Data Science: Statistical Inference - Project Part 1 (Exponential Distribution Simulation)

Jagannatha Reddy

July 25, 2016

Problem Description

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.

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.

Synopsis

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated in R using rexp(n, lambda) method where lambda is the rate parameter.

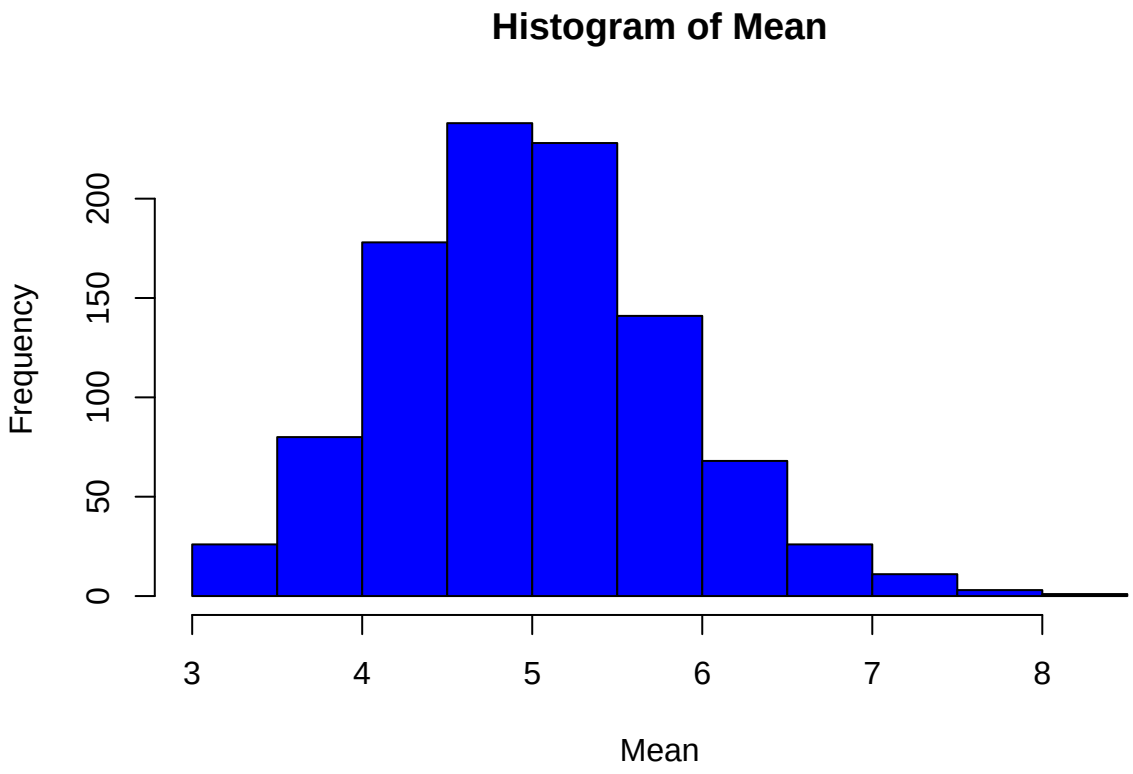
Simulation Data Preparation

```
echo = TRUE
cache = TRUE

numSims <- 1000  # number of simulations
n <- 40          # analysis for 40 exponentials as per problem statement
lambda <- 0.2    # setting lambda as per problem statement

set.seed(1000)    # set the seed to reproduce the same results

# create a matrix of 1000 rows with each row representing 40 random simulations
simulationMatrix <- matrix(rexp(numSims*n, rate=lambda), numSims, n)
```

1) How sample mean compares to the theoretical mean?

Here we show how the sample mean for this simulation data is compared to the theoretical mean of the distribution. The theoretical mean of exponential distribution is 1/lambda

```r
simulationMean   <- rowMeans(simulationMatrix)
hist(simulationMean, main="Histogram of Mean", xlab="Mean", col = "blue")
```




```r
sampleMean      <- round(mean(simulationMean), 3)  # compute the sample mean
theoreticalMean <- round(1/lambda, 3)              # theoretical mean is 1/lambda
print(paste("Simulation mean: ", sampleMean))
```

```
## [1] "Simulation mean:  4.987"
```

```r
print(paste("Theoretical mean: ", theoreticalMean))
```

```
## [1] "Theoretical mean:  5"
```

You can notice that the sample mean of the simulation data is *4.987* which is very close to that of theoretical mean of the distribution which is *5.000*

2) How variance of the sample data compares to the theoretical variance?

