Statistical Inference Project

WendyD

9/8/2020

Overview

In this project, we investigated the exponential distribution in R compared to what we learned in Central Limit Theorem. We worked with 40 randomized variable and conducted 1000 stimulation. We pre-determined lamda is 0.2

#Simulation

```
library(knitr)
```

```
n_simulation<- 1000
n<- 40
lambda<- 0.2
rep<-rexp(n*n_simulation, 0.2)
```

Inputing the result into a matrix for easier manipulation

```
data<- matrix(rep, n_simulation, n)
```

Sample Mean and Theoretical mean

Theoretical Mean

```
theoretical_mean<-1/lambda
```

Sample Mean For each simulation (Row), we calculated the mean

```
sim_mean<-rowMeans(data)
sample_mean<- mean(sim_mean)
```

Making a data.frame for easier comparision

```
data.frame(theoretical_mean, sample_mean)
```

theoretical_mean <dbl>	sample_mean <dbl>
5	4.970358

1 row

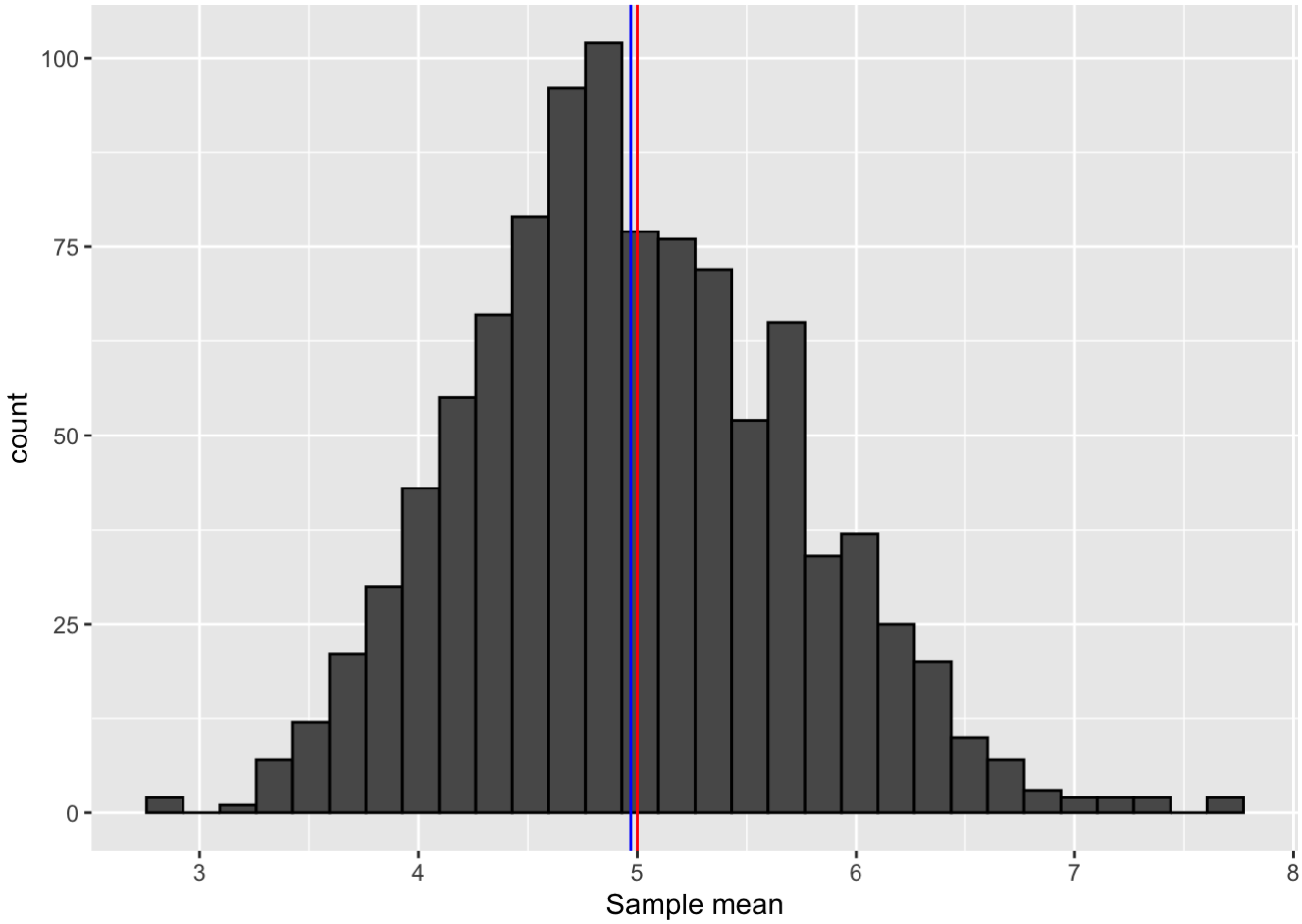
We see that that the sample mean is pretty close to theoretical mean (4.9544 vs 5)

It would be helpful to plot the distribution of the mean over a histogram plot, then input two line representing sample mean(blue) and our theoretical mean(in Red)

```
library(ggplot2)
```

```
sim_meanframe<- as.data.frame(sim_mean)
ggplot(data=sim_meanframe, aes(x= sim_mean))+
       geom_histogram(position="identity", colour="black")+
       xlab("Sample mean") + geom_vline(xintercept = sample_mean, colour="blue")+
       geom_vline(xintercept = theoretical_mean, colour="red")
```

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



Theoretical Variance vs Sample Variance

We first begin by computing the variance from our actual sample, and theoretical variance:

```
sample_var<- var(sim_mean)
theoretical_var<- (1/lambda)^2/n
```

Making a data.frame for easier comparision

```
data.frame(theoretical_var,sample_var)
```

theoretical_var <dbl>	sample_var <dbl>
0.625	0.5575089

1 row

The sample variance is 0.5999 while theoretical variance is 0.625. These two values are close.

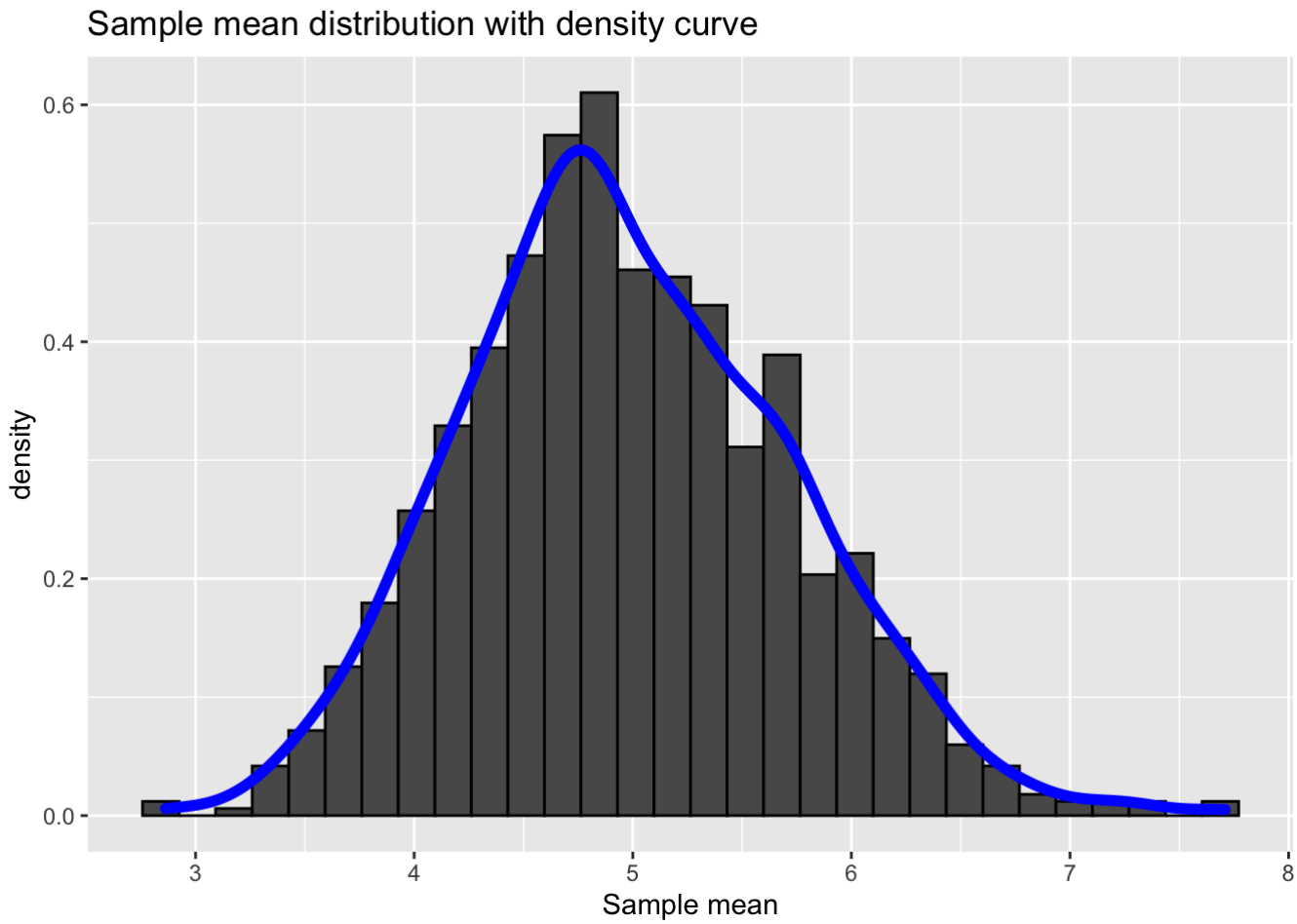
#Distribution

We want to compare our data distribution with the distribution from central limit theorem. According to CLT, the average of our sample will follow normal distribution.

