Statistical Inferences Course Work - Part 1: Simulation Exercise

Ts. Nurul Haszeli Ahmad

11/22/2020

Overview

This is a simulation exercise done to comply with the requirement to complete the Statistical Inferences course. In this simulation exercise, we 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 the exponential distribution is 1/lambda, and the standard deviation is also 1/lambda. We will investigate the distribution of averages of 40 exponentials and a total of a thousand simulations.

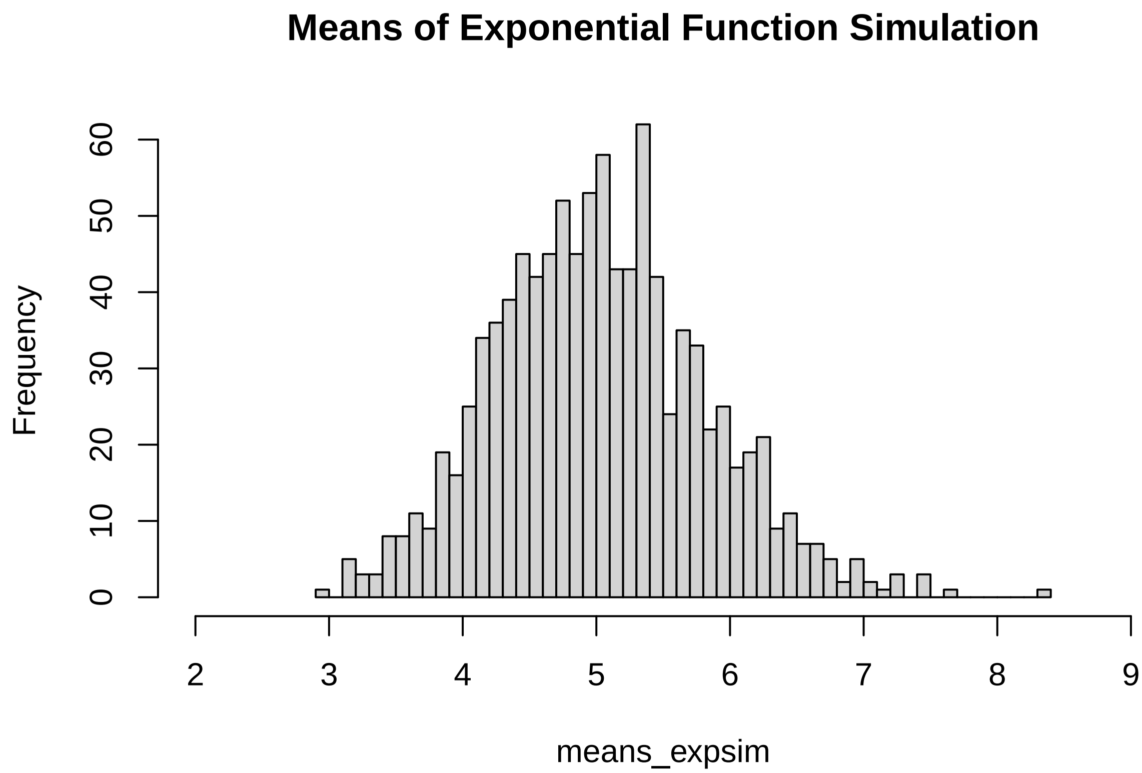
Initial setting

First, we set our seed and parameters based on the given set of data within this course work.

