Statistical Inference Project Part 1

Simulating an Exponential Distribution and comparing it with the CLT

Allan R Brewer Cappellin

May 19, 2015

Introduction

In this part of the project we will report the simulation of an exponential distribution with a $\lambda = 0.2$ for the simulation. We will investigate how the distribution of the mean of 40 exponentials approximates the theoretical mean of the population. Since we are using a known value of λ we know the value of the theoretical mean of the population $\mu_{Theoretical} = 5$ and the the value of the theoretical variance of the population is $\sigma_{Theoretical} = 5$. The objective is to compare the distribution of the means to the theoretical mean $\mu_{Theoretical} = 5$ and the variance of the distribution to the theoretical variance $S_{Theoretical} = \frac{\sigma^2}{n} = \frac{25}{40} = 0.625$ for the sample distribution.

Simulation and Results

We first must simulate the 1,000 exponential of size 40. This simulations will give us 1,000 values of the mean for the 40 samples of the exponential. We have to compare this distribution to see if it follows the CLT and compare the mean and variance to the theoretical.

Simulations

The first part of the project is to simulate the 1,000 exponential. The code for the simulation can be found in the appendix of this document under the name Simulation R Code. This simulation will be done by first setting a seed and then a for loop from 1 to 1,000 for random exponentials of size 40.

The results of this simulation will be a vector of length 10,000 that has a mean and a variance to be compared with the theoretical.

Compare sample mean vs the theoretical mean

As i mentioned in the introduction the theoretical mean for the population $\mu_{Theoretical} = \frac{1}{\lambda} = \frac{1}{0.2} = 5$. With the vector created in the previous section we can determine the sample mean for the simulation which is:

Sample Mean

```
## [1] 4.963023
```

From the comparison we can see the values of the sample approximates almost exactly to the theoretical mean.

Compare the sample variance vs the theoretical variance

As mentioned before the square root of the theoretical variance of the distribution is known as the standard error of the distribution of the mean. This theoretical variance con be calculated with a known λ and it is $S^2 = 0.625$ shown in the formula from the introduction. The variance of the sample is calculated as the variance of the distribution of mean and is equal to:

Sample Variance

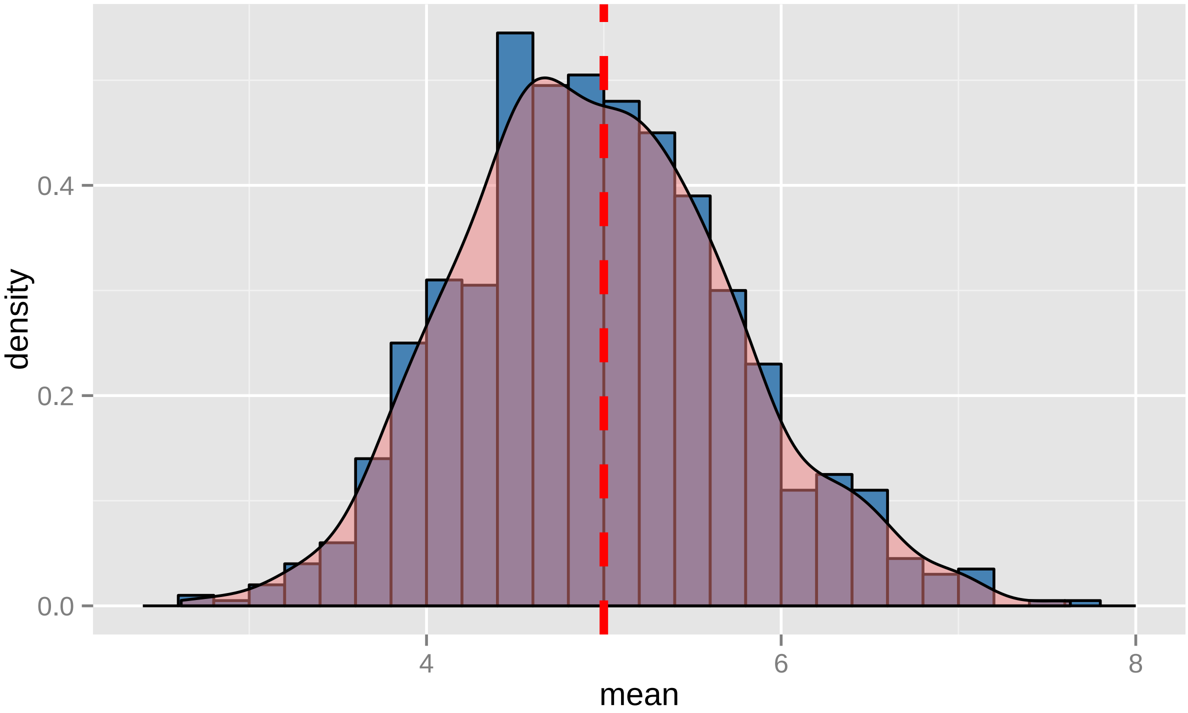
```
## [1] 0.6064594
```

In this comparison we can see that the sample variance approximates very closely to the theoretical value. The error is a bit bigger than with the mean because the value is squared.