Here we show how the variance of the sample data for this simulation is compared to the theoretical variance of the distribution. The theoretical variance of exponential distributions is (1/lambda)^2/n

```r
sampleVar      <- round(var(simulationMean), 3)  # compute the variance of the simulation data
theoreticalVar <- round((1/lambda)^2/n, 3)       # theoretical variance is (1/lambda)^2/n
print(paste("Simulation variance: ", sampleVar))
```

```
## [1] "Simulation variance:  0.658"
```

```r
print(paste("Theoretical variance: ", theoreticalVar))
```

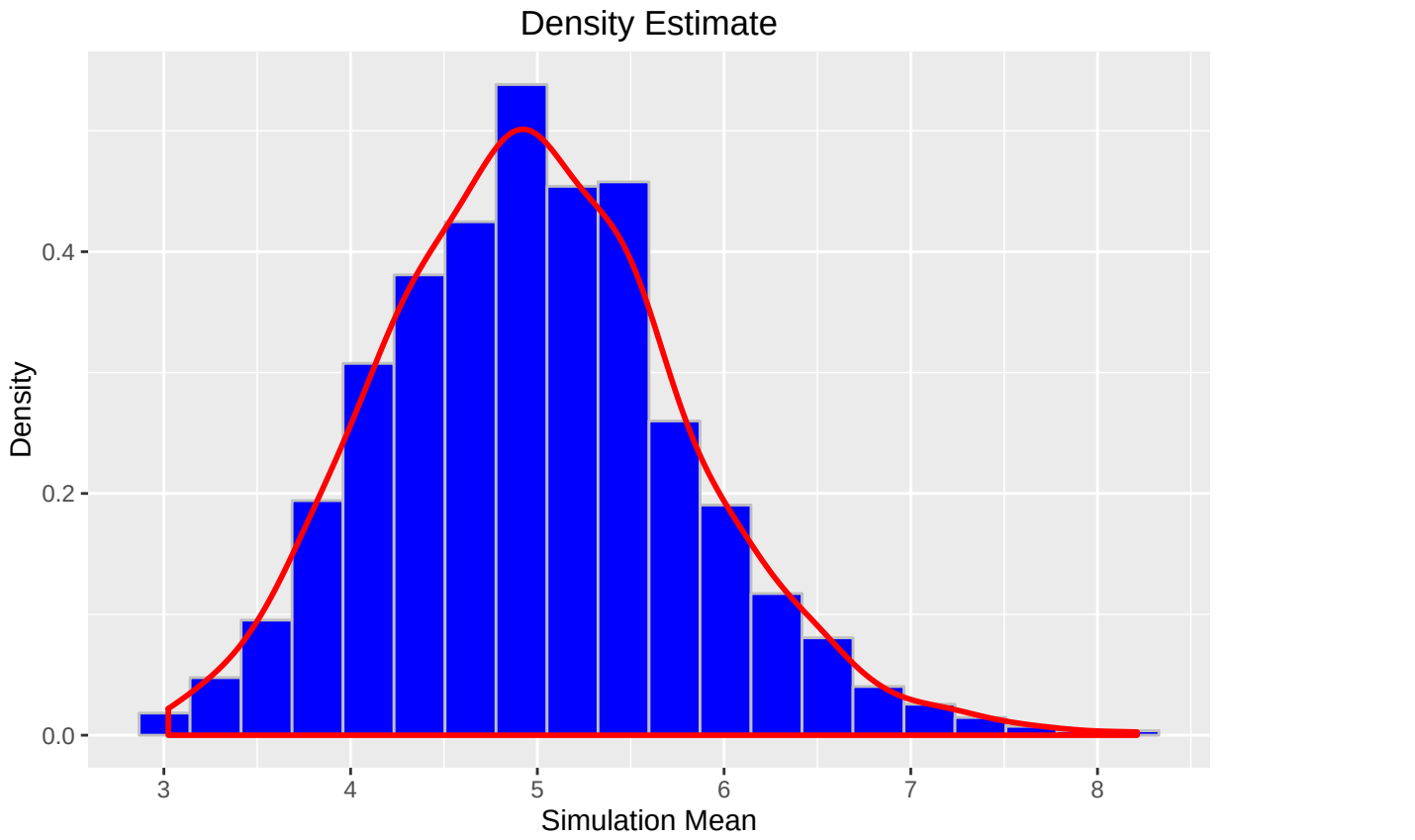
```
## [1] "Theoretical variance:  0.625"
```

You can notice that the variance of the simulation data is **0.658** which is very close to that of theoretical variance of the distribution which is **0.625**

3) Show that the distribution is approximately normal

Here we show that distribution is approximately normal

```r
library(ggplot2) # load ggplot2 library
g<-ggplot(data.frame(simulationMean), aes(x=simulationMean))
g<-g+geom_histogram(aes(y=..density..), colour="grey", fill = "blue", bins=20)
g+geom_density(colour="red", size=1) +labs(x="Simulation Mean", y="Density", title="Density Estimate")
```



By looking at the smooth density estimate we observe that curve closely follows the histogram data

Let us also derive the confidence internals for the simulation data and theoretical exponential distribution data

```r
sampleSD <- sqrt(sampleVar)
round(sampleMean+c(-1,1)*qnorm(0.975)*sampleSD/sqrt(n), 3)          # confidence interval
```

```
## [1] 4.736 5.238
```

```r
theoreticalSD <- theoreticalVar
round(theoreticalMean+c(-1,1)*qnorm(0.975)*theoreticalSD/sqrt(n), 3)  # confidence internal
```

```
## [1] 4.806 5.194
```

You can observe that the confidence interval of the simulation data is [**4.736, 5.238**] which is very close to that of confidence interval of the theoretical distribution which is [**4.806, 5.194**]