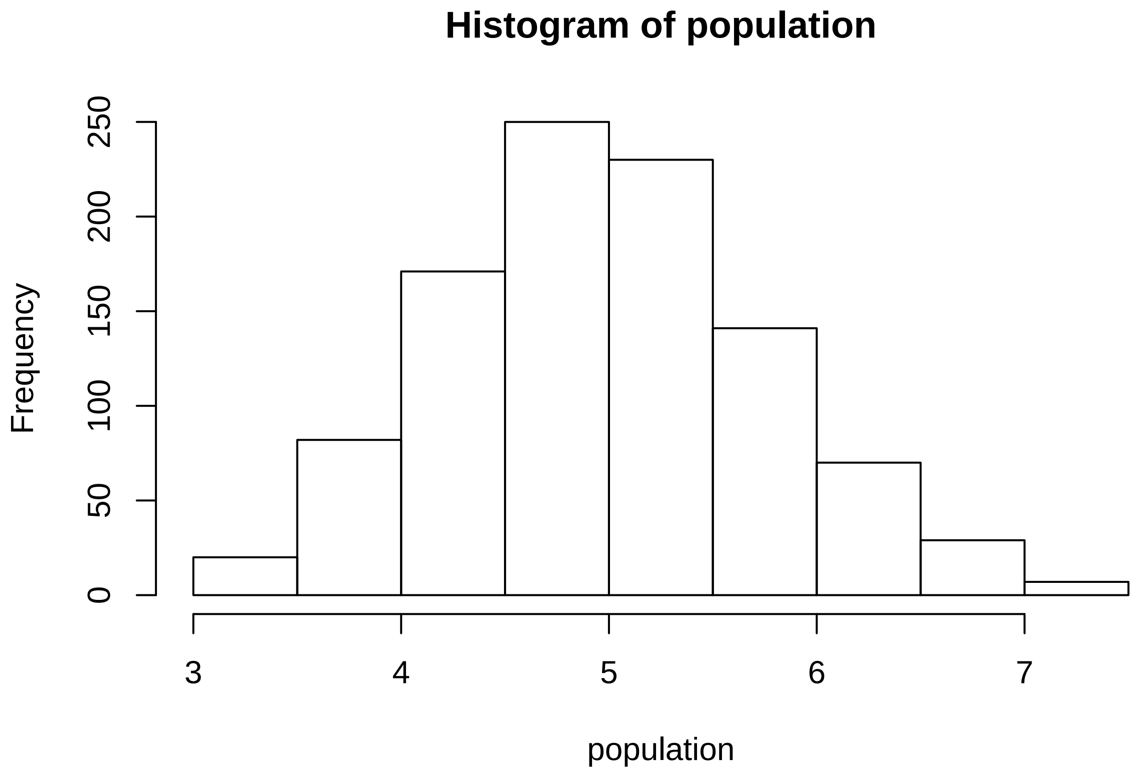
2.1.1. Set pre-defined parameters

As stated in the exercise instructions, the following parameters (assumptions) are set:

```r
lambda <- 0.2
n <- 40
n_simulations <- 1000
set.seed(1)
```

2.1.2. Plot histogram to compare the distribution of 1000 simulations

```r
population <- NULL
for (i in 1 : n_simulations)
        population <- c(population, mean(rexp(n, lambda)))
hist(population)
```



2.1.3. Sample Mean vs. Theorical Mean

```
sample_mean <- mean(population)
theorical_mean <- 1 / lambda
print(sample_mean)
```

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

```
print(theorical_mean)
```

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

The means are very close. Looking for confidence interval:

```
t.test(population)[4]
```

```
## $conf.int
## [1] 4.941515 5.038536
## attr(,"conf.level")
## [1] 0.95
```

2.2. Sample Variance vs. Theorical Variance

```
sample_variance <- var(population)
theorical_variance <- ((1 / lambda) ^ 2) / n
print(sample_variance)
```

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

```
print(theorical_variance)
```

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

The variances are very close too.

