Statistical Inference Course Project Part 1 - Simple inferential data analysis

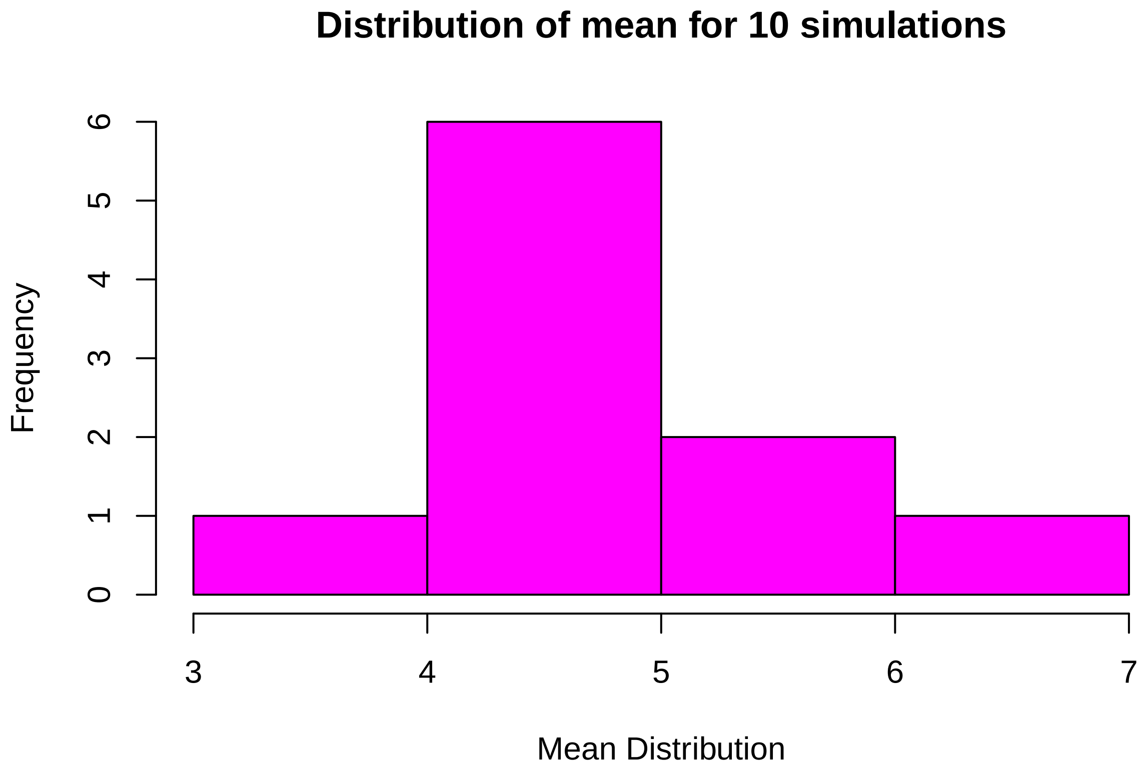
Author: Ramesh Natarajan

Overview

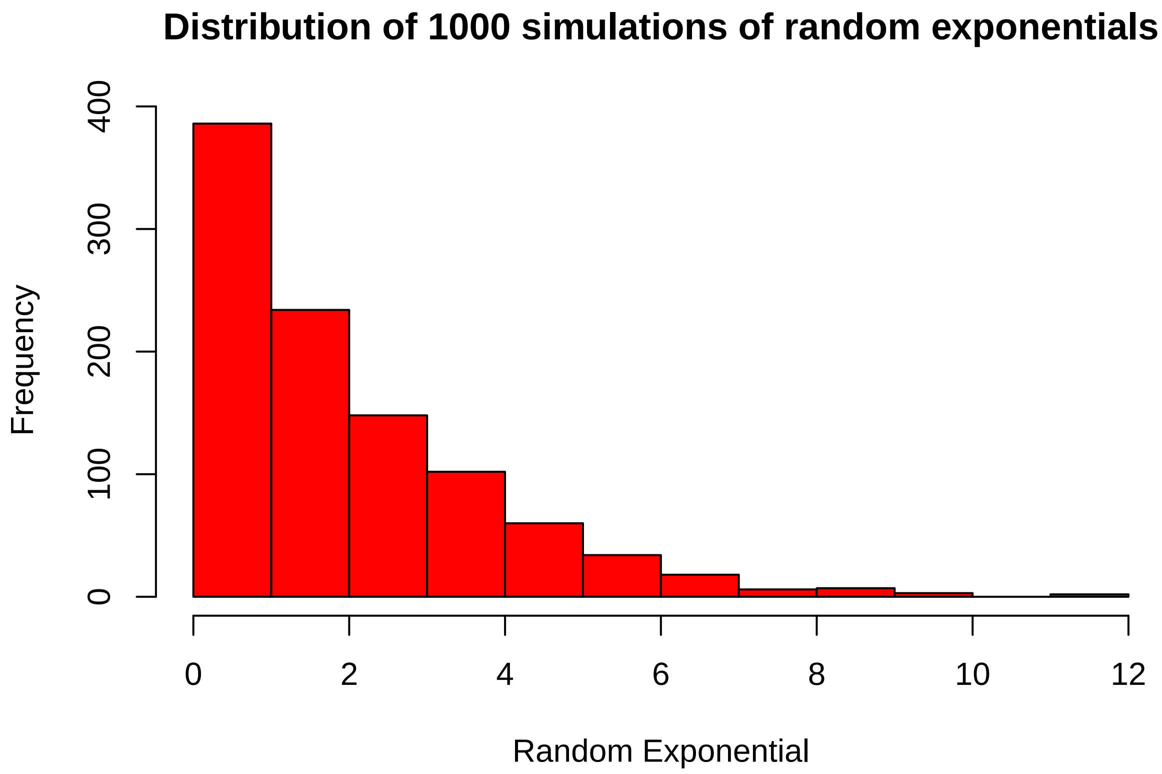
In this project, we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. We will set lambda – the rate parameter as 0.2 and do a the simulations for the analysis.

The objective of this exercise is to illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. We will: 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.

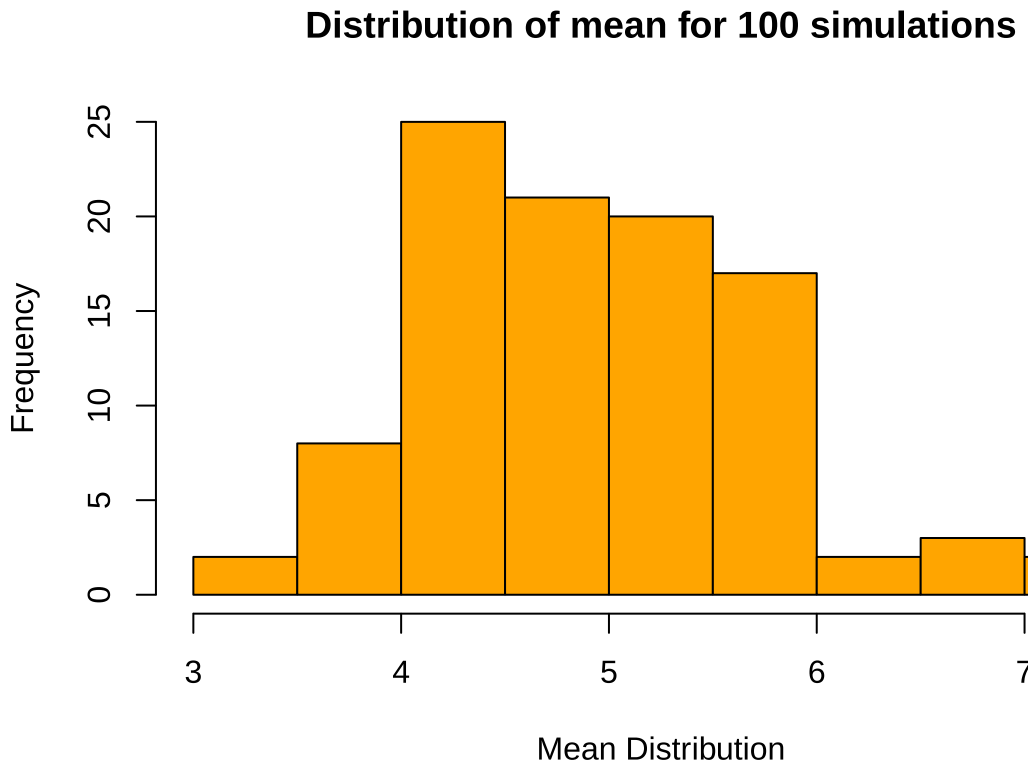
```
# Let's create a dataframe of 10, 100 and 1000 simulations of sample means and
# explore the properties of the distribution formed by these means
# (which will be our random variables)

rmeans10 = data.frame(X = sapply(1:10, function(X) {mean(rexp(40, 0.2))}))
hist(rmeans10$X, col="magenta", main="Distribution of mean for 10 simulations",
     xlab="Mean Distribution", ylab = "Frequency")
```



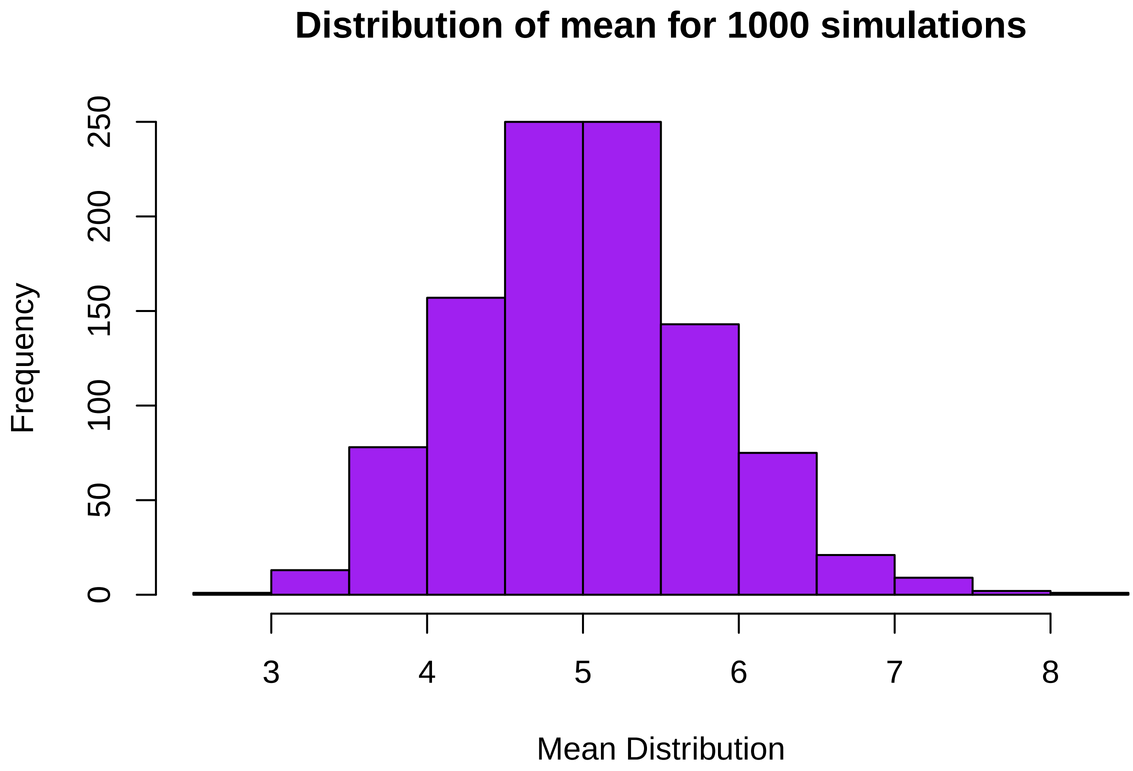
```r
hist(rexp(1000, 0.5), col="red", main="Distribution of 1000 simulations of random exponentials",
     xlab="Random Exponential", ylab = "Frequency")
```



```r
rmeans100 = data.frame(X = sapply(1:100 , function(X) {mean(rexp(40, 0.2))}))
hist(rmeans100$X, col="orange", main="Distribution of mean for 100 simulations",
     xlab="Mean Distribution", ylab = "Frequency")
```



```
rmeans1000 = data.frame(X = sapply(1:1000, function(X) {mean(rexp(40, 0.2))}))
hist(rmeans1000$X, col="purple", main="Distribution of mean for 1000 simulations",
     xlab="Mean Distribution", ylab = "Frequency")
```



```
# Let's view some records
head(rmeans1000)
```

```
##          X
## 1 5.086949
## 2 4.912822
## 3 6.030887
## 4 5.440128
## 5 4.663961
## 6 6.428544
```

Notice that the distribution of 1000 generated random exponentials is skewed and unbalanced compared to the distribution of the 100 averages of random expoentials

Also notice that the distribution of means of 1000 generated random exponentials is more Gaussian compared to the distribution of the 100 generated random exponentials supporting the CLT.

Question 1

```
# Theoretical Mean of the distribution is 1/lambda = 5
# Mean of the the random variables of means (rmeans1000)

mean(rmeans1000$X)
```

```
## [1] 5.021594
```

We see that the means of the exponential of 1000 simulations center around the theoretical mean 1/lambda = 5

Question 2

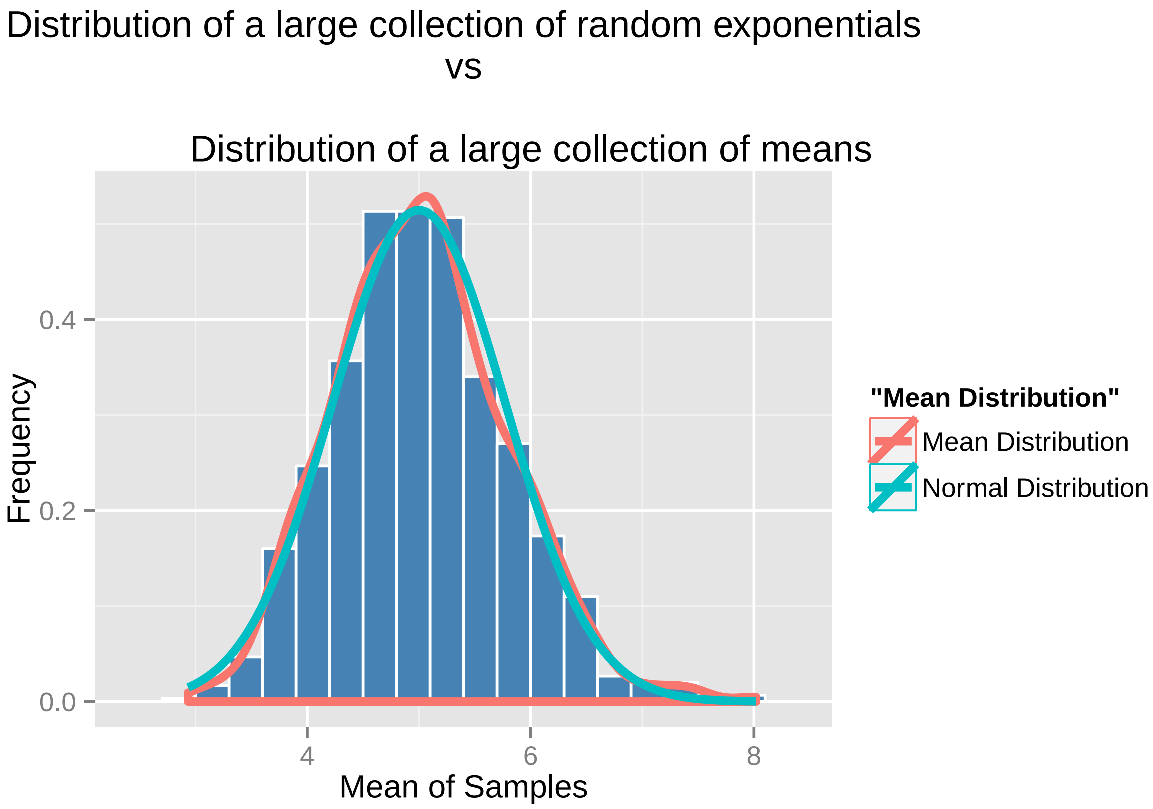
```r
# Theoretical variance of the distribution (1/lambda^2)/40 = .625
# Variance of the the random variables of means (rmeans1000)
var(rmeans1000$X)
```

```
## [1] 0.6009932
```

Again, we see that the variance of the means of the exponential of 1000 simulations is very close to the the theoretical variance (1/lambda^2)/40 = .625

Question 3

```r
library(ggplot2)
ggplot(data = rmeans1000, aes(x = X)) +
    geom_histogram(aes(y=..density..), fill = I('steelblue'),
                   binwidth = 0.30, color = I('white')) +
        geom_density(aes(colour="Mean Distribution"), size=1.5) +
        stat_function(fun = dnorm, aes(colour='Normal Distribution'),
        size=1.5, arg = list(mean = 5, sd = sd(rmeans1000$X))) +
        labs(title="Distribution of a large collection of random exponentials\n vs \n
             Distribution of a large collection of means") +
        labs(x="Mean of Samples", y="Frequency")
```



The Graph suggests that the distribution of a large collection of random exponentials centers around distribution of a large collection of means both of which are approximately normal.