Show that the distribution is approximately normal

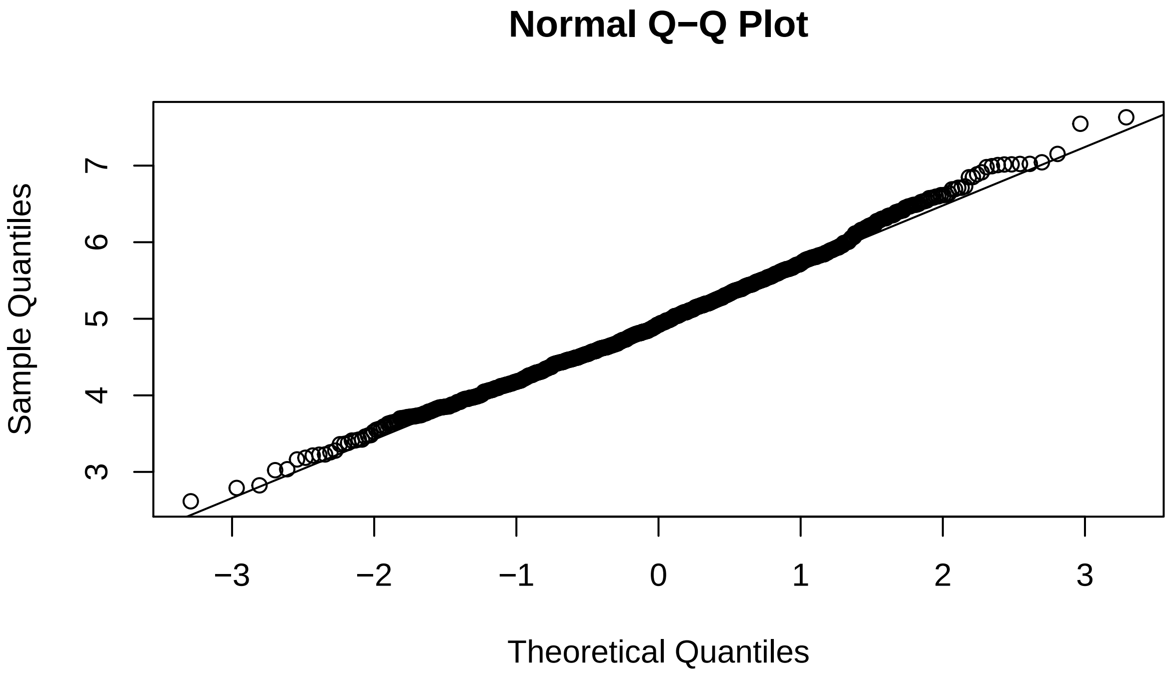
After comparing the sample mean and variance with the theoretical values we will compare the sample distribution with a normal distribution to determine if the CLT predicts the distribution of the 10,000 samples.

The first logical step to the comparison is to create a histogram and density plot of the 1,000 sample means to take a look of its distribution. The R code for the plot can be found in the appendix, the R code is under the name Histogram and Density Code.



From this plot we can see that the distribution of the sample means approximates to a normal distribution but it is not completely symmetrical. We can determine it is skewed to the right and that it is not completely centered around the theoretical mean in the dashed red line.

After creating the density and histogram we will create a plot for testing the normality of the distribution. This plots are called qqnorm and qqline. We will compare the values of our distribution to that of a normal distribution. When this values coincide with the line printed in the graph then we can say that the sample distribution is approximately normal. This R code can be found in the appendix.



This last graph creates a line that indicates normality and the plotted dots which in this case are in over the line in almost all of its length. We can see in the tails that the values are not approximating to a normal distribution but this is to be expected from this plot.

So for this section we can say that the distribution of means for the 1,000 simulations approximates a normal distribution very well for most of the samples.

Conclusions

From the simulation, R code and graphs we can conclude that the simulation of 1,000 samples size 40 approximates a normal distribution as predicted by the CLT. We can see from the first two sections of the body as the population mean and variance are almost identical to the theoretical value, this evidence is considered enough with the plots that the distribution of means of a random variable approximates to a normal distribution that is slightly skewed to the right, this is because as the population is exponential it can not take values lower than zero. This can be seen in the first plot of the appendix where the distribution is not totally symmetrical.

Appendix

In this section we will show all the R code, figures and graphs that support the document and conclusions drawn in it.

R code

In this first section of the appendix you can find all of the R code that is not shown in the main body of the document.

This first R code correspond to the simulation of the 1,000 mean values of samples of size 40.

R Code #1 : Simulation R Code

```r
## Set a variable with the number of simulations
nsims <- 1000
n <- 40
lambda <- 0.2
## Create an empty vector where the sample means will be stored.
simmeans <- NULL
## We do 1000 simulation where the mean of 40 exponentials are recorded
set.seed(2345) ## We first set the seed
for (i in 1:nsims) {simmeans <- c(simmeans, mean(rexp(n,lambda)))}
mn <- mean(simmeans)
sdv <- sd(simmeans)
var <- sdv^2
```

The succeeding R code corresponds to the creation of the histogram and dentistry plot of the sample mean distribution with a vertical line indicating the theoretical mean value.

R Code #2 : Histogram and Density Code

```r
simmeans <- as.data.frame(simmeans)
names(simmeans) <- "mean"
plot1 <- ggplot(simmeans, aes(x=mean)) + geom_histogram(aes(y=..density..), binwidth = 0.20, colour = "
plot1 <- plot1 + geom_density(alpha = 0.5, fill = "lightcoral")
plot1 <- plot1 + geom_vline(aes(xintercept=5), color = "red", linetype = "dashed", size =1.5)
plot1
```

The next R code if for the qqnorm and qqline graph for normality tests.

R Code #3 : qqnorm R Code

```r
qqnorm(simmeans$mean)
qqline(simmeans$mean)
```