Stastical Inference Part 1: Simulation Exercise

Rhay

3/14/2021

Overview of Requirements

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

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

Data Processing

Load libraries Set variables defined in problem statement

Load data

```r
set.seed(100)
#number of exponentials
n<-40

#lambda for rate
lambda<- 0.2

#number of simulations
numsim <- 1000

#Confidence Interval 95th quantile
quantile<-1.96
```

Compare Means

```r
sampledata<-matrix(rexp(n*numsim, rate=lambda), numsim)

#calculate mean for sampledata row
samplemean<-rowMeans(sampledata)

#calculate mean of means for each row
MeanOfsamplemean <- mean(samplemean)
```

Find mean of theoretical data

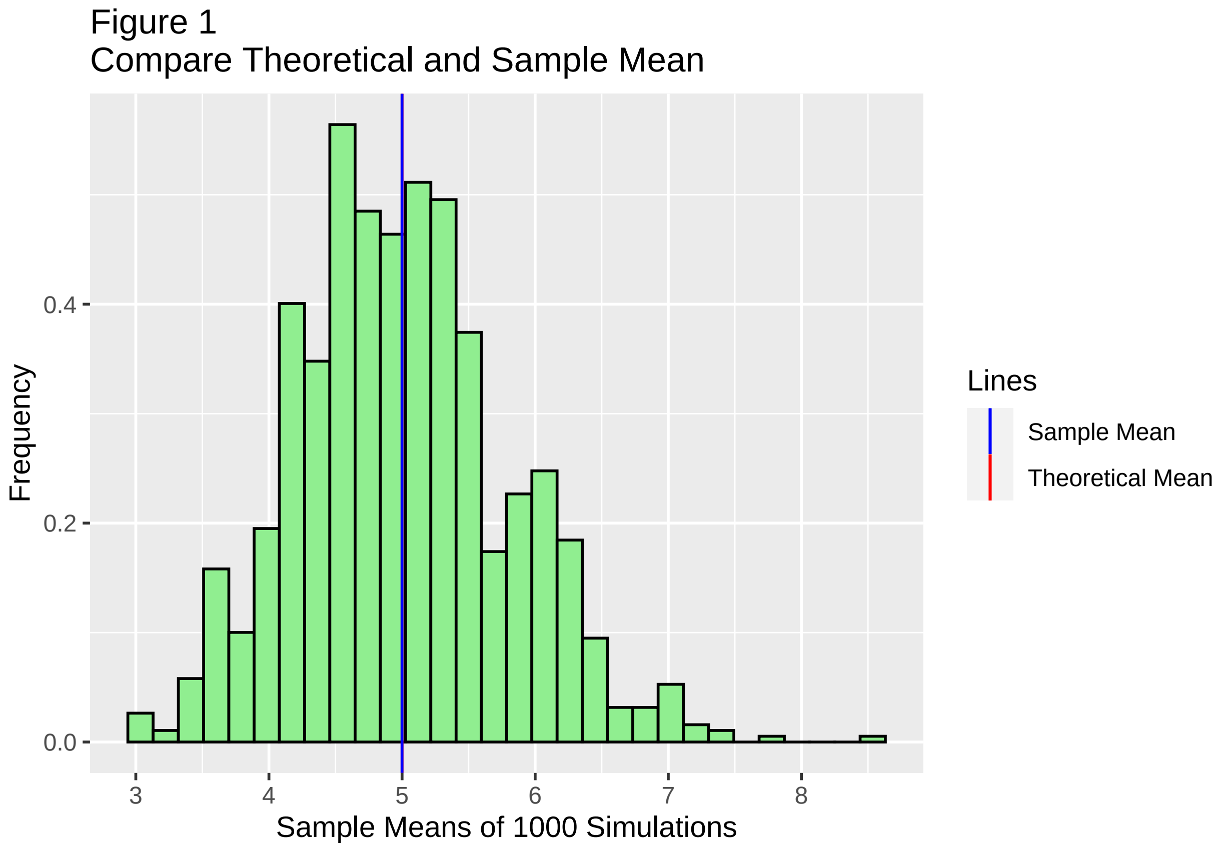
```r
theorMean<- 1/lambda

cat("The sample mean is: ", MeanOfsamplemean, "and the theoretical mean is:", theorMean)
```

```
## The sample mean is:  4.999702 and the theoretical mean is: 5
```

Show histogram to compare same and theoretical means

```r
plotdata<-data.frame(samplemean)
g1 <- ggplot(plotdata, aes(samplemean))
g1 <- g1 + geom_histogram(aes(y=..density..), colour="black",
                        fill = "lightgreen")
g1 <- g1 + geom_vline( aes(xintercept = theorMean,colour="Theoretical Mean"))
g1 <- g1 + geom_vline( aes(xintercept =mean(samplemean), colour="Sample Mean"))
g1 <- g1 + scale_colour_manual(name='Lines', values = c("Theoretical Mean"="red",
                                                        "Sample Mean"="blue"))
g1 <- g1 + labs(x = "Sample Means of 1000 Simulations")
g1 <- g1 + labs(y = "Frequency")
g1 <- g1 + labs(title = "Figure 1 \nCompare Theoretical and Sample Mean")
g1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Both means are very close and it's hard to distinguish between then on the graph.

Compare variance and std. deviation between sample and theoretical

```
#sample e
samplevar<-var(samplemean)

samplesd<-sd(samplemean)

#theoretical
theorVar<-(1/lambda)^2/(n)

theorSD<-1/(lambda * sqrt(n))

