Part 1: Simulation Exercise

Simone

2/22/2020

Overview

The goal of this project is to investigate the exponential distribution in R and compare it with the Central Limit Theorem. In particular, the distribution of averages of 40 random exponentials will be investigated by running 1000 simulations.

Simulations

First, all the libraries needed for the analysis are loaded, some constants related to the simulation are set, and the random seed is also set.

```
# Dependencies
library(ggplot2)

# Set constants
nosim <- 1000
n <- 40
lambda <- 0.2

# Set the random seed
set.seed(0704)
```

Then, the random exponentials are drawn, and the averages (in groups of 40) are calculated.

```
# Draw the random exponentials and calculate the average
sample <- rexp(nosim * n, lambda)
sample_avg <- apply(matrix(sample, nosim), 1, mean)
```

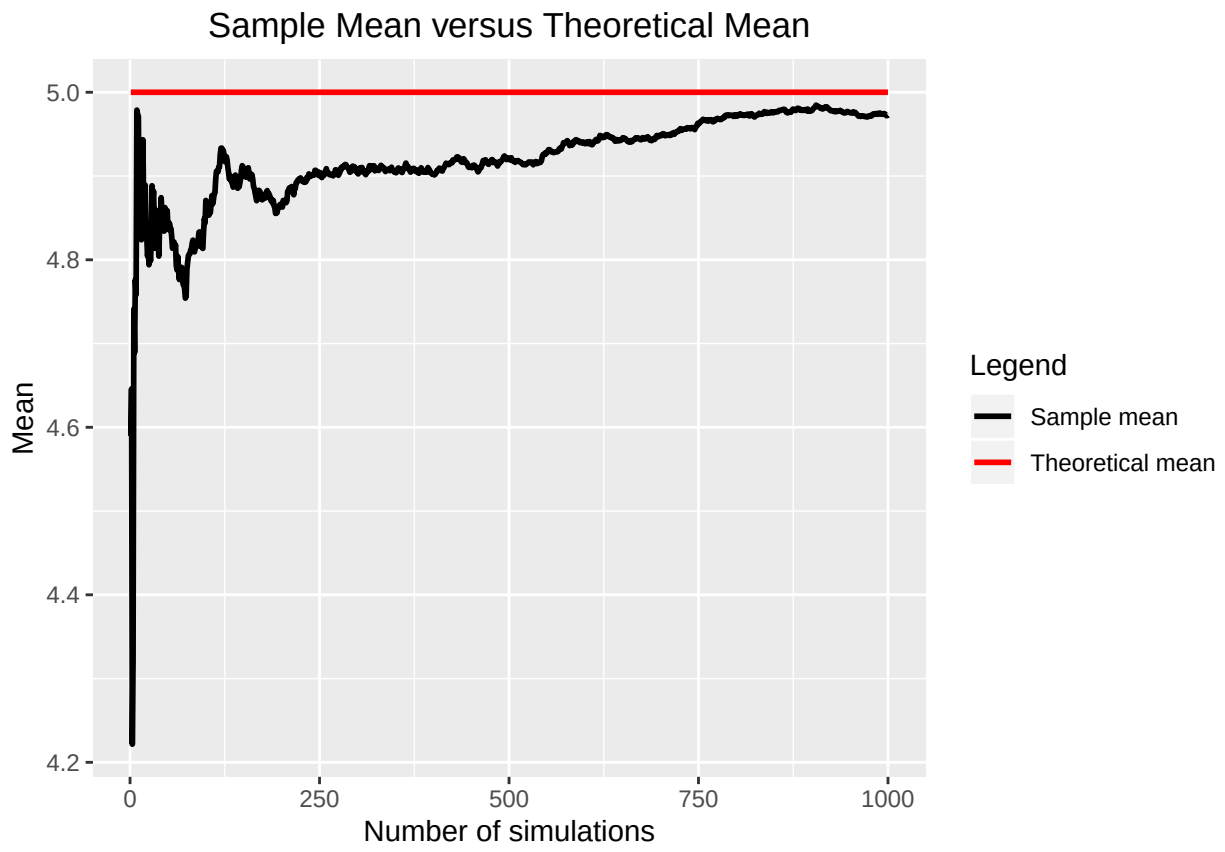
Finally, both the simulated statistics of the distribution of averages (mean and variance) and their theoretical values are calculated.

```
# Calculate the sample mean with progressively larger sample sizes
means <- cumsum(sample_avg) / (1:nosim)

# Calculate the final sample mean and its theoretical value
sample_mean <- mean(sample_avg)
theoretical_mean <- 1 / lambda

# Calculate the sample variance with progressively larger sample sizes
variances <- sapply(2:nosim, function(i) {sum((sample_avg[1:i] - means[i])^2)/(i-1)})

# Calculate the final sample variance and its theoretical value
sample_variance <- var(sample_avg)
theoretical_variance <- 1 / lambda^2 / n
```

Sample Mean versus Theoretical Mean

Below is a plot showing the sample mean (with progressively larger sample sizes) and its theoretical value (when the sample size tends to infinity).

```
ggplot(data.frame(x = rep(1:nosim, times = 2), y = c(means, rep(1/lambda, times = nosim)),
                  type = factor(rep(1:2, each = nosim))), aes(x = x, y = y, col = type)) +
  geom_line(size = 1) +
  scale_color_manual(name = "Legend", labels = c("Sample mean", "Theoretical mean"),
                     values = c("black", "red")) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Sample Mean versus Theoretical Mean",
       x = "Number of simulations", y = "Mean")
```



We can see from the plot that as the sample size increases, the sample mean converges to its theoretical value.

