Statistical Inference | Simulation Excercise

Alex Lee

10/3/2019

Overview

The goal of this project is to investigate the relationship between an exponential distribution (using R) and the Central Limit Theorom (CLR).

We know that the Central Limit Theorom can take in any probability distribution, given a well-defined mean and variance, and determine the mean. The mean is obtained by taking a random number of samples in the distribution and calculating the sample mean. The resulting distribution is a normal distribution where the mean should near equal to the original distribution mean.

Exponential Distribution is a probability distribution that determines the probabilities of events occuring given a specific time unit. This category of distribtion is (typically) asymmptotic in nature where the area under the curve be 1.

Simulations

The simulation will rely on R to generate a set of random distributions. 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. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a 1000 simulations.

The below R code initializes the simulation given the parameters above:

```
set.seed(5)
lambda     <- 0.2
n          <- 40
sim        <- 1000
exp_data   <- replicate(sim, mean(rexp(n, lambda)))
```

In this section, we set the variable lambda (or rate) to 0.2, then define the number of sample variables to 40, defin the number of simulations to run and, lastly, use rexp() function to generate our exponential data distribution and assign it to the exp_data data frame.

Sample Mean versus Theoretical Mean

According to CLT, the sample mean should be roughly close to the theoretical mean. We will use R to run simple mean calculations against the exp_data set and compare it to the theoretical mean (1/lambda).

```
th_mean    <- 1/lambda
sim_mean   <- mean(exp_data)

th_mean

## [1] 5

sim_mean

## [1] 5.043053
```

As expected, the simulated mean (CLT), given a sample set size of 40 and 1000 total samples does provide a very close mean approximation to the theoretical mean. In this case, the simulated mean returns a mean value of 5.04 while the theoretical mean gives us 5.0.

Sample Variance versus Theoretical Variance

CLT also provides a very close variance approximation when compared to the theoretical variance. In this case, we calculate the simulated variance to be 0.602 while the theoretical variance is 0.625.

```
sim_var    <- var(exp_data)
th_var     <- 1/(lambda^2*n)

sim_var

## [1] 0.6026047

th_var

## [1] 0.625
```

We expect a close variance because of the nature of CLT in how it produces a normal distribution of sample means, given a large enough sample set size and number of samples.

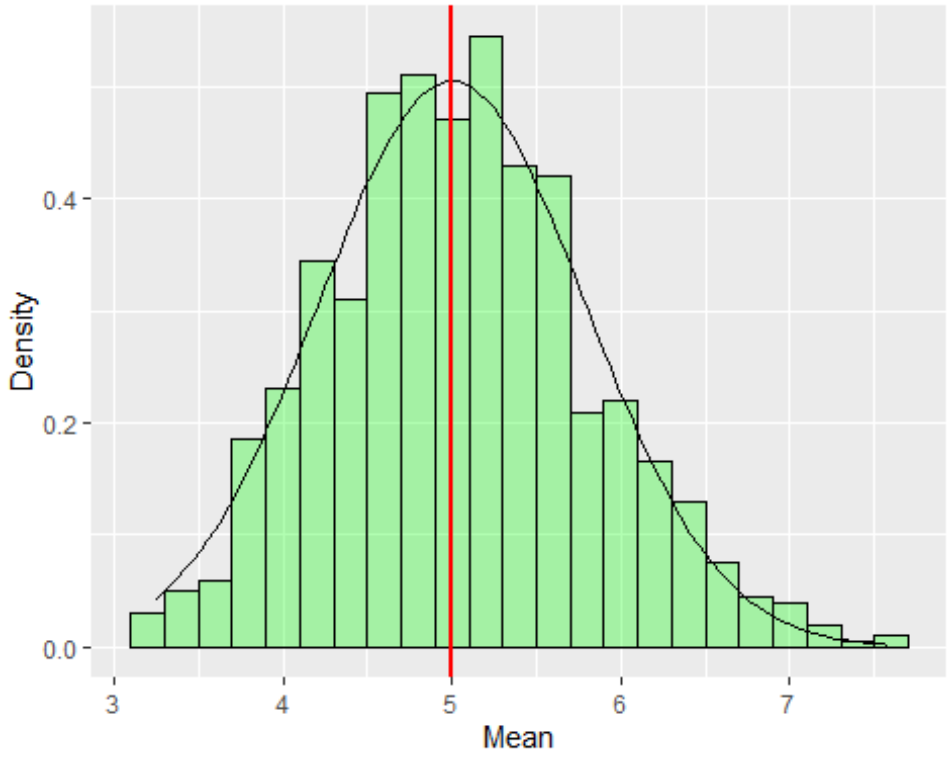
Distribution

Up until now, we only provided a numerical view of the results. To better see CLT in action, we need to generate a plot of the sample means as this relates to a normal distribtuion line. We will also need to calculate the theoretical standard deviation as part of this plotting excercise.

```
library(ggplot2)

th_sd   <- (1/lambda)/sqrt(n)

hist   <- ggplot(data.frame(exp_data), aes(x = exp_data))
hist   <- hist + geom_histogram(aes(y = ..density..), colour = "black", fill
= "green", alpha = .3,binwidth=.2)
hist   <- hist + stat_function(fun = "dnorm", args = list(mean = th_mean, sd
= th_sd))
hist   <- hist + geom_vline(xintercept=th_mean,size=1,colour="red")
hist   <- hist + xlab("Mean")+ylab("Density")
hist   <- hist + scale_x_continuous(breaks = c(1:10))
hist
```



In the figure above, we can see the CLT distribution from the green bars and compare this distribution against the theoretical mean (vertical) line. We can see that CLT is doing a pretty good job at estimating the exponential distribution mean, given the sample set size and sample count parameters.

Other observations from the figure is the tails of the distribution curve appears non-continuous (gaps). In order for CLT to produce a more “normal” distribution, increasing the number of iterations can help CLT avoid producing non-normal-looking distrubtions and avoiding gaps, skew and kurtosis.