Statistical Inference Course Project - Simulation Exercise

Prykhodko Pavel

June 6, 2018

Overview

In this simulation exercise we'll look into the distribution of averages of 40 exponentials in R and compare it with the Central Limit Theorem.

In particular, we will illustrate:
1. Sample mean and theoretical mean
2. Sample variance and theoretical variance
3. Normality of the distribution

To reproduce this exercise you will need *ggplot2* library.
To get the same results you can set the seed to *87310*.

```
library(ggplot2)
set.seed(87310)
```

Simulation

We will simulate 1000 exponential distributions with R function rexp(n, lambda).
lambda will be 0.2 for all our simulations.
The mean of the exponential distribution is 1/lambda, and the standard deviation is also 1/lambda.

Let's make our simulations and store them into a matrix:

```
lambda <- 0.2
sampleSize <- 40
simulationsCount <- 1000
simulations <- rexp(simulationsCount * sampleSize, rate = lambda)
simulationsMatrix <- matrix(simulations, simulationsCount, sampleSize)
```

Sample Mean versus Theoretical Mean

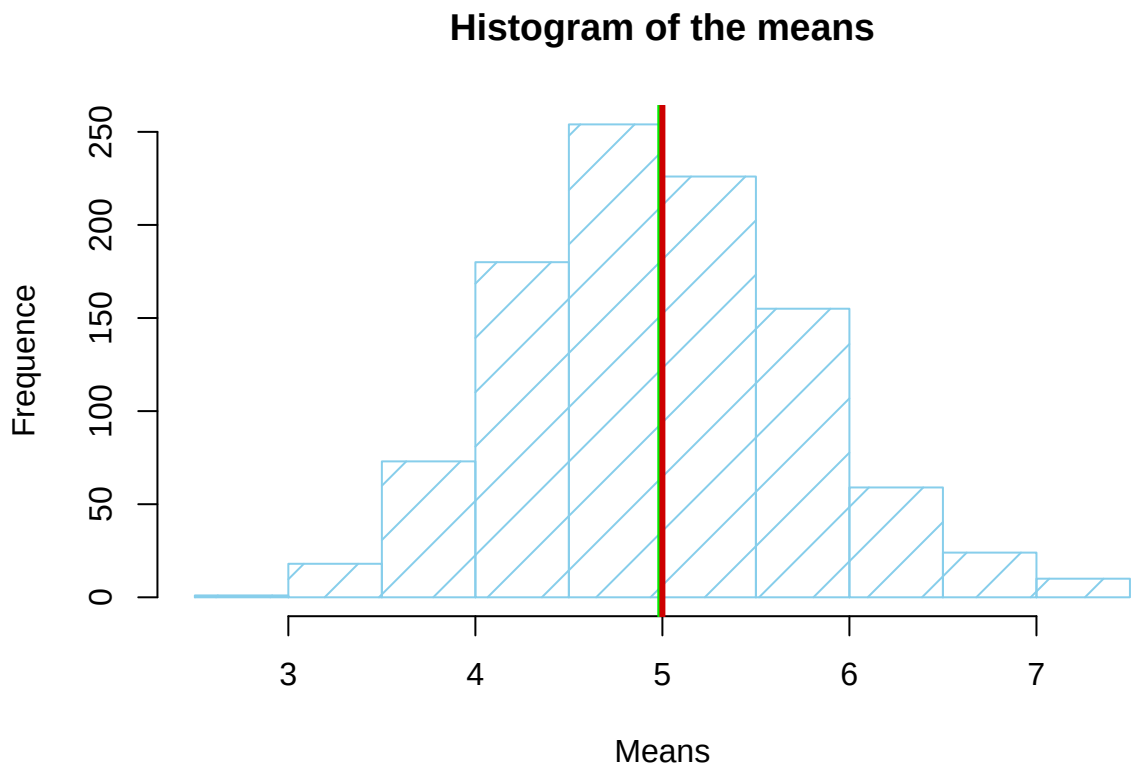
The exponential distribution mean is 1/lambda, so for this distribution theoretical mean should be `1 / lamba(0.2) = 5`.
Let's calculate the means of our simulations.

```
simulationsMean <- rowMeans(simulationsMatrix)
sampleMean <- mean(simulationsMean)
theoreticalMean <- (1 / lambda)
```

We will plot the means histogram and draw a sample mean with green line and a theoretical mean with red line on that plot for comparison.

```
hist(
  simulationsMean,
  xlab = "Means",
```

```
  ylab = "Frequence",
  main = "Histogram of the means",
  col = "skyblue",
  density = 5
)
abline(v = sampleMean, lwd = 3, col = "green2")
abline(v = theoreticalMean, lwd = 3, col = "red3")
```



As you see - the red and green lines are nearly same, because of sample mean `mean(simulationsMatrix)` is 4.9886958 and that is very close to 5.