We first plot our data and a density curve to see if our data align with the normal distribution

```
ggplot(data=sim_meanframe, aes(x=sim_mean))+
       geom_histogram(aes(y = after_stat(density)), position="identity", colour="black")
+
       xlab("Sample mean")+
        ggtitle("Sample mean distribution with density curve")+
        geom_density(colour="blue", size=2)
```

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



We now have compared the theoretical mean vs sample mean, theoretical variance vs sample variance. We will investigate the last element: confidence interval

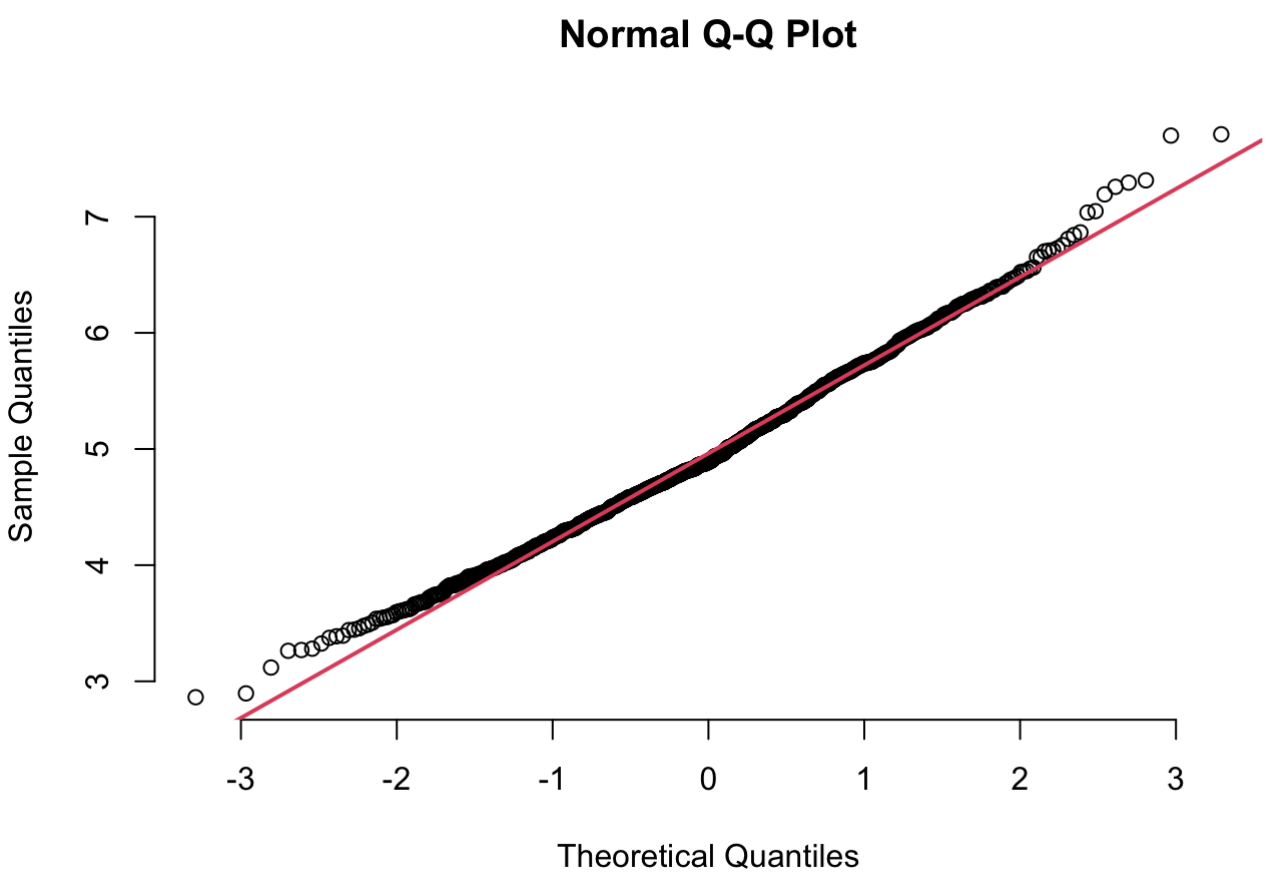
```
sample_conf_interval <- round (mean(sim_mean) + c(-1,1)*1.96*sd(sim_mean)/sqrt(n),3)


theoretical_conf_interval <- theoretical_mean+ c(-1,1)*1.96*sqrt(theoretical_var)/sqrt
(n)
```

The actual 95% confidence interval is `sample_conf_interval` and the theoretical confidence interval `theoretical_conf_interval` are pretty close to one another also

Lastly, we used qqplot to draws the correlation between a given sample and the normal distribution. If the data is normally distributed, the points in the QQ-normal plot lie on a straight diagonal line

```
qqnorm(sim_mean, pch = 1, frame = FALSE)
qqline(sim_mean,col ="2",lwd = 2)
```



The difference between the dots and the line is minimal. Thus we can conclude that the sample is relatively normally distributed.