Statistical Inference Course Project: An experiment about the Central Limit Theorem(CLT)

Emily

Overview

In this project we will first investigate the exponential distribution and then compute the distribution of sample means. With the CLT we predict this distribution would be a normal distribution with certain known mean and standard deviation when the number of simulations is large enough. We compared the theoretical and simulation results to verify our prediction.

Simulations

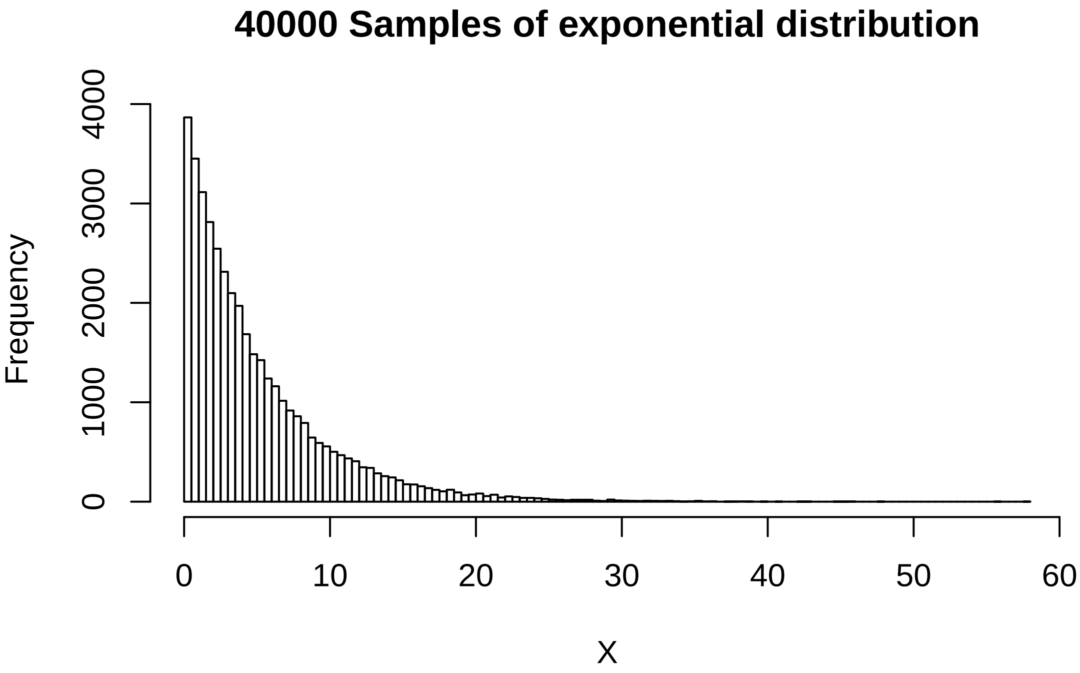
First we load useful libraries and set several global variables.

```r
library(knitr)
library(ggplot2)
set.seed(1)
lambda <- 0.2
sample_size <- 40
nosim <- 1000
n <- sample_size*nosim
```

