Statistical Inference Course Project - Part 1 & Appendix

Roshan Riazi

Overview

In this project we want to investigate the exponential distribution in R and compare it with the Central Limit Theorem. Infact we want to investigate the distribution of averages of 40 exponentials and according to Central Limit Theorem, show that this distribution is approximately normal.

Simulations

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. We should set lambda = 0.2 for all of the simulations and investigate the distribution of averages of n = 40 exponentials. Note that we will need to do a thousand simulations (nsim = 1000) and store the result of these simulations in "mns" variable. We will set the seed, so that our results would be reproducible.

```
set.seed(1)
lambda = 0.2
nsim = 1000
n = 40
mns <- NULL
for(i in 1:nsim){
    mns[i] <- mean(rexp(n, lambda))
}
```

Sample Mean versus Theoretical Mean

We want to show the sample mean and compare it to the theoretical mean of the distribution. According to wikipedia page of central limit theorem:

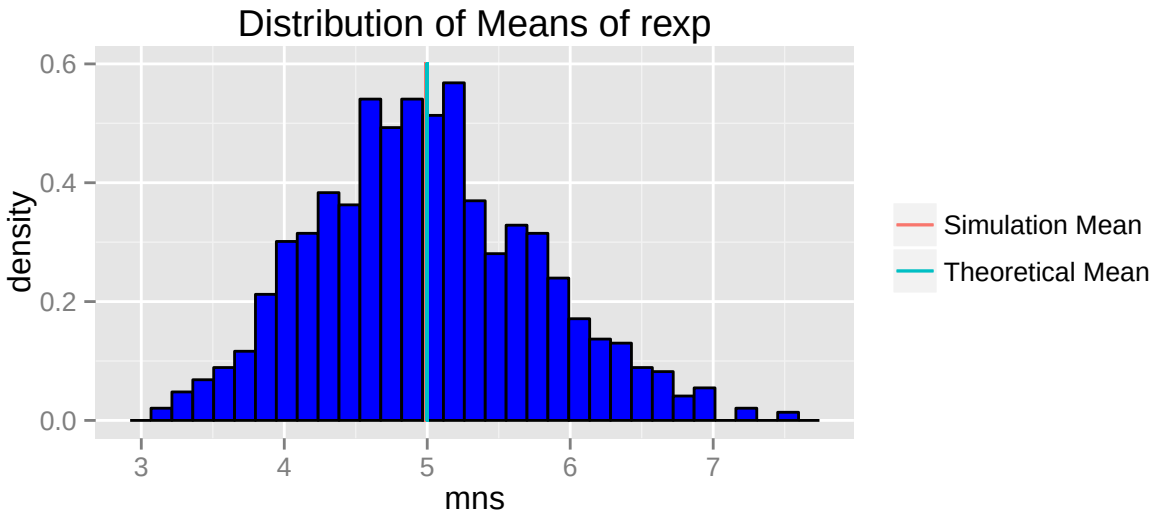
> In probability theory, the central limit theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

According to CLT, and with the knowledge that theoretical mean of exponential distribution is 1/lambda, the theoretical mean of averages of n exponential iid variables would also be 1/lambda.

```
theoMean <- 1/lambda
simMean <- mean(mns)
c(theoMean, simMean)
```

```
## [1] 5.000000 4.990025
```

We can see that theoretical and simulation means are very close, with theoretical mean of 5 and simulation mean of 4.99. We can also plot these averages and visually compare theoretical and simulation means.