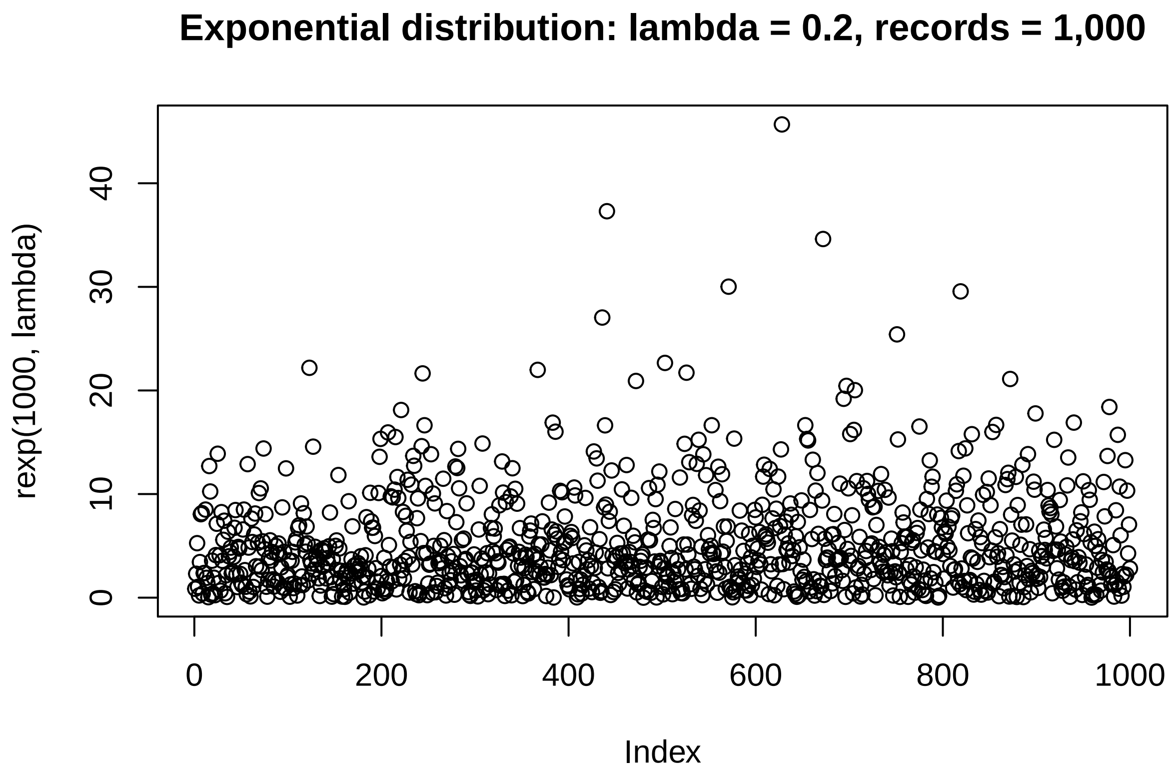
```
#set the seed
set.seed(1976)

#Set lambda = 0.2 for all of the simulations.
lambda <- 0.2

#Distribution of averages of 40 exponentials.
total_exponentials <- 40

#Total simulations is 1000
total_simulation <- 1000
```

Execute the simulation and plot into histogram and distribution

```
# run simulations
running_sim <- replicate(total_simulation, rexp(total_exponentials,lambda))

# Calculate means of exponential simulations
means_expsim <- apply(running_sim,2, mean)

# Plot means histogram
hist(means_expsim, breaks = total_exponentials, xlim = c(2,9),
     main = "Means of Exponential Function Simulation")
```




```r
# Exponential distribution can be simulated in R with rexp(n, lambda)
plot(rexp(1000,lambda),
     main="Exponential distribution: lambda = 0.2, records = 1,000")
```



