SI Course Project (Part 1)

Cecilia Cruz-Ram, MD DPCOM

4/2/2018

Introduction:

In this project 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 exponential distribution is 1/lambda and the standard deviation is also 1/lambda.

Provided Data:

a. lambda = 0.2 for all of the simulations

b. Distribution of averages of 40 exponentials

c. Number of simulations = 1000

Simulations:

A. Setwd

```
setwd("/Users/sexybaboy/Documents/Files/Zetch/Online Courses/Data Science Specialization Feb18/R/Statistical I
```

B. Set seed of reproducibility

```
set.seed(2018)
```

C. Set sample size

```
n <- 40              # sample size
lambda <- 0.2        # number of exponentials
simNum <- 1000       # number of simulations
```

D. Run simulations

```
simMeans = NULL
for (i in 1 : 1000) {
  simMeans = c(simMeans, mean(rexp(n,lambda)))
}

head(simMeans)
```

```
## [1] 4.851478 5.734640 4.144173 4.848036 4.766554 5.062200
```

E. Instructions

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

Run theoretical mean

```
theosampMean <- round(1/lambda,3)
theosampMean
```

```
## [1] 5
```

Run actual mean

```
actualMean <- round(mean(simMeans),3)
actualMean
```

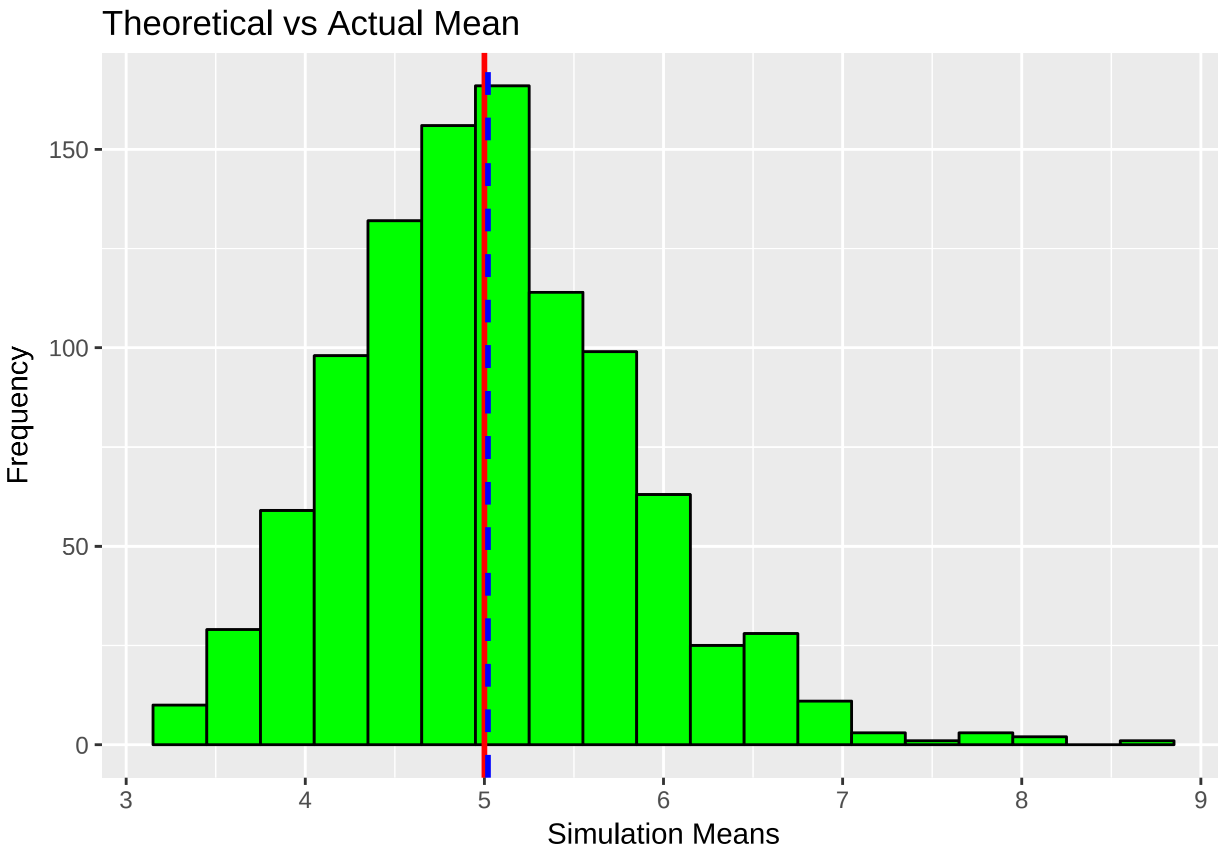
```
## [1] 5.02
```

Plot showing both means

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
simMeansDf <- as.data.frame(simMeans)
g <- ggplot(simMeansDf, aes(x = simMeans))
g <- g + geom_histogram(binwidth = .3, color = "black", fill = "green") +
  geom_vline(xintercept = theosampMean, color = "red", size = 1, linetype = 1) +
  geom_vline(xintercept = actualMean, color = "blue", size = 1, linetype = 2) +
  labs(x = "Simulation Means", y = "Frequency",
  title = "Theoretical vs Actual Mean")
g
```



The red dashed vertical line indicate the theoretical sample mean, 1/lambda = 5, while the green dashed vertical line is the calculated average sample mean size of 40 of 1000 samples showing very close proximity.

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

Run theoretical variance

```r
theosampVar <- round((1/lambda)^2/n,3)
theosampVar
```

```
## [1] 0.625
```

Run actual variance

```r
actualVar <- round(var(simMeans),3)
actualVar
```

```
## [1] 0.626
```

Table showing both theoretical and actual mean and variance

```r
tab <- matrix(c(theosampMean, actualMean,
                theosampVar,actualVar),
```

```
              ncol = 2, byrow = TRUE)
colnames(tab) <- c("Theoretical","Sample")
rownames(tab) <- c("Mean","Variance")
tab <- as.table(tab)
tab
```

```
##          Theoretical Sample
## Mean           5.000  5.020
## Variance       0.625  0.626
```

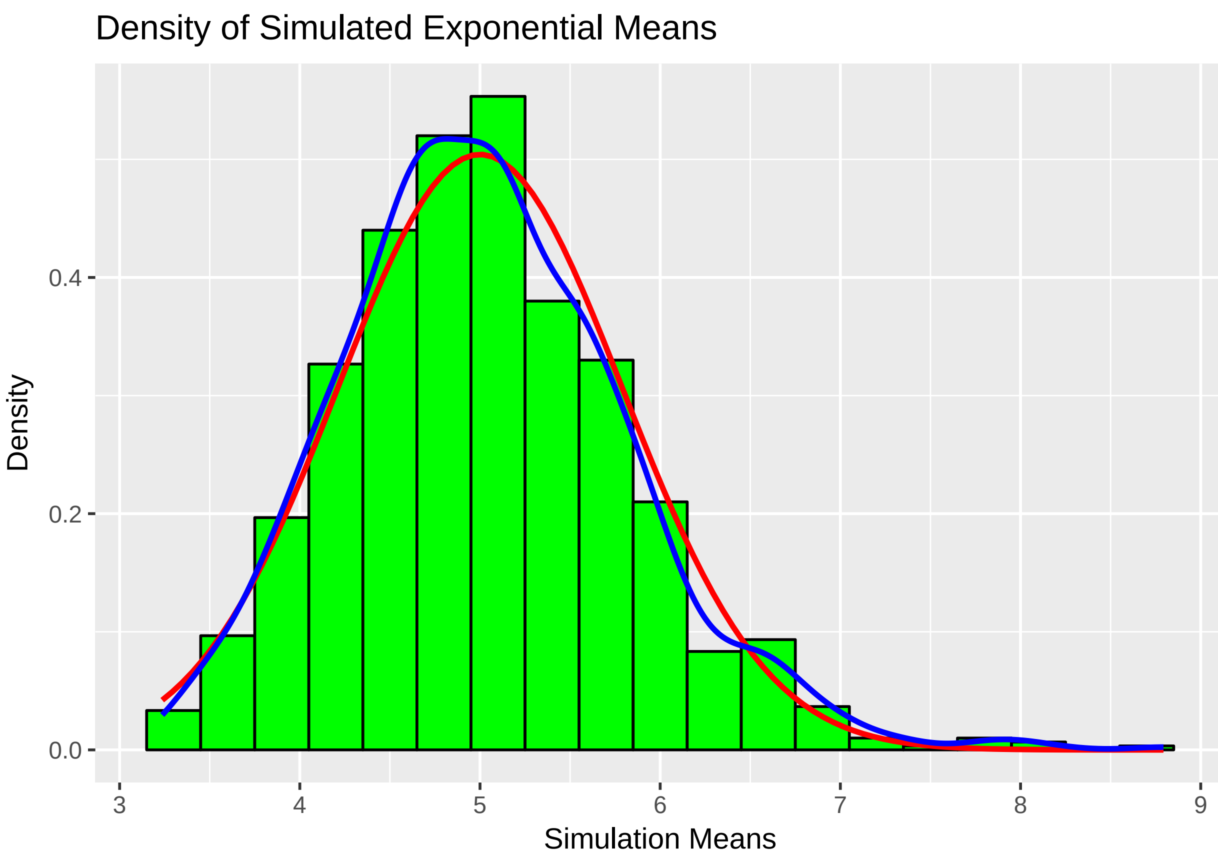
Preceding table shows near identical values.

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.

Make a histogram with the density and sample means. Add density curve of the normal distribution and the sample distribution:

```
g <- ggplot(simMeansDf, aes(x=simMeans))
g <- g + geom_histogram(binwidth = .3, color = "black", fill = "green" , aes(y = ..density..)) +
  stat_function(fun = dnorm, args = list(mean = theosampMean, sd = sd(simMeans)),
                color = "red", size = 1) +
  stat_density(geom = "line", color = "blue", size = 1)  +
  labs(x = "Simulation Means", y = "Density",
       title = "Density of Simulated Exponential Means")
g
```



Plot2 shows the distribution of means of the sampled exponential distributions which appear to follow a normal distribution, due to the Central Limit Theorem. An increase in the number of samples (currently 1000) will create a distribution that would be even closer to the standard normal distribution. The red line above is the normal distribution curve which closely approximates the blue colored sample curve.