Sample Mean versus Theoretical Mean

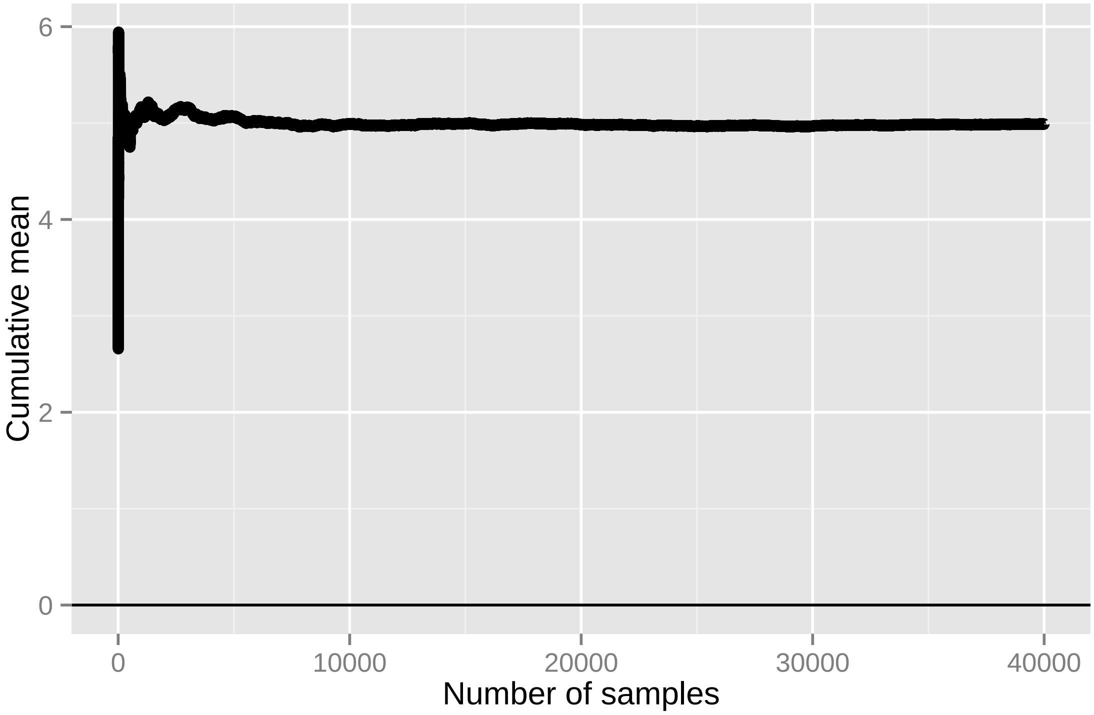
The exponential distribution has density function $f(x) = \lambda e^{-\lambda x}$ and it can be simulated in R with `rexp(n, lambda)`. The mean of exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$. Here we set $\lambda = 0.2$ for all of the simulations. First we draw 40000 samples from exponential distribution and plot the distribution.

```r
data <- rexp(n,lambda)
hist(data,breaks=100,main="40000 Samples of exponential distribution", xlab="X")
```



We can see the distribution is asymmetric and the probability density function decays exponentially when x increases. Next we plot the cumulative mean versus sample size to see how the sample mean approaches a fixed value which will be shown later that is the theoretical mean 5(by the Law of Large Numbers).

```
means <- cumsum(data)/(1:n)
g <- ggplot(data.frame(x = 1 : n, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0) + geom_line(size = 2)
g <- g + labs(x = "Number of samples", y = "Cumulative mean")
g
```



We divide the 40000 draws into 1000 groups of simulations, each contains 40 samples of exponential distribution. Then we calculate means of each 40 exponentials and compute the mean of those means. We find the mean of sample means is 4.990025 which is every close to the theoretical mean stored in pop_mean.

```r
mns <- apply(matrix(data,nosim),1,mean)
mean(mns)
```

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

```r
pop_mean <- 1/lambda
pop_mean
```

```
## [1] 5
```

Sample Variance versus Theoretical Variance

Similarly, we compute the variance of the distribution of sample means. From theory, we know it is Var(S)= $\frac{\sigma^2}{40}$ where $\sigma = \frac{1}{\lambda}$ is the variance of the exponential distribution. We find the variance of sample means is 0.6177072 which is close to the theoretical variance 0.625.

```r
var(mns)
```

```
## [1] 0.6177072
```

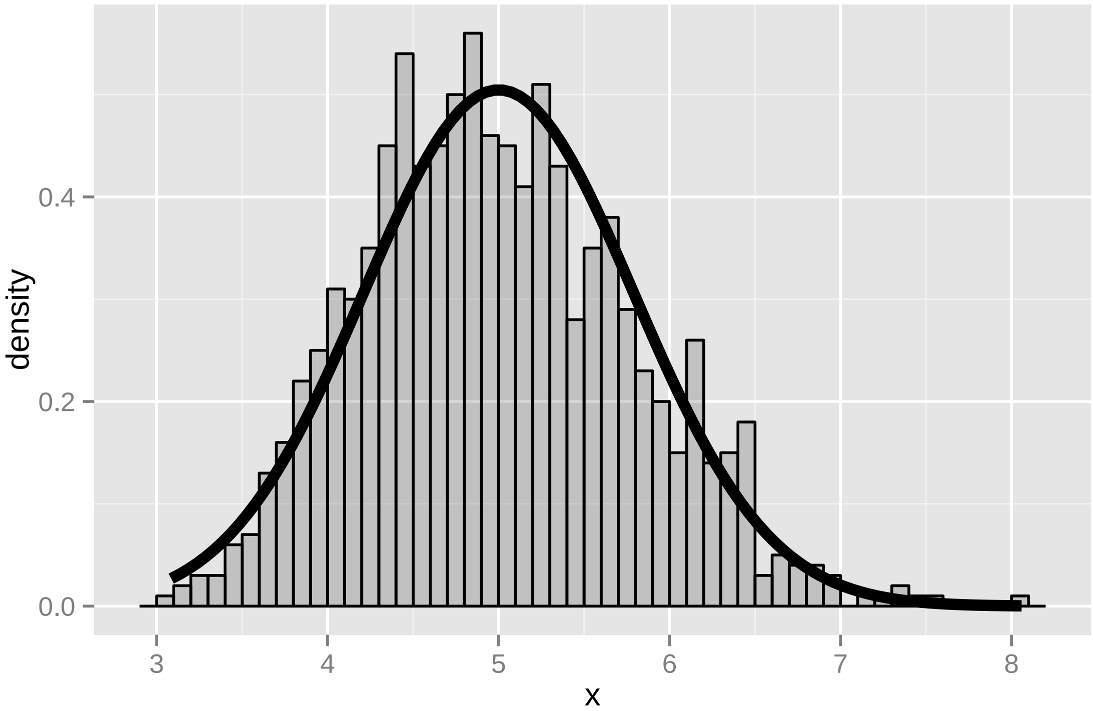
```r
pop_sd <- 1/lambda
# theoretical variance of the distribution
mns_var <- pop_sd^2/sample_size
mns_var
```

```
## [1] 0.625
```