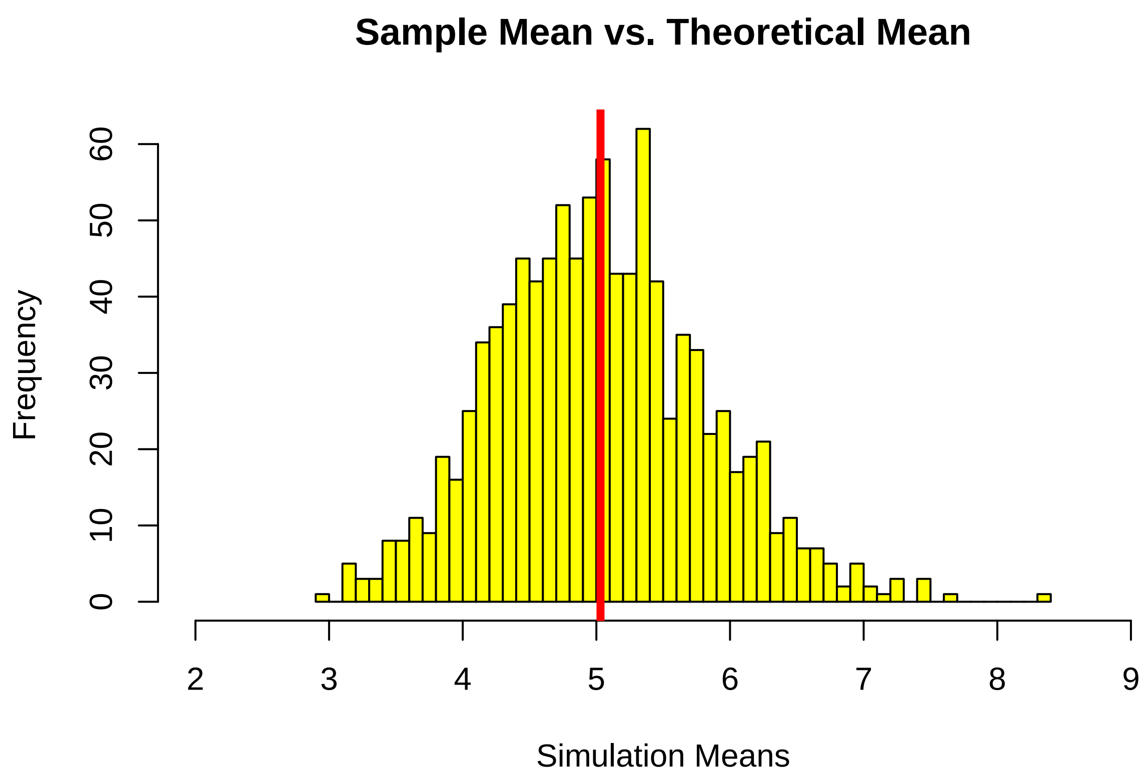
Note: The last plot does a quick plot to familiarize with the rexp. It is easy to see from plot that the majority of the values fall below 10, approximately around 5. There are no non-zero values so expect a mean around 5.

Question

Question 1: Sample Mean vs. Theoretical Mean

The exponential distribution mean is 1/lambda. For this simulation, lambda is 0.2. The theoretical mean should be 5 (i.e., 1 / 0.2).

```
# plot histogram of the sample means
hist(means_expsim, main = "Sample Mean vs. Theoretical Mean", xlim =  c(2,9),
     breaks = total_exponentials, xlab = "Simulation Means", col = "yellow")

# plot vertical blue line at mean of samples
abline(v = mean(means_expsim), lwd = "4", col = "red")
```




```r
# calculate the mean of our samples
mean(means_expsim)
```

```
## [1] 5.030864
```

Result: Based on the plot and calculation, the sample mean is **5.02**; of which it is very close to our theoretical mean of **5**.

Question 2: Sample Variance vs. Theoretical Variance

The standard deviation of the exponential distribution is (1/lambda) / sqrt(n). We will compare this to our simulations.

```r
# theoretical variance vs. simulated variance
print(paste ("Theoretical variance is: ", round( (1/lambda)^2/total_exponentials, 3)))
```

```
## [1] "Theoretical variance is:  0.625"
```

```r
print(paste("Actual variance is: ", round( var(means_expsim),3)))
```