This behavior is expected, and confirms the validity of the Central Limit Theorem.

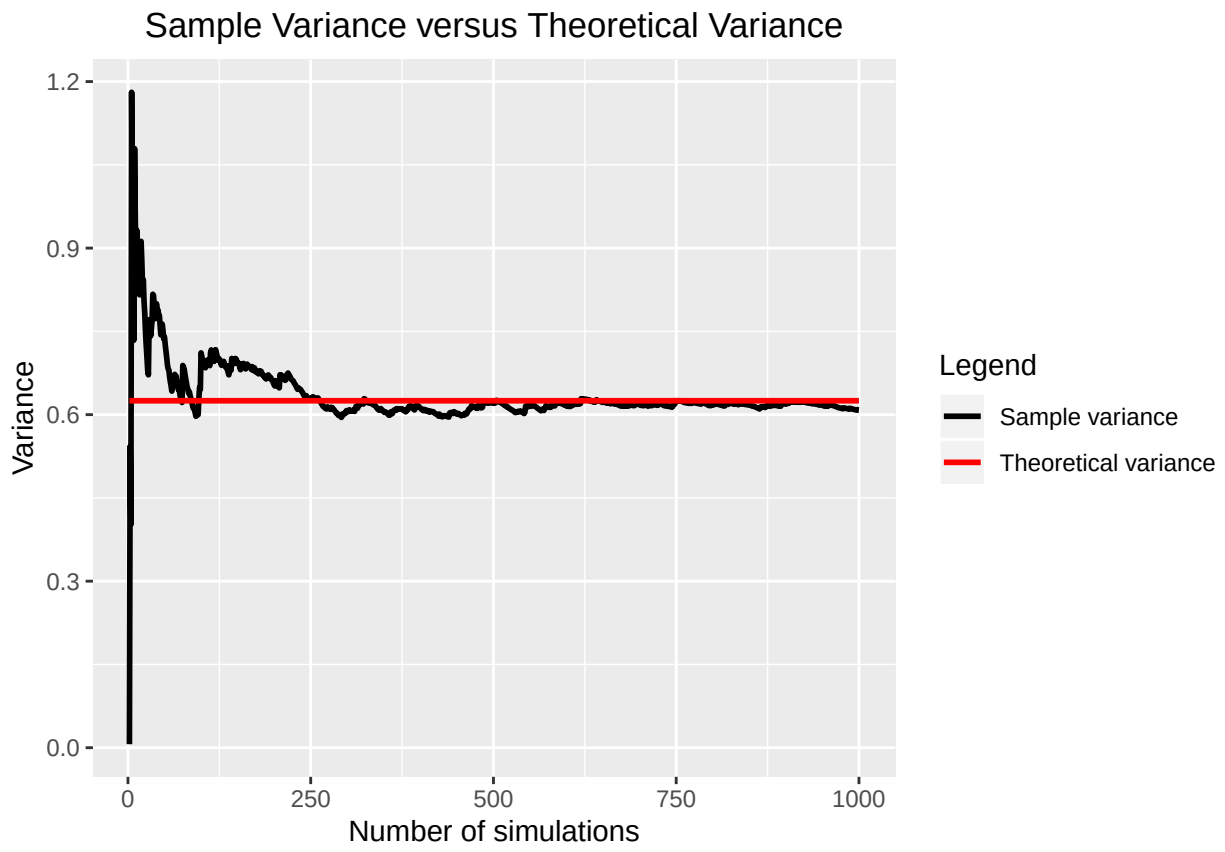
The final results, including all the 1000 simulations, are:

- sample mean: ***4.97***
- theoretical mean: ***5.00***

Sample Variance versus Theoretical Variance

Below is a plot showing the sample variance (with progressively larger sample sizes) and its theoretical value (when the sample size tends to infinity).

```
ggplot(data.frame(x = rep(2:nosim, times = 2), y = c(variances, rep((1/lambda)^2/n, times = nosim-1)),
                  type = factor(rep(1:2, each = nosim-1))), aes(x = x, y = y, col = type)) +
  geom_line(size = 1) +
  scale_color_manual(name = "Legend", labels = c("Sample variance", "Theoretical variance"),
                     values = c("black", "red")) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Sample Variance versus Theoretical Variance",
       x = "Number of simulations", y = "Variance")
```



Also in this case, we can see from the plot that as the sample size increases, the sample variance converges to its theoretical value.

This behavior is expected, and confirms the validity of the Central Limit Theorem.

