Simulations on Exponential Distribution with Comparison to Central Limit Theorem

Coursera Statistical Inference Course Project I

Xinyu W

7/25/2020

Overview

This project will investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). The investigation include 1,000 simulations on the distribution of averages of 40 exponentials. Essentially, the project is going to work around the sample/theoretical mean, sample/theoretical variance and the distribution, which at last we are going to see if it is normal.

Simulations

Before work: set environment and variables

We will set the echo to be true to show our code in the final report and turn off the warnings. Next, we assign our default values in this project to variables.

```
library(knitr)
opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
```

```
set.seed(234) #insure reproducibility
n <- 40 #number of exponentials
lambda <- .2 #default lambda values
sim <- 1000 #number of simulations
z_value <- 1.96 #95% confidence interval
```

Creating simulations

Here, we create a matrix with 40 columns by 1000 rows to store the data of our exponential simulation data.

```
#create a matrix to hold up data for exponential distribution
data <- matrix(rexp(n*sim, rate = lambda), sim)
```

Sample Mean Versus Theoretical Mean

After we calculate the means for each of our simulation, we can get our sample mean by further taking the mean of those means. Since we know our default lambda and the way to get our theoretical mean (1/lambda). We can first do comparison simply by print those numbers.

```r
#calculate the mean through each row and assign all means to simMn
simMn <- rowMeans(data)
#the mean of the simMn is what we want to get for comparison
MnMean <- mean(simMn)
paste("The sample mean of our simulations is", round(MnMean, 5))
```

```
## [1] "The sample mean of our simulations is 5.00157"
```

```r
#For comparison, we print our theoretical mean
theoMean <- 1/lambda
paste("By comparison, our theoretical mean is", theoMean)
```

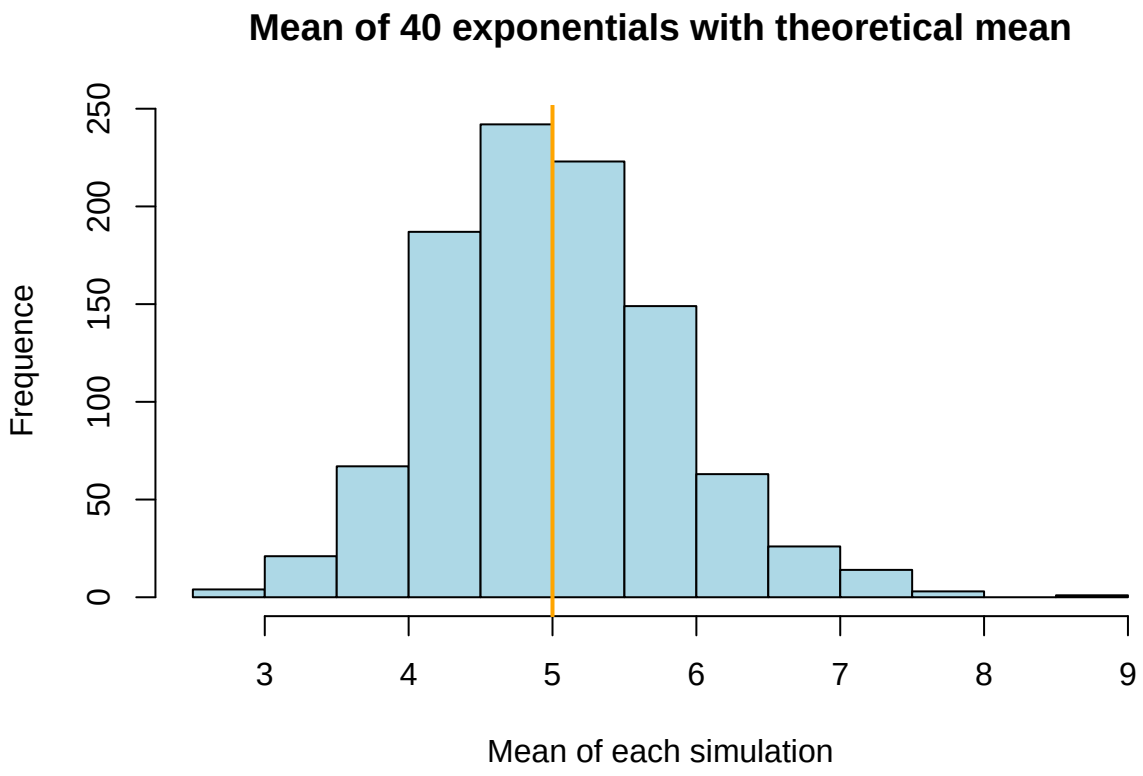
```
## [1] "By comparison, our theoretical mean is 5"
```

```r
paste("The difference is", round(MnMean - theoMean, 5))
```

```
## [1] "The difference is 0.00157"
```