Sample Variance versus Theoretical Variance

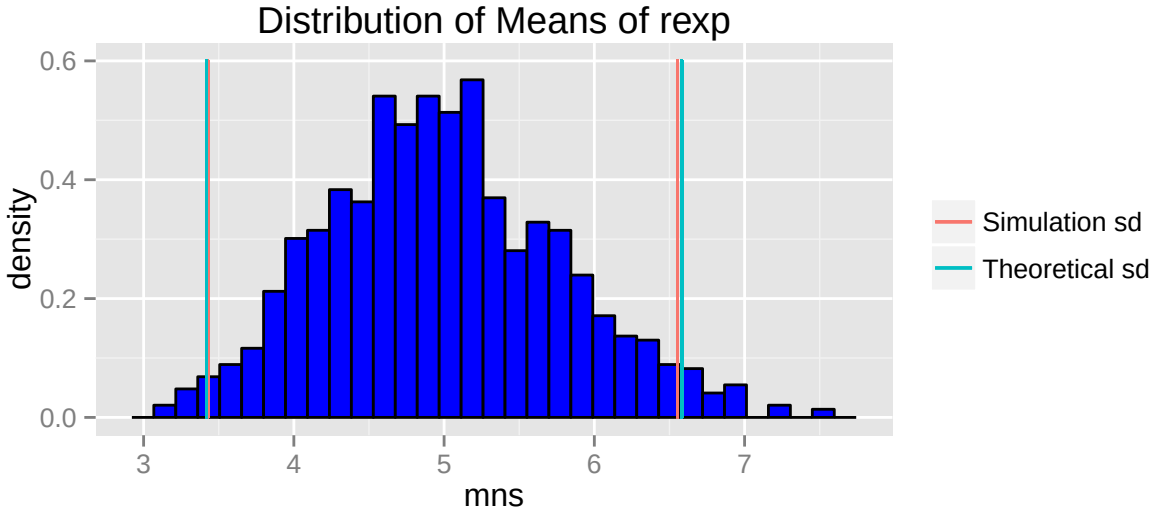
We want to Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. According to CLT, and with the knowledge that theoretical standard deviation of exponential distribution is 1/lambda, the theoretical standard deviation of averages of n exponential iid variables would be 1/(lambda*sqrt(n)).

```
theoVar <- 1/((lambda^2)*n)
simVar <- var(mns)
theoSd <- sqrt(theoVar)
simSd <- sqrt(simVar)
c(theoVar, simVar)
```

```
## [1] 0.6250000 0.6111165
```

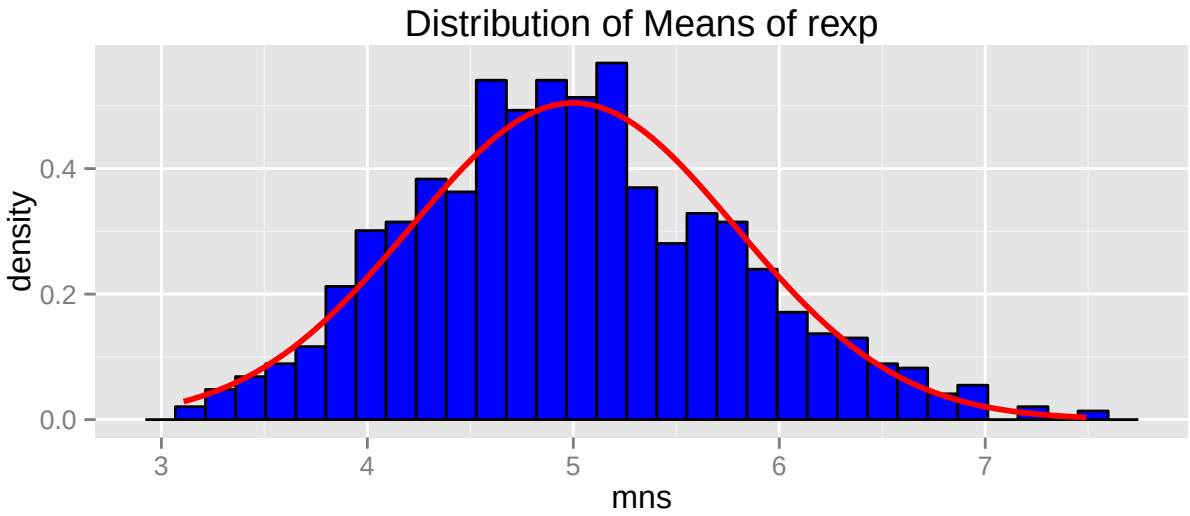
We can see that theoretical and simulation variances are very close, with theoretical variance of 0.625 and simulation variance of 0.6111. We expect that as the number of n and simulations gets bigger, these values get closer to eachother.

We can also plot this distribution and draw horizontal lines indicating intervals with, for example, 2 standard deviations from mean (about 95% interval) to visually compare theoretical and simulation standard deviations.



