Statistical Inference - Simulation Project - Part 1

Pierre Baudin

February 7, 2017

Part 1: Simulation Exercise Instructions

Overview

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, lambda)** where lambda 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.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

Show the sample mean and compare it to the theoretical mean of the distribution.

Simulation

The code in the following section performs a simulation to generate a 1000*40 matrix containing 1000 simulations (draw) of 40 exponential distributions observations (sampling size).

The sample mean of the sample distribution is calculated by applying the mean function to each row and then to the resulting vector of means.

Theoretically both mean and variance have value 1/lambda.

```
# set.seed for reproducibility
set.seed(122333)

# rate parameter
lambda = 0.2
# sample size
n <- 40
# Number of draw
Nsim <- 1000

# Simulation of Nsim number of draw of sample size n
SimSample <- matrix(rexp(n * Nsim, rate = lambda), nrow = Nsim, ncol = n)

# Computation of the Sample Mean of the distribution
SimSampleMeanVec <- apply(SimSample, 1, mean)
SimSampleMean <- mean(SimSampleMeanVec)

# Theoritical Mean of the distribution
mu <- 1 / lambda
```

Sample Mean versus Theoritical mean result

```
data.frame(Theoritical = mu, Simulation = SimSampleMean)
```

```
##   Theoritical Simulation
## 1           5   4.988875
```

We can see that the theoritical mean and the sample mean are very close to each other. To observe a better result, increasing the sample size and/or the number of draw can help.

Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

Computation

```
# Theoritical variance
sigma2 <- (1 / lambda^2) / n

# Computation of the variance of the Simulation set
SimSigma <- var(SimSampleMeanVec)
```

Sample variance versu theoritical variance

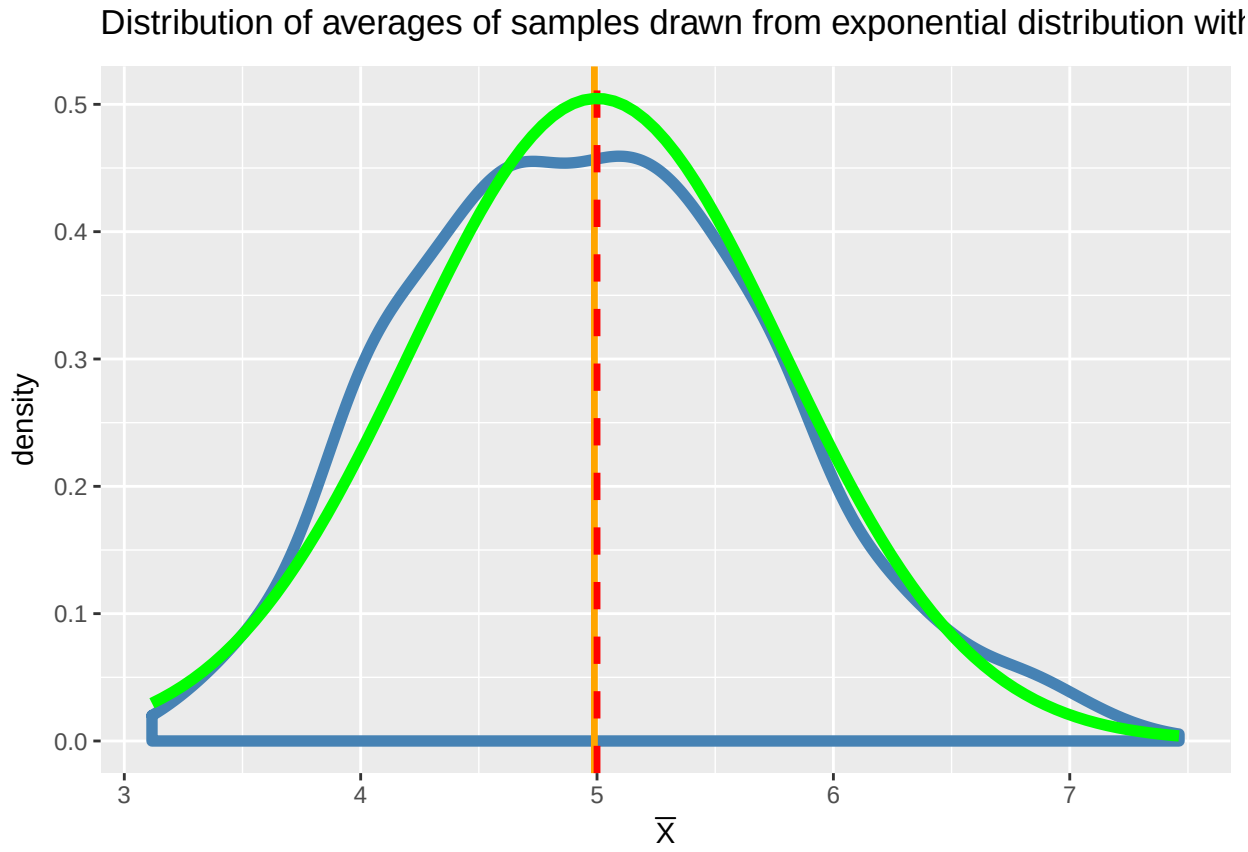
```
# Result
data.frame(Theoritical = sigma2, Simulation = SimSigma)
```

```
##   Theoritical Simulation
## 1       0.625  0.6173667
```

The comparison of the theoritical variance and the variance calculated from the simulation data set is shown above. We can see that the output values are similar.

Show that the distribution is approximately normal.

```
library(ggplot2)
ggplot(data.frame(SimSampleMeanVec), aes(x = SimSampleMeanVec)) +
      geom_density(colour = "steelblue", size = 2) +
      geom_vline(xintercept = mean(SimSampleMeanVec), size = 1.2, lty = 1, colour = "orange") +
      geom_vline(xintercept = mu, size = 1.2, colour = "red", lty = 2) +
      stat_function(fun = dnorm, args = list(mean = mu, sd = sqrt(sigma2)), colour = "green", size = 2)
      ggtitle(expression(paste("Distribution of averages of samples drawn from exponential distribution
      xlab(expression(bar(X)))
```



The plot of the thoritical normal distribution and the calculated sample mean distribution from our simulation are shown in green and blue, respectively. The theoritical mean represented by the red dotted line and the calculated sample mean from the simulation represented by the yellow line are overlapping each other. Visually, we can consider that the output of the simulation is approximately normal. Once again, by increasing the sample size and/or the number of draw, the simulated sample mean distribution will get closer to the normal distribution.