The final results, including all the 1000 simulations, are:

- sample variance: ***0.608***
- theoretical variance: ***0.625***

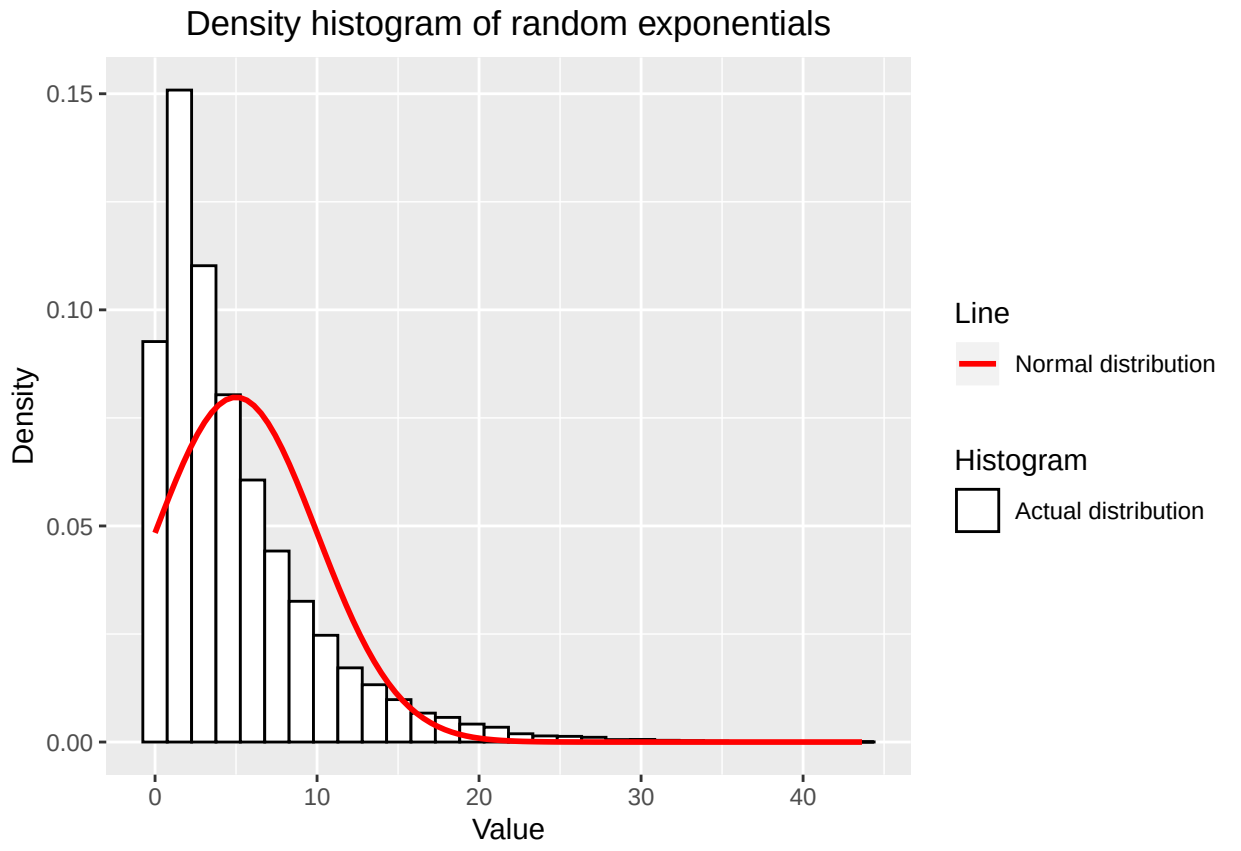
Distribution

To check the normality of the distribution, we can plot its density histogram and compare it with a normal distribution (the red line) having same mean and variance (here we used the theoretical values).

In particular to highlight even more the Central Limit Theorem, we show the difference between the distribution of a large collection of random exponentials (our entire sample) and the distribution of the averages of 40 exponentials.

In the first case, the original distribution is not normal (the histogram doesn't match the red line).

```
ggplot(mapping = aes(x = sample)) +
  geom_histogram(colour = "black", aes(y = ..density.., fill = "Actual distribution")) +
  stat_function(fun = dnorm, args = list(mean = 1/lambda, sd = 1/lambda), size = 1,
                aes(col = "Normal distribution")) +
  scale_fill_manual("Histogram", values = "white") +
  scale_color_manual("Line", values = "red") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Density histogram of random exponentials", x = "Value", y = "Density")
```



In the second case, instead, it is almost perfectly normal (the histrogram follows the red line very well).

```
ggplot(mapping = aes(x = sample_avg)) +
  geom_histogram(colour = "black", aes(y = ..density.., fill = "Actual distribution")) +
  stat_function(fun = dnorm, args = list(mean = 1/lambda, sd = 1/lambda/sqrt(n)), size = 1,
                aes(col = "Normal distribution")) +
  scale_fill_manual("Histogram", values = "white") +
```

```
  scale_color_manual("Line", values = "red") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Density histogram of average of 40 random exponentials", x = "Value", y = "Density")
```