Distribution

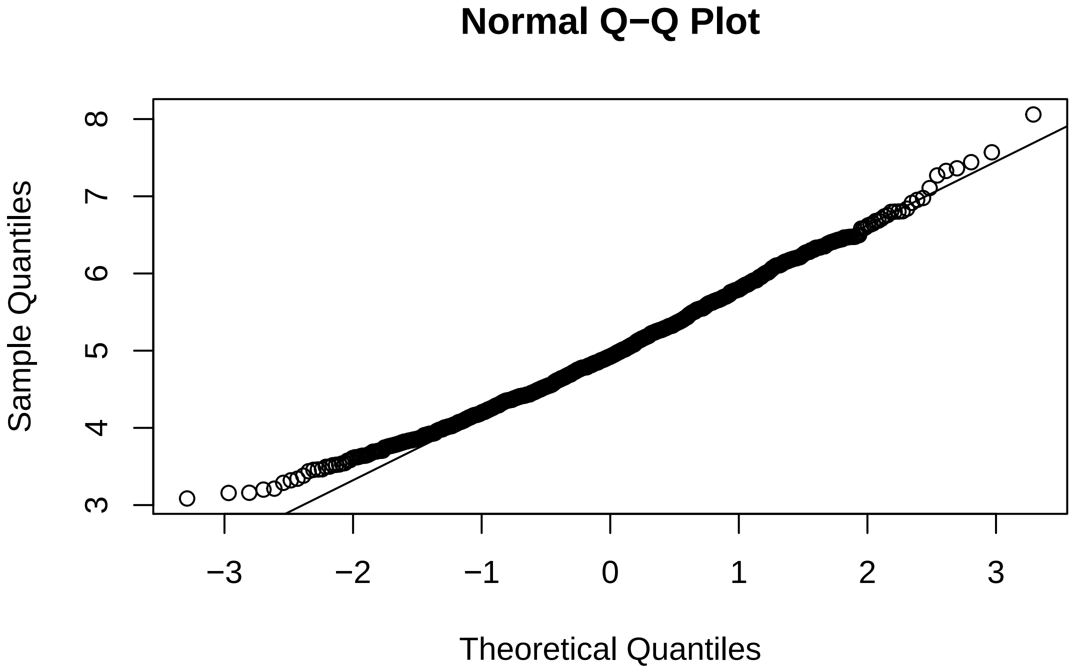
Lastly, we compared the distribution of means with the normal distribution. From CLT we know the sample means follow normal distribution. The following plot of sample means together with normal distribution(with theoretical mean and variance) verifies this prediction.

```r
# Compare the distribution of sample means to the normal distribution
mns1 <- data.frame(x=mns)
g <- ggplot(mns1,aes(x=x)) + geom_histogram(alpha = .20, binwidth=.1, colour = "black",
aes(y = ..density..))
g <- g + stat_function(fun = dnorm, arg=list(mean=pop_mean,sd=sqrt(mns_var)),size = 2)
g
```



Another way to compare two distributions is through the comparison of quantiles. In the following, the qq-plot shows the quantiles of the obtained distribution of means compared with the quantiles of the normal distribution. It also suggests the normality of the sample mean distribution.

```r
qqnorm(mns)
qqline(mns)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles