Simulation

Amy Lee

November 8, 2019

Simulation Project

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.

Overview

This project was completed to observe the effect of Central Limit Theorem. Central Limit Theorem states, the distribution of the sample means approximate a normal distribution as the sample size increases. In this project, we will use the exponential distribution as our sample where lambda=0.2 and each sample size=40.

Before we start the project, packages must be imported first.

Necessary packages

```
library(ggplot2)
library(Hmisc)
library(gridExtra)
```

And now, we need to make the samples (1000 samples of n=40, lambda=0.2 exp distributions)

Making Samples

```
set.seed(1)
exp<-c()
for (i in 1:1000){
  exp<-rbind(exp,rexp(40,0.2))
}
```

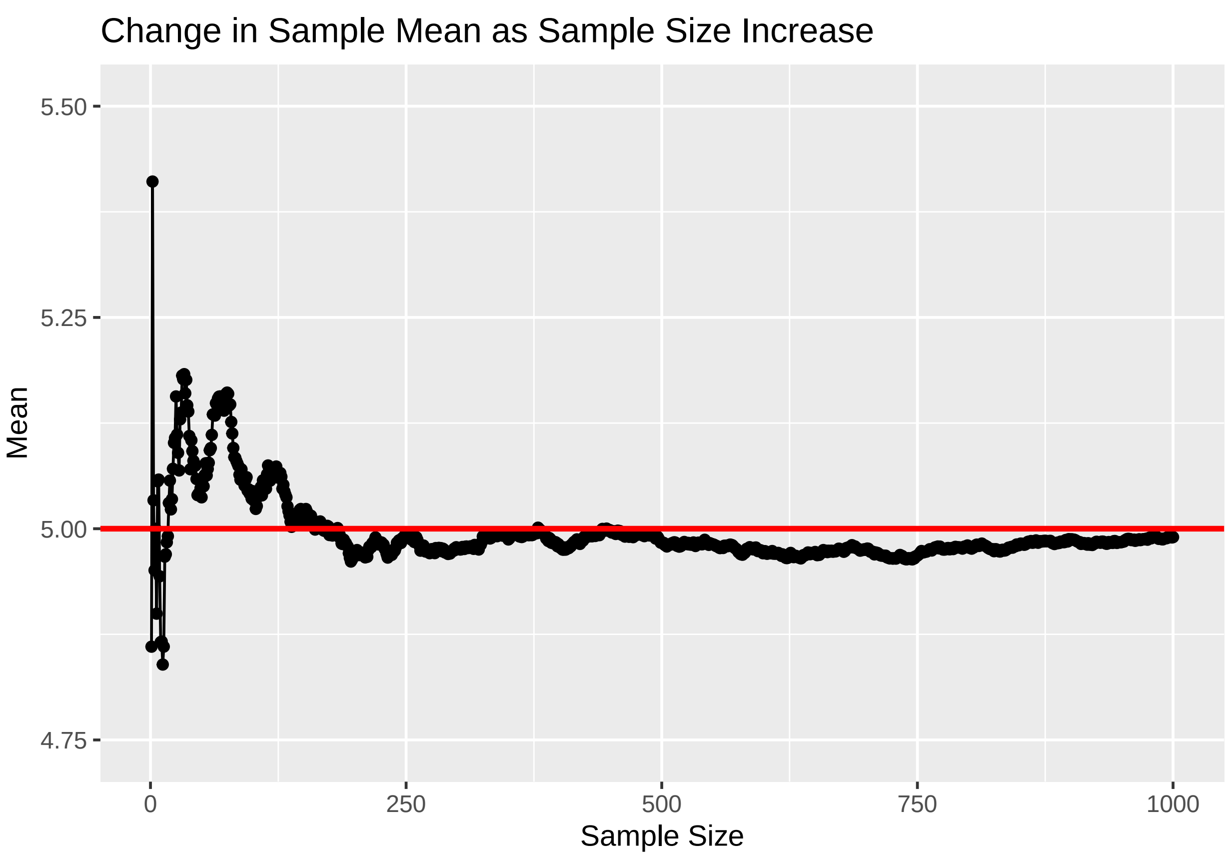
Here is the sample data. Each row is one sample exponential distribution with lambda = 0.2 and n = 40. There are 1000 sample datas (1000 rows) Now, we can start answering question.

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

```
mns<-c()
for (i in 1:1000) {
  mns[i]<-mean(exp[i,])
}
means<-cumsum(mns[1:1000])/(1:1000)
```

First, I made the mns vector which contains the means of all distribution of the 1000 samples. And made the cumulative sum vector called means. Then, I calculated the cumulative means to see how the mean changes as the sample size increased.

```
g_m<-ggplot(data.frame(x=1:1000, y=means), aes(x=x, y=y))+geom_line(size=0.5)+geom_point()
g_m<-g_m+ylim(min(means)-0.1,max(means)+0.1)+geom_hline(yintercept=5, size=1, color="red")
g_m<-g_m+labs(title="Change in Sample Mean as Sample Size Increase", x="Sample Size", y="Mean")
g_m
```



As it can be observed on the graph, the sample mean converges to the theoretical mean of 5.