Then, we can have a plot to look better. We can see the distribution of our sample means and where the theoretical mean lies (the red line).

```r
hist(simMn, xlab = "Mean of each simulation", ylab = "Frequence",
     col = "lightblue", breaks = 20,
     main = "Mean of 40 exponentials with theoretical mean")
abline(v=theoMean, col = "orange", lwd = 2)
```



We can see from text and plot that our sample mean and theoretical mean are very close (0.00157). The theoretical mean lies in the middle area of the sample mean distribution (where there are most number of means).

Sample Variance Versus Theoretical Variance

We can first simply calculate the sample variance (from the means of our 1000 simulations) and theoretical variance, since we have already known that the standard deviation of the mean of exponential distribution is *1/lambda*.

```
#calculate the variance of all averages
variance <- var(simMn)

#calculate the theoretical variance
theoVar <- (1/lambda)^2/n

#We paste those two variance for comparison
paste("The sample variance of our simulations is", round(variance,5))
```

```
## [1] "The sample variance of our simulations is 0.66315"
```

```
paste("By comparison, the theoretical variance of our simulations is", theoVar)
```

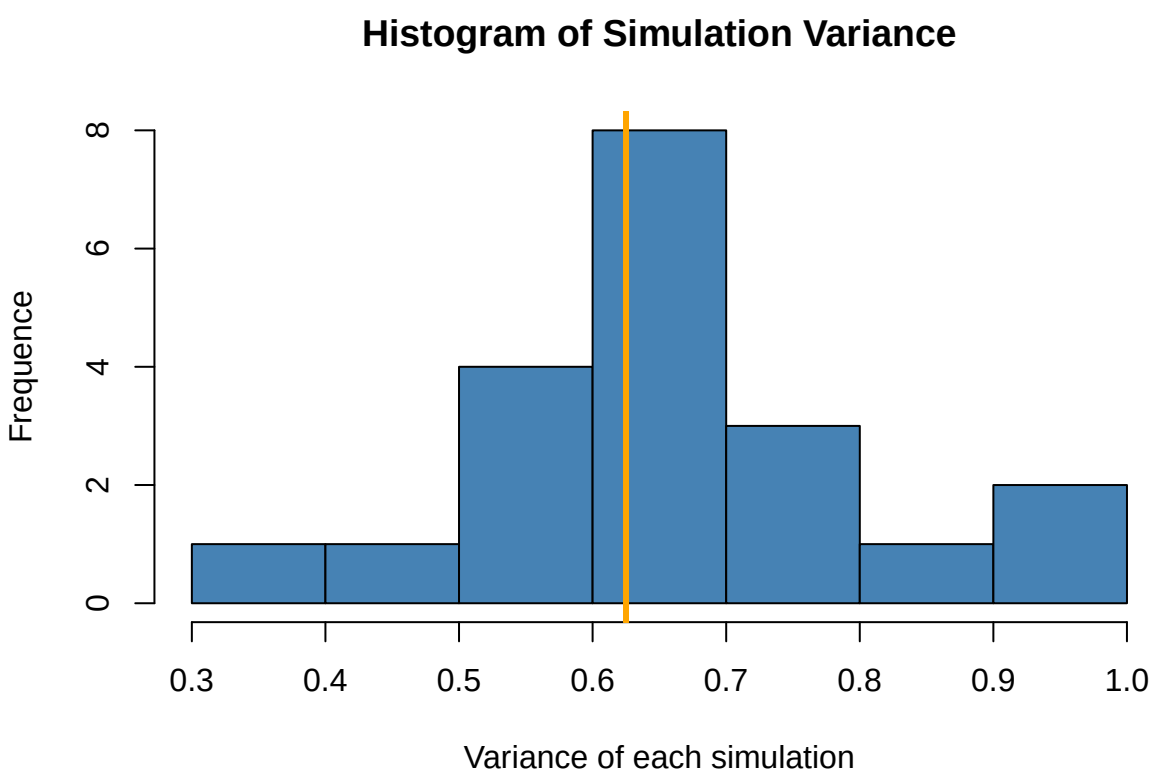
```
## [1] "By comparison, the theoretical variance of our simulations is 0.625"
```

```
paste("The difference is", round(variance - theoVar, 5))
```

```
## [1] "The difference is 0.03815"
```

