Statistical Inference Course Project 1

Anuar Imanbayev

January 30, 2016

GitHub Link (PDF, Markdown, Html): https://github.com/anuarimanbayev/datasciencecoursera/tree/master/06_StatisticalInference/SI_Project

Comparison of Exponential Distribution to the Central Limit Theorem

Overview

In this project, we 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. We will investigate the distribution of averages of 40 exponentials. Note that we 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. We 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.

Simulations

Sample Mean vs Theoretical Mean

We will run a series of 1000 simulations to create a data set for comparison to theory. Each simulation will contain 40 observations and the exponential distribution function will be set to "rexp(40, 0.2)".

Known values: lambda = 0.2, n = 40, simulations = 1000

```
# Load neccesary libraries
library(ggplot2)

# Set constants
lambda = 0.2
n = 40
nosim = 1000

set.seed(756)
```

The following code performs the simulations to collect necessary data, then plots the data:

```
exp_sim <- function(n, lambda)
{
          mean(rexp(n,lambda))
}

sim <- data.frame(ncol=2,nrow=1000)
names(sim) <- c("Index","Mean")
```

```
            for (i in 1:nosim)
            {
                      sim[i,1] <- i
                      sim[i,2] <- exp_sim(n,lambda)
            }
```

Mean of n = 1000

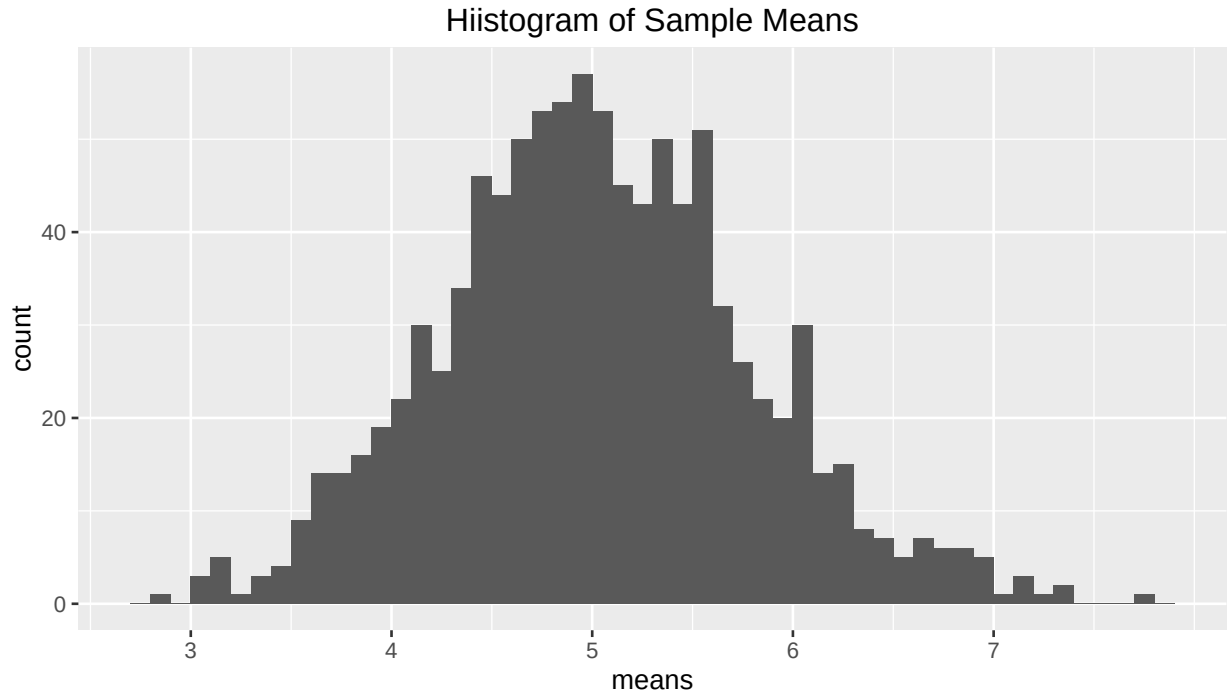
```
        sample_mean <- mean(sim$Mean)
        sample_mean
```

```
## [1] 4.972894
```

Histogram of sample means

```
exp_dis <- matrix(data=rexp(n * nosim, lambda), nrow=nosim)
exp_dis_means <- data.frame(means=apply(exp_dis, 1, mean))


ggplot(data = exp_dis_means, aes(x = means)) +
        geom_histogram(binwidth=0.1) +
        ggtitle("Hiistogram of Sample Means") +
        scale_x_continuous(breaks=round(seq(min(exp_dis_means$means), max(exp_dis_means$means), by=1)))
```



Theoretical exponential mean of exponential distribution

The expected mean μ of a exponential distribution of rate λ is

$\mu = \frac{1}{\lambda}$

```
        theor_mean <- 1/lambda
        theor_mean
```

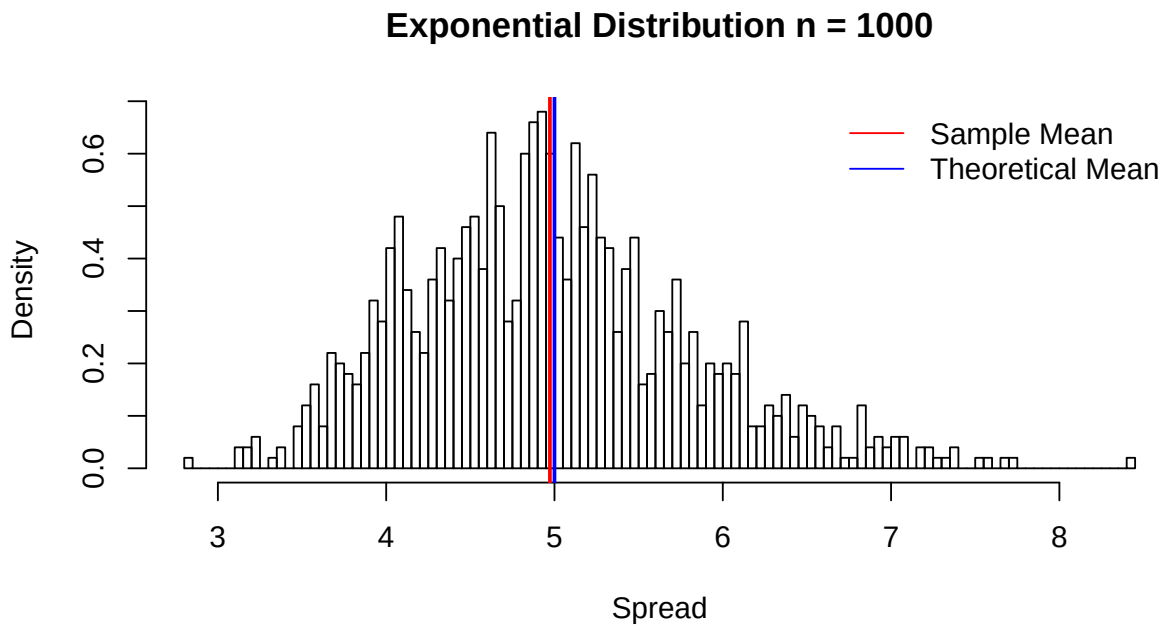
```
## [1] 5
```

The simulation mean of 4.972894 is close to the theoretical value of 5. In other words, the expected theoretical mean of 5 and the average sample mean of 4.97 are very close.

Histogram plot of the exponential distribution n = 1000

```
        hist(sim$Mean,
                breaks=100,
                prob=TRUE,
                main="Exponential Distribution n = 1000",
                xlab="Spread")
                        abline(v = theor_mean,
                                col= 4,
                                lwd = 2)
                        abline(v = sample_mean,
                                col = 2,
                                lwd = 2)

                legend('topright', c("Sample Mean", "Theoretical Mean"),
                                lty=c(1,1),
                                bty = "n",
                                col = c(col=2, col=4))
```



Sample Variance vs. Theoretical Variance

We now turn our attention to the variance. We will compare the variance present in the sample means of the 1000 simulations to the theoretical variance of the population.

The variance of the sample means estimates the variance of the population by using the variance of the 1000 entries in the means vector times the sample size, 40.

The expected standard deviation σ of a exponential distribution of rate λ is

$\sigma = \frac{1/\lambda}{\sqrt{n}}$

The variance Var of standard deviation σ is

$Var = \sigma^2$

This nets us the following: σ^2=Var(samplemeans)×N.

```
sample_var <- var(sim$Mean)
theor_var <- ((1/lambda)^2)/n
```

The theoretical variance of the population is given by σ^2=(1/lambda)2.

```
sample_var
```

```
## [1] 0.6912115
```

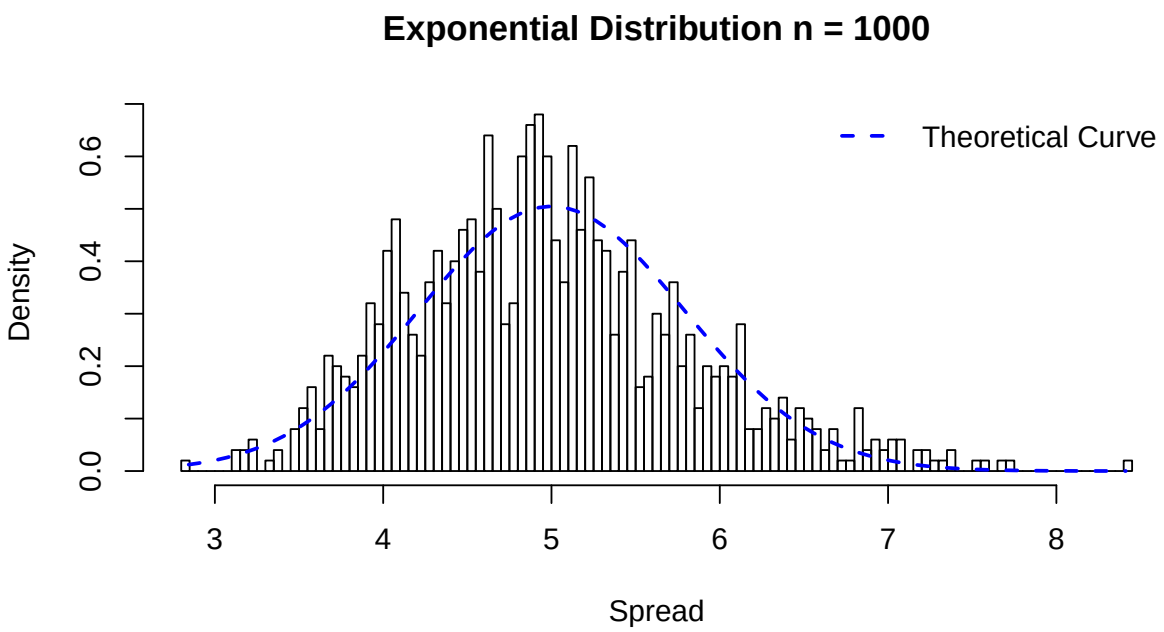
```
theor_var
```

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

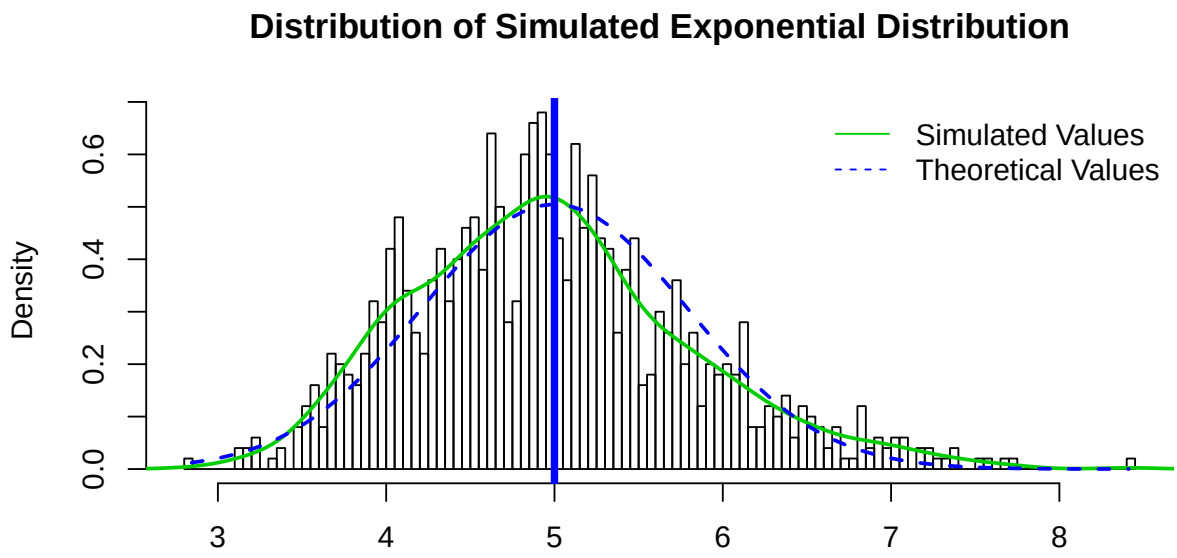
Thus, the variances are still fairly close, even given the fact that variance is the square of the standard deviations which enhances minor differences.