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

Variance of the sample mean is the variance of the population over the sample size. So, the variance of the exponential distribution is 1/lambda^2 which is 25. And the sample size is 40. This makes the variance of the sample mean 0.625 So the variance of the sample should be similar to 0.625.

```r
var(mns)
```

```
## [1] 0.6111165
```

As shown above, it can be seen that the sample variance is similar to the theoretical variance.

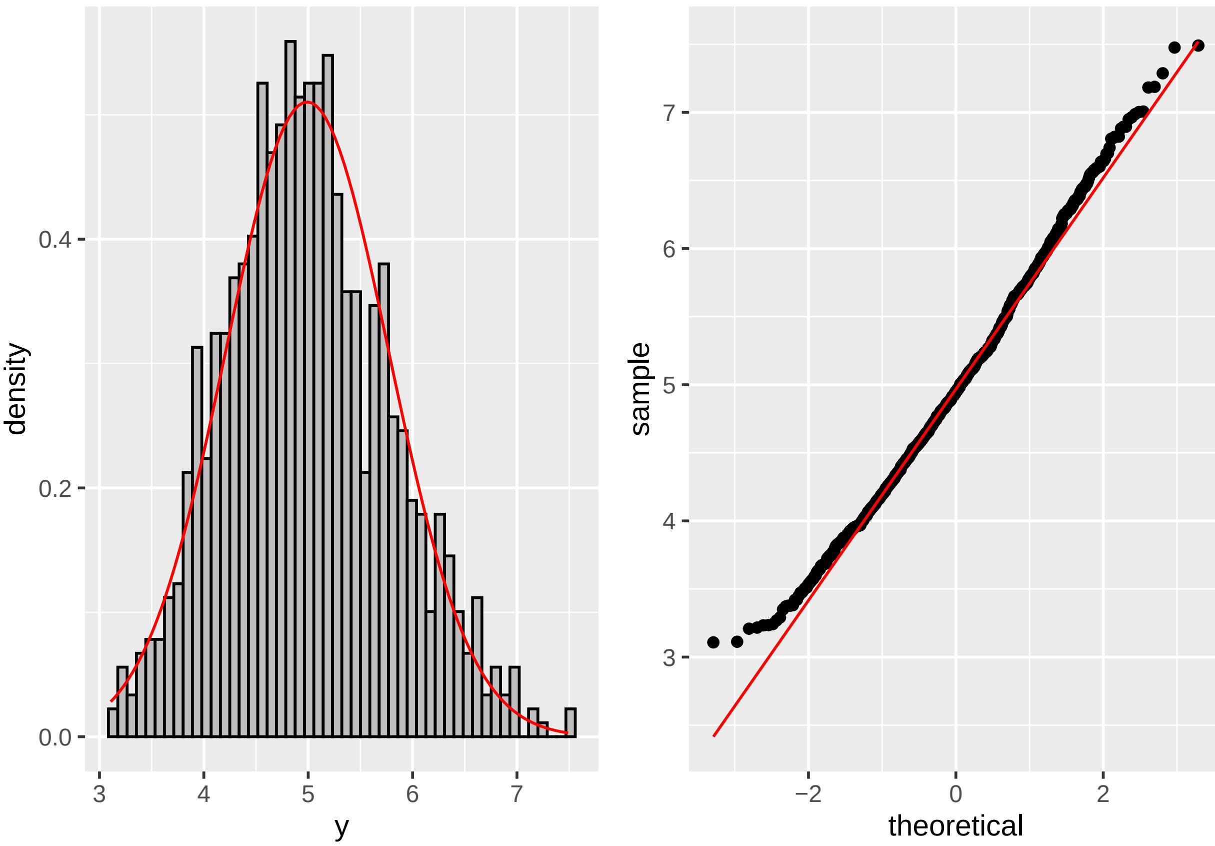
3) Show that the distribution is approximately normal.

One of the most common way of showing normality of data is to see the histogram and QQ plot.

```r
par(mfrow=c(1,2))
g_h<-ggplot(data.frame(x=1:1000, y=mns), aes(x=y))+geom_histogram(aes(y=..density..), bins=50,color="bla
g_h<-g_h+stat_function(fun = dnorm, args = list(mean = mean(mns), sd = sd(mns)), color="red")

g_q<-ggplot(data.frame(x=1:1000, y=mns), aes(sample=mns))+stat_qq()+stat_qq_line(color="red")

grid.arrange(g_h, g_q, nrow=1)
```



We can see that the histogram is approximating the bell curve. Another way of showing normality of a sample is by observing the Q-Q plot. In the QQ plot, we can observe the straight line, which means the sample distribution is basically normal.