To generate the plot, I grouped every 50 simulations of the total 1000 simulations, calculated the variance of each group, and created *simVar* to store every variance. By doing this, I can see some changes and distributions of variance from different part of our data. Then as I plot the grouped variance, I add a line for theoretical variance to see where it lies. As a result, it concentrated in where there are most variance(s). In comparison, our sample variance is also between 0.6 and 0.7, which makes sense.

```
#calculate the variance of each simulation
simVar <- c()
for (i in seq(from=1, to = 1000, by = 50)) {
        temp <- var(rowMeans(data[i:(i+49), ]))
        simVar <- c(simVar, temp)
}
hist(simVar, xlab = "Variance of each simulation", ylab = "Frequence",
     col = "steelblue",
     main = "Histogram of Simulation Variance")
abline(v = theoVar, lwd = 3, col = "orange")
```

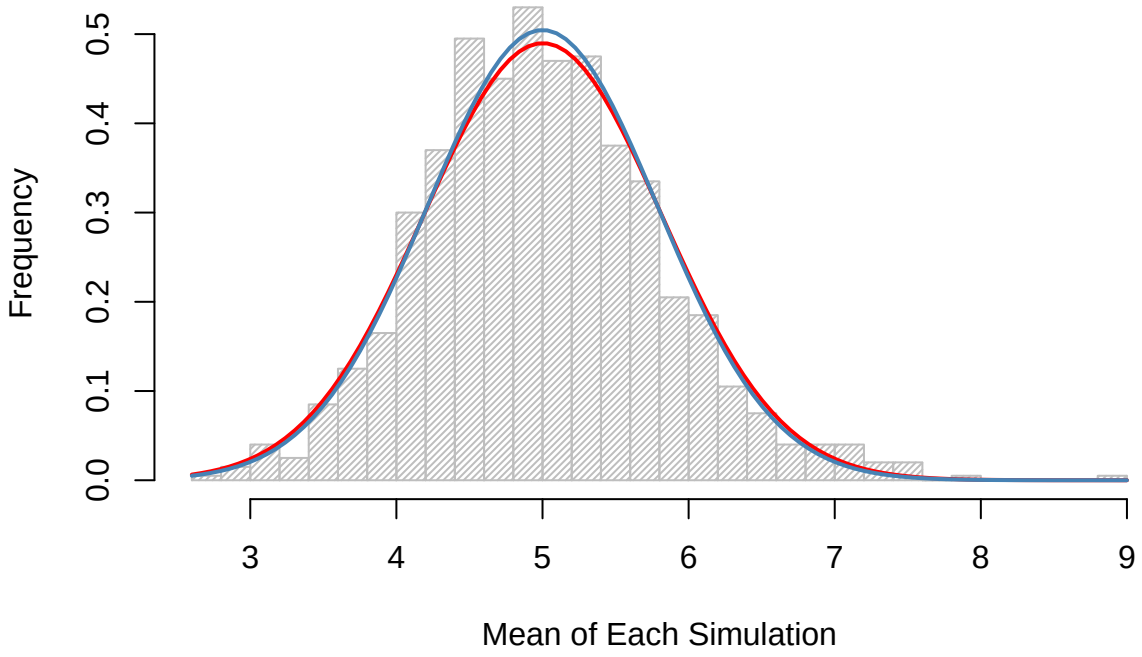


Distribution

Finally, we came to this section which shows you the distribution of the means of all our simulations. We also add two lines to illustrate the distribution of sample mean/variance and theoretical mean/variance.

Here, the red line represents the sample data, while the blue line represents the theoretical data. We find that the distribution is close to a normal distribution.

```
hist(simMn, xlab = "Mean of Each Simulation",
     ylab = "Frequency", breaks = 40,
     probability = TRUE, density = 40, col = "grey",
     main = "Distribution of Means of 40 exponentials")
curve(dnorm(x, mean = MnMean, sd = sqrt(variance)), col = "red", lwd = 2, add = TRUE)
curve(dnorm(x, mean = theoMean, sd = sqrt(theoVar)), col = "steelblue", lwd = 2, add = TRUE)
```



Then we calculate the confidence interval for both sample and theoretical data and do comparison between them. In all, they are pretty close by a difference within the absolute value of 0.02. Here, we can be pretty sure the distribution is normal.

```
#The CI value for sample
simCI <- MnMean + c(-1,1)*z_value*variance/sqrt(n)
#The CI value for theory
theoCI <- theoMean + c(-1,1)*z_value*theoVar/sqrt(n)
#calculate the difference
difference <- simCI-theoCI

simCI
```

```
## [1] 4.796060 5.207085
```

```
theoCI
```

```
## [1] 4.80631 5.19369
```

```
difference
```

```
## [1] -0.01025009  0.01339579
```