Sample Distribution vs. Theoretical Distribution

```
hist(sim$Mean,
                breaks = 100,
                prob = TRUE,
                main ="Exponential Distribution n = 1000",
                xlab ="Spread")
                xfit <- seq(min(sim$Mean), max(sim$Mean), length = 100)
                yfit <- dnorm(xfit, mean = 1/lambda, sd = (1/lambda/sqrt(40)))
                lines(xfit, yfit, pch=22, col=4, lty=2, lwd=2)
                legend('topright', c("Theoretical Curve"), bty="n", lty=2,lwd=2, col=4)
```




```
        hist(sim$Mean,
                     breaks = 100,
                     prob = TRUE,
                     main = "Distribution of Simulated Exponential Distribution", xlab="")
                     lines(density(sim$Mean), col=3, lwd=2)
                     abline(v = 1/lambda, col = 4, lwd=4)
                     xfit <- seq(min(sim$Mean), max(sim$Mean), length = 100)
                     yfit <- dnorm(xfit, mean = 1/lambda, sd = (1/lambda/sqrt(40)))
                     lines(xfit, yfit, pch=22, col=4, lty=2, lwd=2)
                     legend('topright', c("Simulated Values", "Theoretical Values"),
                     bty="n", lty=c(1,2), col=c(3,4))
```

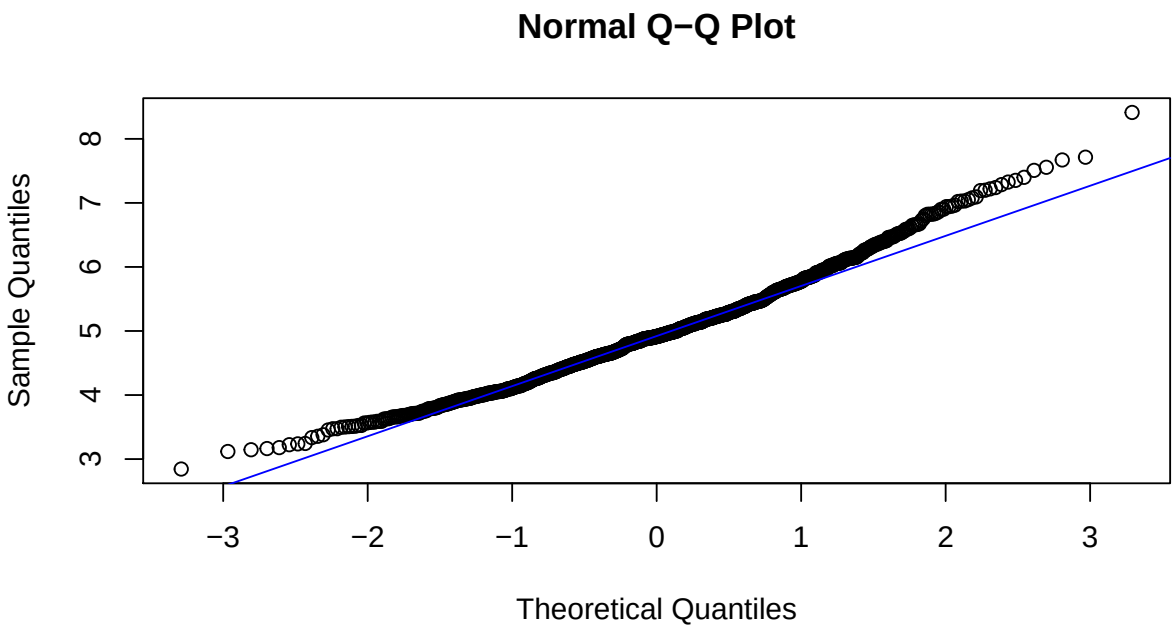


Thus, the calculated distribution of means of random sampled exponential distributions overlaps fairly closely with the normal distribution of expected theoretical values based on the given lambda.

Conclusion: Distribution is approximately normal

The close values of both the sample mean and variances to theoretical expected mean and variances suggest normality. The bell-shaped distribution shape of the simulated data that closely matches the theoretical curve also suggests normality. Additionally, the q-q plot below displays that theoretical quantiales again match closely with the actual quantiles. Together, these four methods (means, variances, distribution shape, quantiles) prove that the distribution is approximately normal.

```
qqnorm(sim$Mean,
                main ="Normal Q-Q Plot")
                qqline(sim$Mean,
                col = "4")
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
3
4
5
6
7
8
−3
−2
−1
0
1
2
3