Distribution

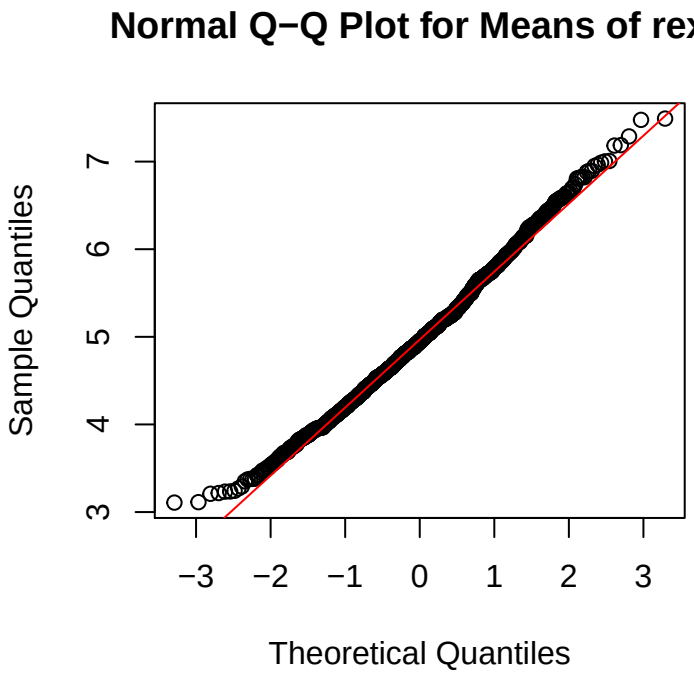
We want to show that the distribution is approximately normal, so the CLT is right. We can plot distribution of averages of 40 exponential iid exponential variables and overlay density of a normal distribution to check it.



We can see that this distribution is approximately normal, and we expect that as number of n and simulations increases, this distribution will get closer to normal distribution.

We can also check qqplot for this distribution:

```
qqnorm(mns, main = "Normal Q-Q Plot for Means of rexp")
qqline(mns, col = "red")
```



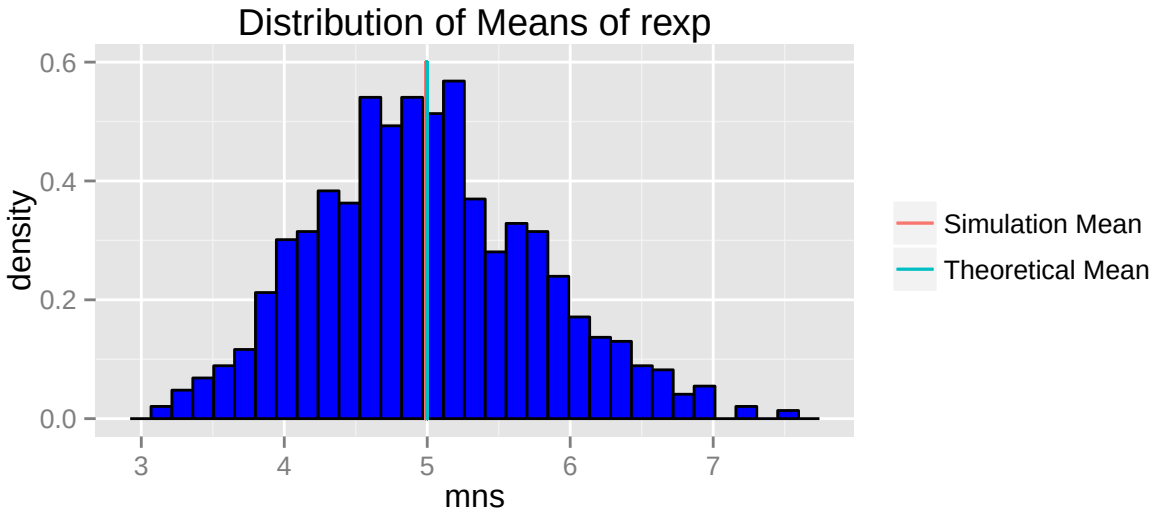
It's mostly on the line, which indicates that it is approximately normal.

