Statistical Inference Project 2 Part 1

Nicolas Moreno Andrade

August 14, 2017

Overview

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. In particular we'll calculate the mean of 1000 simulated data sets each containing 40 samples of an exponential distribution. According to the Central Limit Theorem, the distribution of these means should approximate that of a normal distribution. By comparing the theoretical and simulated mean and variance and by plotting relevant figures we'll conclude that the simulated data conforms to the expectation of the Central Limit Theorem.

1 Simulations

The exponential distribution is simulated in R with **rexp(n,lambda)**, where $\lambda = 0.2$, sample size n is 40, and the number of simulations is 1000. The 1000 sample means are stored in simMeans.

```
set.seed(9)  # we set a seed for reproducibility

# set constants
lambda <- 0.2  # lambda for rexp
n <- 40  # number of exponentials
nsim <- 1000  # number of simulations

# store the results of the test in a nsim * n matrix
simMatrix <- matrix(data = rexp(n * nsim, rate = lambda), nsim)
# compute the mean of each simulation
simMeans <- rowMeans(simMatrix)
head(simMeans)

## [1] 5.622956 6.079371 4.721247 4.193506 5.762213 4.926966
```

2 Theoretical vs simulated mean and variance

The exponential distribution has the following probability density function:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

With mean λ^{-1} and variance λ^{-2}. Since we are taking $\lambda = 0.2$ in our case, according to the Central Limit Theorem the theoretical mean of the distribution of averages is also $\mu = \lambda^{-1}$, that is: $\mu = 5$. Now we compute the average of our simulated means:

```
simulatedavg <- round(mean(simMeans),3)
simulatedavg

## [1] 4.993
```

On the other hand, according to the Central Limit Theorem, the theoretical value of the variance of the distribution of averages from an exponential is $\frac{\lambda^{-2}}{n}$. In R we compute the theoretical and simulated variances as follows:

```
theorvar  <- lambda^(-2) / n
simvar <- var(simMeans)
theorvar

## [1] 0.625

simvar

## [1] 0.682878
```

We see that the mean from the averages (4.993) is quite similar to the theoretical mean ($\mu = 5$). Also the simulated variance and the theoretical variance are quite close.

3 Distribution

In figure 1 we plotted the histogram of the mean of our 1000 simulations. The vertical lines show how the theoretical mean of the distribution is almost identical to the simulated mean. The curves are the density lines corresponding to normal distributions with mean and standard deviation corresponding to our theoretical and simulated values. We readily see that the distribution closely resembles that of a normal distribution. Finally we verify this by plotting the Q-Q for quantiles (figure 2)to verify that the sample quantiles match the theoretical quantiles.

After examining the plots, we conclude that the distribution of means of simulated samples closely resembles that of a normal distribution.

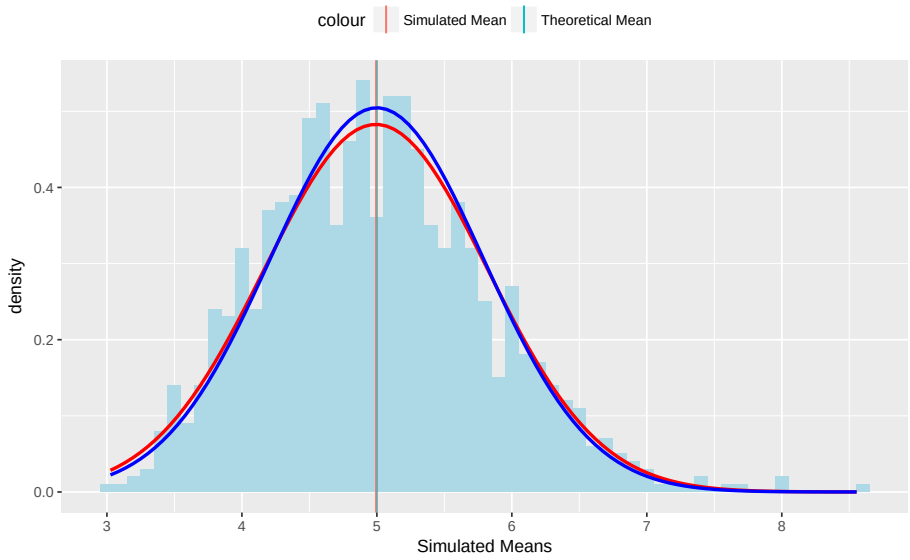

Figure 1: Distribution of the simulated means

```
# plot the means
mu = lambda^(-1)
ggplot(data = data.frame(simMeans), aes(x = simMeans)) +
geom_histogram(binwidth=0.1,
               aes(y=..density..),
               fill="lightblue") +
geom_vline(aes(xintercept=mu,
               color="Theoretical Mean"))+
geom_vline(aes(xintercept=simulatedavg,
               color="Simulated Mean"))+
stat_function(fun = dnorm,
              args = list(mean = simulatedavg, sd = sqrt(simvar)),
              color = "red", size = 1.0)+
stat_function(fun = dnorm,
              args = list(mean = mu, sd = sqrt(theorvar)),
              color = "blue", size = 1.0)+
xlab("Simulated Means")+
theme(legend.position = "top")
```

```
# use qqplot and qqline to compare the distribution of averages of 40 exponentials
# to a normal distribution
qqnorm(simMeans)
qqline(simMeans, col = 2)
```

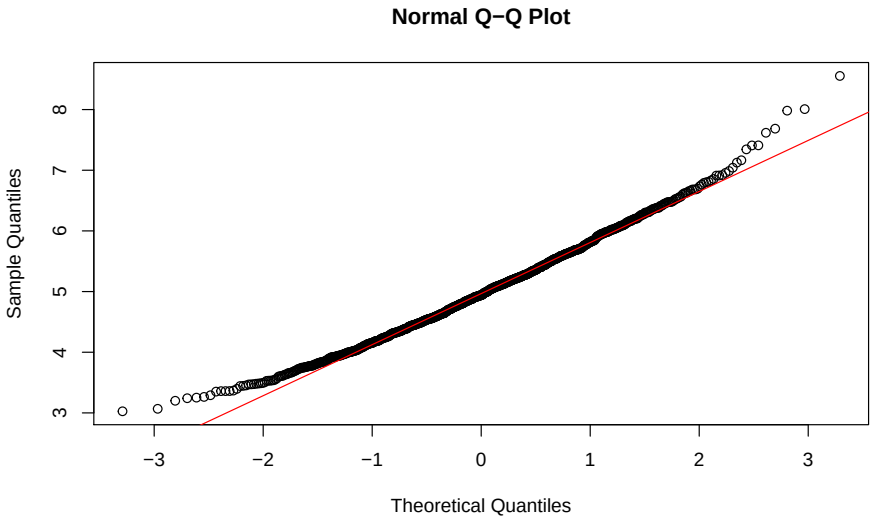

Figure 2: The points lie mostly on the line so the distribution has the same shape as the thoretical distribution