Statistical Inference - Week 4 Course Project - Part 1

Sanjay Lonkar

14 Jun 2018

A simulation exercise

Overview

This assignment is Part 1 of Statistical Inference - Week 4 course project. Goal of this assignment is to perform A simulation exercise with following objectives:

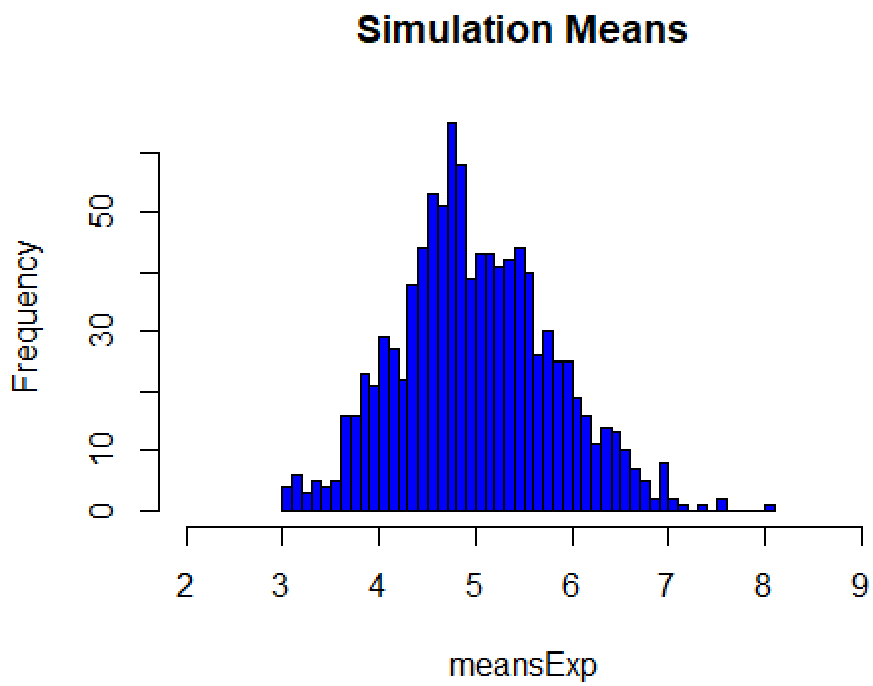
In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp (n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation 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

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.
4. In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Set the given constants in R variables

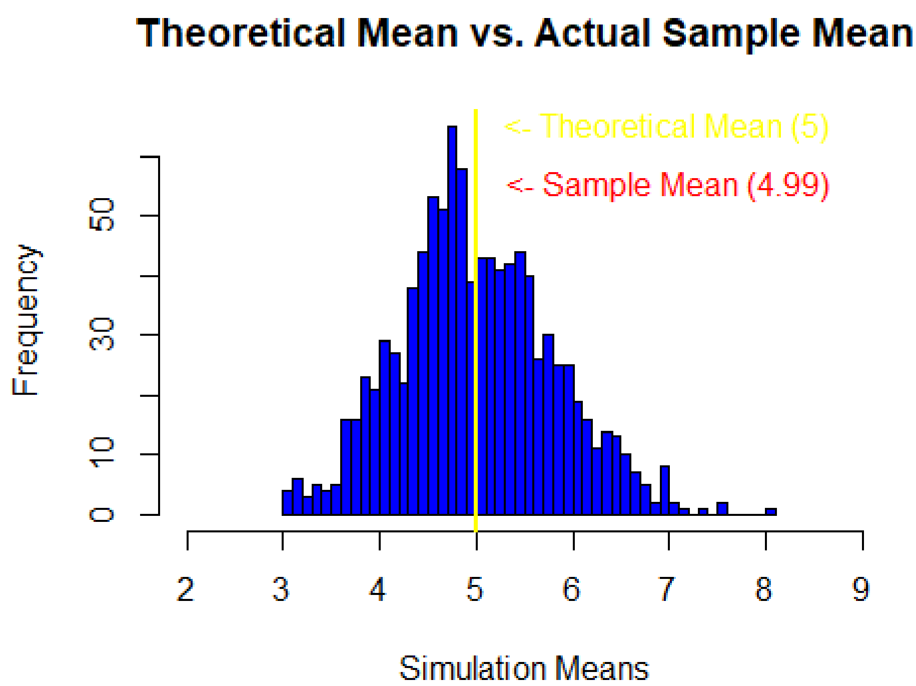
```
set.seed(100)
lambda <- 0.2
n <- 40
simulations <- 1000

# Replicate to prepare simulations and calculate their mean
simulationsExp <- replicate (simulations, rexp(n, lambda))
meansExp <- apply (simulationsExp, 2, mean)

# Plot histogram
hist(meansExp, breaks = 40, xlim = c(2, 9), main = "Simulation Means", col =
"blue")
```