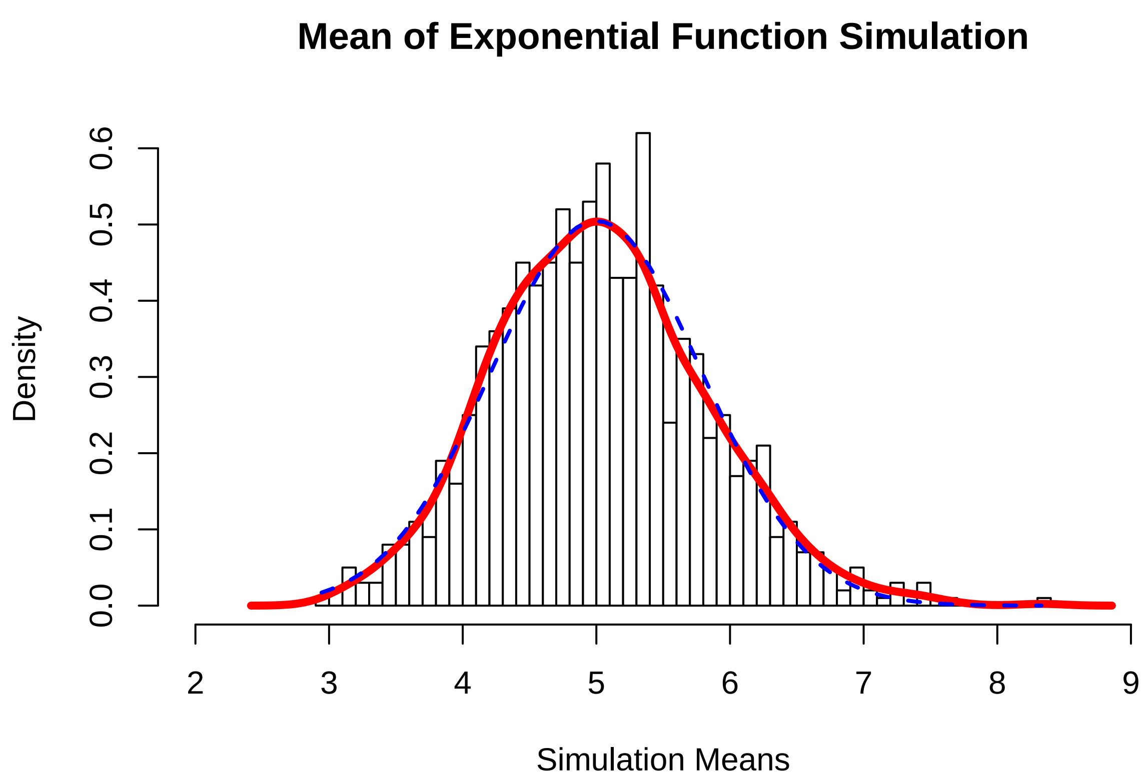
```
## [1] "Actual variance is:  0.625"
```

Result: Based on the above result, both variances are close to each other indicating that the theoretical can be considered giving a true value of an actual result.

Question 3: Distribution

The following simulate whether the exponential distribution is approximately normal. Due to the Central Limit Theorem, the means of the sample simulations should follow a normal distribution.

```
# Histogram with distribution curve included
hist(means_expsim, prob=TRUE, main = "Mean of Exponential Function Simulation",
     breaks = 40, xlim = c(2,9), xlab = "Simulation Means", col = "white")
lines(density(means_expsim), lwd=4, col="red")

# Normal distribution line
x <- seq(min(means_expsim), max(means_expsim), length = 2*total_exponentials)
y <- dnorm(x, mean = 1/lambda, sd = sqrt(((1/lambda)/sqrt(total_exponentials))^2))
lines(x,y, pch = 20, lwd = 2, lty = 2, col="blue")
```



Conclusion

As shown on the last plot, the distribution of means of the simulated exponential distributions follows a normal distribution due to the Central Limit Theorem. The higher the number of samples increase (currently at 1000), the closer the simulated distribution should get to the standard normal distribution (the dotted blue line, above). The red line is the simulated curve.