Sample Variance versus Theoretical Variance

Let's compare sample and theoretical variances.

```
theoreticalVariance <- ((1 / lambda) ^ 2) / sampleSize
sampleVariance <- var(simulationsMean)
```

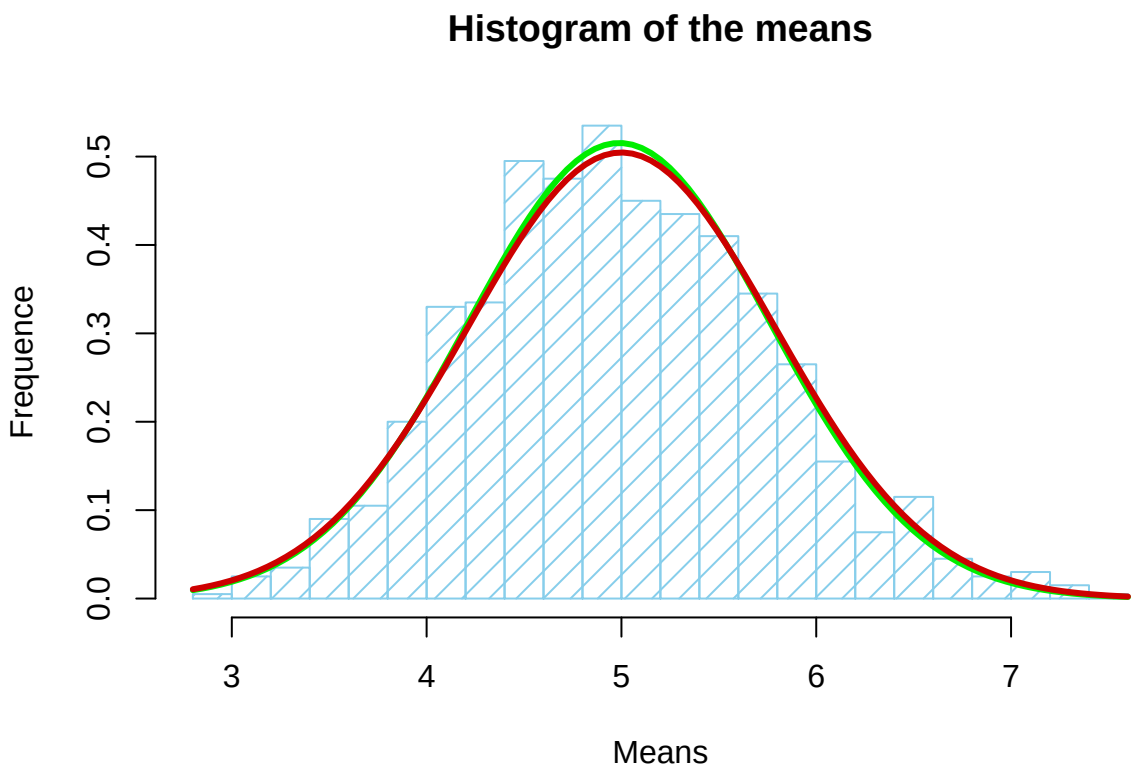
Variance is `((1 / lambda) ^ 2) / sampleSize`, so theoretical one will be 0.625.
The variance of our sample `var(simulationsMean)` is 0.599169.
That is not as good as in previous comparison, but still very close.

Distribution

In this question we need to demonstrate, that our simulations distribution is approximately normal.
The best way to do it is to draw a sample distribution and a normal distribution lines in one plot.
We will mark distributions with green line for sample and red line for normal as we did in the first plot.

```
hist(
  simulationsMean,
  xlab = "Means",
  ylab = "Frequence",
  main = "Histogram of the means",
  col = "skyblue",
  breaks = 20,
  prob = TRUE,
  density = 10
)

curve(
  dnorm(x, mean = sampleMean, sd = sqrt(sampleVariance)),
  col = "green2", lwd = 3, add = TRUE
)
curve(
  dnorm(x, mean = theoreticalMean, sd = sqrt(theoreticalVariance)),
  col = "red3", lwd = 3, add = TRUE
)
```



Again, red and green curves are close, so we can make a conclusion that distribution is approximately normal.