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

```
sampleMean <- mean(meansExp)
theoreticalMean <- 1 / lambda

print(paste("Sample Mean of simulated data is", sampleMean, " while
Theoretical Mean of same data is", theoreticalMean))

## [1] "Sample Mean of simulated data is 4.9997019268744  while Theoretical
Mean of same data is 5"

print(paste("Difference between these two is just", (theoreticalMean -
sampleMean), "It indicates that Sample Mean and Theoretical Means are preety
close."))

## [1] "Difference between these two is just 0.000298073125599707 It
indicates that Sample Mean and Theoretical Means are preety close."

# Show above in a plot
hist (meansExp, breaks = 40, xlim = c (2, 9), main = "Theoretical Mean vs.
Actual Sample Mean", xlab = "Simulation Means", col="blue")

# plot a vertical red line at the mean of the sample means
abline (v = sampleMean, lwd = "2", col = "red")
abline (v = theoreticalMean, lwd = "2", col = "yellow")
```

```r
text (x = 7, y = 55, "<- Sample Mean (4.99)", col = "red", lwd = 4)
text (x = 7, y = 65, "<- Theoretical Mean (5)", col = "yellow", lwd = 4)
```



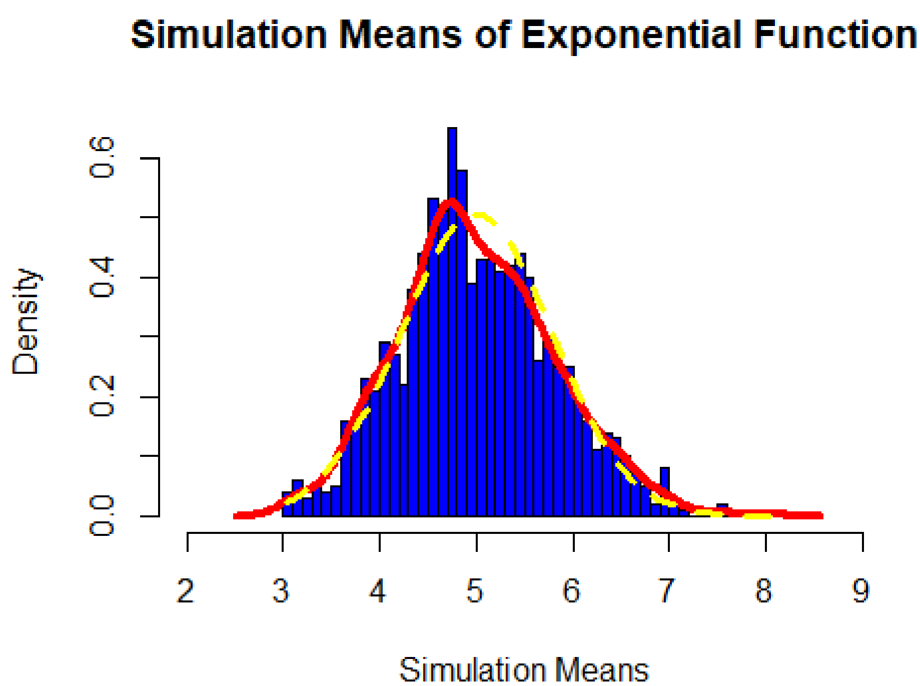
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution

```r
theoreticalVar <- round (((1/lambda)/sqrt(n))^2 ,4)
simulatedVar <- round (sd(meansExp)^2 ,4)

print(paste("Simulated variance of this data is", simulatedVar, " while
Theoretical variance of same data is", theoreticalVar))

## [1] "Simulated variance of this data is 0.6432  while Theoretical variance
of same data is 0.625"

print(paste("Difference between these two is just", (theoreticalVar -
simulatedVar), "It indicates that Sample Variance and Theoretical Variance
are preety close."))

## [1] "Difference between these two is just -0.0182 It indicates that Sample
Variance and Theoretical Variance are preety close."
```

3. Show that the distribution is approximately normal

```r
#General Plot with ditribution curve drawn
hist (meansExp, prob = TRUE, col = "blue", main="Simulation Means of
Exponential Function", breaks = 40, xlim = c (2, 9), xlab = "Simulation
Means")
```

```
lines (density (meansExp), lwd = 4, col = "red")

# Normal distribution line creation
x <- seq (min (meansExp), max (meansExp), length =2*n)
y <- dnorm (x, mean = 1 / lambda, sd = sqrt (((1/lambda)/sqrt(n))^2))
lines(x, y, pch = 22, col = "yellow", lwd = 3, lty = 2)
```




```
print("Sample curve is red line and yellow dotted line is normal
distribution. Both these are very close as one can see in the plot above.")

## [1] "Sample curve is red line and yellow dotted line is normal
distribution. Both these are very close as one can see in the plot above."
```

4. In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

If we increase number of samples from current 1000 to a higher number, the distribution will be very close to standard normal distribution.