Appendix

Codes

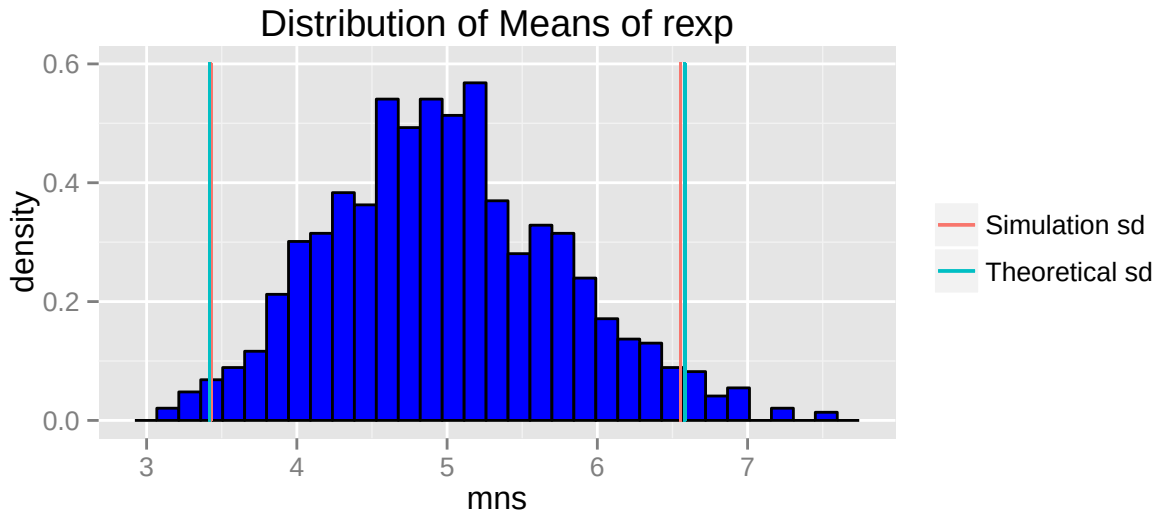
Code of plot 1

```
library(ggplot2)
plotData <- data.frame(mns)
plot1 <- ggplot(data = plotData, aes(x = mns))
plot1 <- plot1 + geom_histogram(aes(y = ..density..), fill = "blue", color = "black")
plot1 <- plot1 + geom_line(aes(x = c(simMean, simMean), y = c(0, .6),
                               col = "Simulation Mean"))
plot1 <- plot1 + geom_line(aes(x = c(theoMean, theoMean), y = c(0, .6),
                               col = "Theoretical Mean"))
plot1 <- plot1 + guides(col = guide_legend(title = NULL)) #removing legend title
plot1 <- plot1 + labs(title = "Distribution of Means of rexp")
plot1
```



Code of plot 2

```
plot2 <- ggplot(data = plotData, aes(x = mns))
plot2 <- plot2 + geom_histogram(aes(y = ..density..), fill = "blue", color = "black")
plot2 <- plot2 + geom_line(aes(x = c(simMean+2*simSd, simMean+2*simSd), y = c(0, .6),
                               col = "Simulation sd"))
plot2 <- plot2 + geom_line(aes(x = c(simMean-2*simSd, simMean-2*simSd), y = c(0, .6),
                               col = "Simulation sd"))
plot2 <- plot2 + geom_line(aes(x = c(theoMean+2*theoSd, theoMean+2*theoSd), y = c(0, .6),
                               col = "Theoretical sd"))
plot2 <- plot2 + geom_line(aes(x = c(theoMean-2*theoSd, theoMean-2*theoSd), y = c(0, .6),
                               col = "Theoretical sd"))
plot2 <- plot2 + guides(col = guide_legend(title = NULL)) #removing legent title
plot2 <- plot2 + labs(title = "Distribution of Means of rexp")
plot2
```



Code of plot 3

```
plot3 <- ggplot(plotData, aes(x = mns))
plot3 <- plot3 + geom_histogram(aes(y = ..density..), fill = "blue", color = "black")
plot3 <- plot3 + stat_function(fun = dnorm, args = list(mean = theoMean, sd = theoSd),
                               color = "red", size = 1)
plot3 <- plot3 + labs(title = "Distribution of Means of rexp")
plot3
```