#add normally distributed plots of sample and theoretical curves to plot
g2 <- ggplot(plotdata, aes(samplemean))
g2 <- g2 + geom_histogram(aes(y=..density..), colour="black",
                          fill = "lightyellow")
g2 <- g2 + geom_vline( aes(xintercept = theorMean,colour="Theoretical Mean"))
g2 <- g2 + geom_vline( aes(xintercept =mean(samplemean), colour="Sample Mean"))
g2 <- g2 + scale_colour_manual(name='Lines', values = c("Theoretical Mean"="red",
                                                         "Sample Mean"="blue"))
g2 <- g2 + labs(x = "Sample Means of 1000 Simulations")
g2 <- g2 + labs(y = "Frequency")
g2 <- g2 + labs(title = "Figure 2 \nCompare Theoretical and Sample Distribution")
```

```
g2<- g2+stat_function(fun = dnorm, args = list(mean = theorMean, sd = theorSD), colour = "red", size = 
g2<- g2+stat_function(fun = dnorm, args = list(mean = MeanOfsamplemean, sd = samplesd), colour = "black
g2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

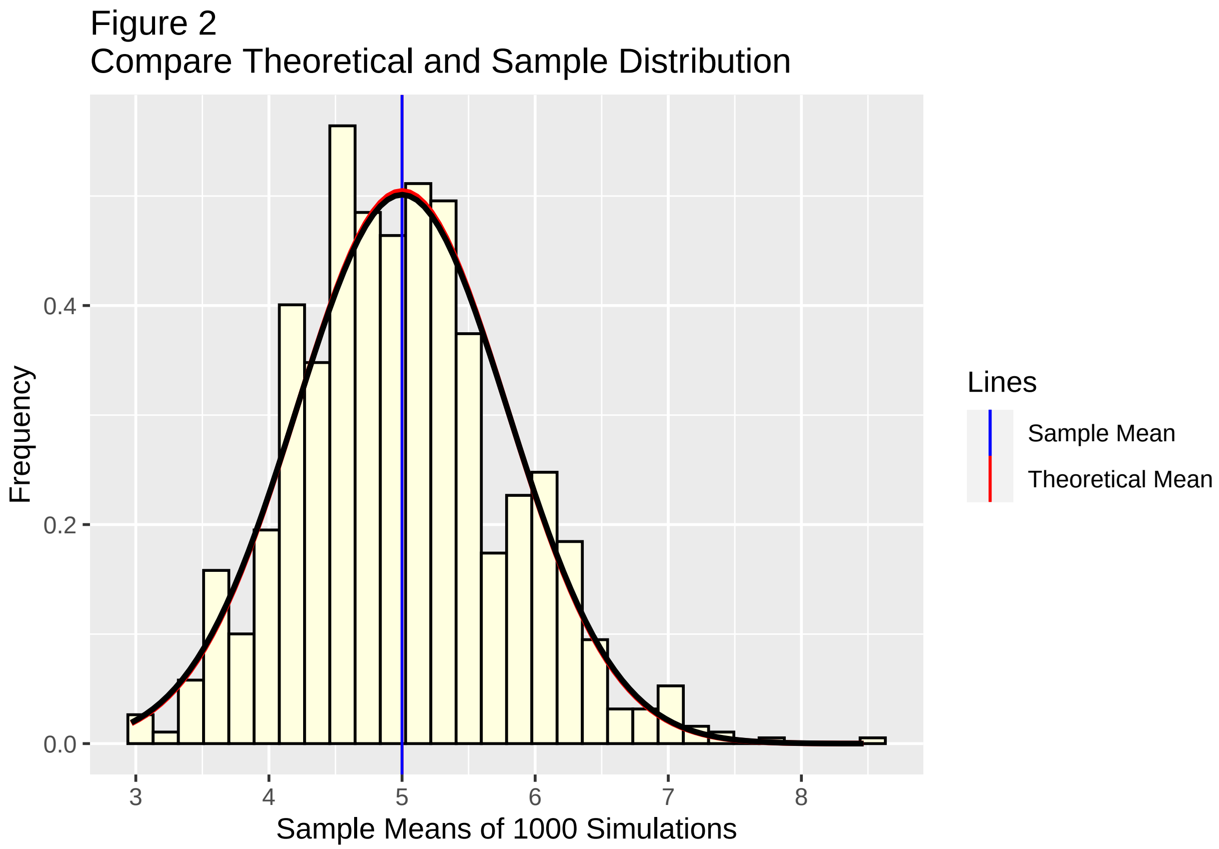


Figure 2
Compare Theoretical and Sample Distribution

The bars on the graph show the actual data. It's difficult to distinguish between the sample and theoretical means as they overlap. The blue line shows the sample data while the red line shows the theoretical. Both distributions are very close to the normal distribution.

Confidence Interval Comparison

```
#Calculate the Conficend Interval for sample and theoretical
sampleCI<-round (mean(samplemean) + c(-1,1)*1.96*sd(samplemean)/sqrt(n),3)

theorCI<-theorMean + c(-1,1) * 1.96 * sqrt(theorVar)/sqrt(n)
cat("The sample confidence interval is:     ", sampleCI, " and \nthe theoretical confidence interval is
```

```
## The sample confidence interval is:      4.753 5.246  and
## the theoretical confidence interval is: 4.755 5.245
```

Both confidence intervals match closing indicating the the distributions are approximately normal.

Conclusion

From looking at the distributions of the sample and theoretical data, we can conclude that both the data shows a bell curve indicative the the Central Limit Theorom.