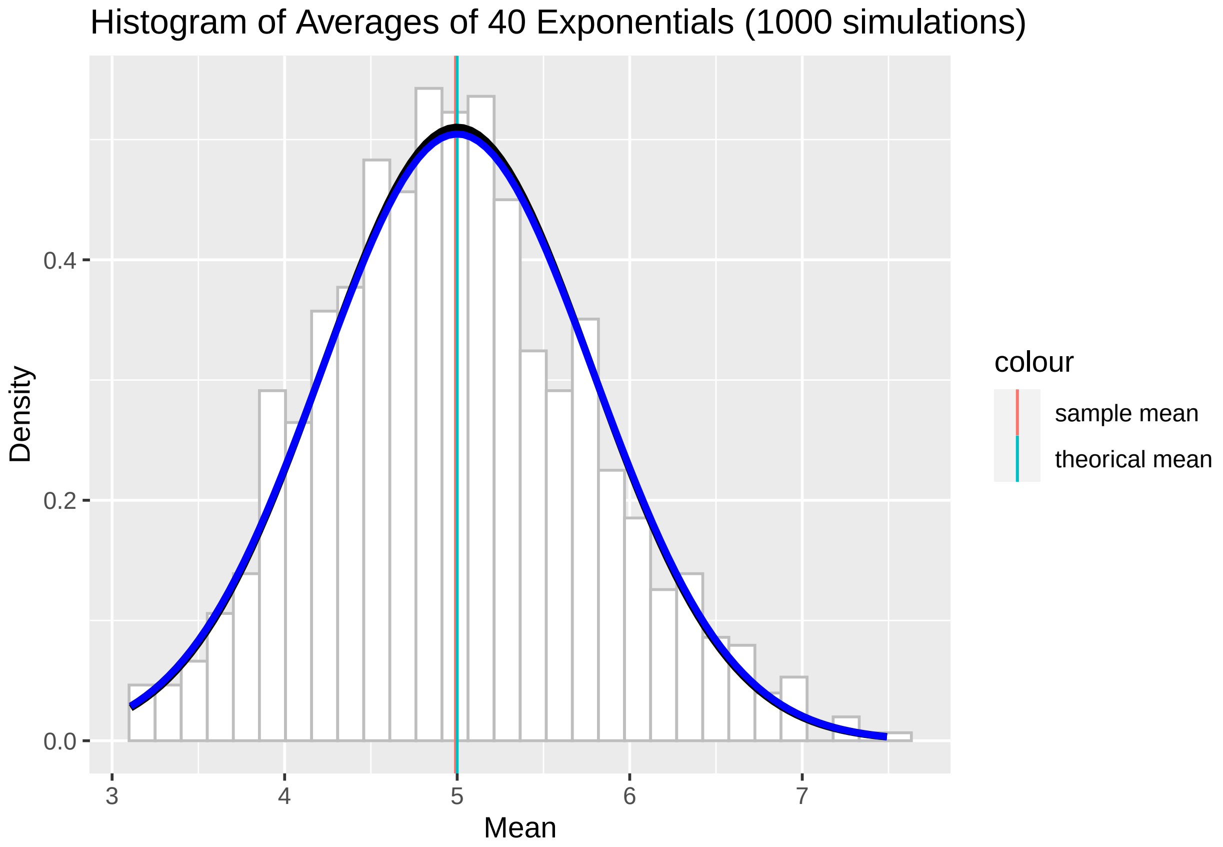
2.3. Distribution

```
library(ggplot2)

population <- data.frame(population)

distribution_plot <- ggplot(population, aes(x=population)) +
        geom_histogram(aes(y=..density.., fill=..density..), colour="grey", fill="white") +
        labs(title="Histogram of Averages of 40 Exponentials (1000 simulations)", y="Density", x="Mean"
        geom_vline(aes(xintercept=sample_mean, colour="sample mean")) +
        geom_vline(aes(xintercept=theorical_mean, colour="theorical mean")) +
        stat_function(fun=dnorm, args=list(mean=1/lambda, sd=sqrt(sample_variance)), colour="black", si
        stat_function(fun=dnorm, args=list(mean=1/lambda, sd=sqrt(theorical_variance)), colour="blue",
distribution_plot
```

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



The sample mean for 40 exponential distributions, simulated 1000 times, are very close to the theorical mean for a normal distribution.