Investigating the Central Limit Theorem Using the Exponential Distribution

Jason Foster

## Overview

The Central Limit Theorem (CLT) basically says that if independent samples of size n are repeatedly taken from any population, then when n is large the distribution of sample means will approach a normal distribution. In this report, we investigate the CLT by simulation using R to generate random independent samples of the exponential distribution.

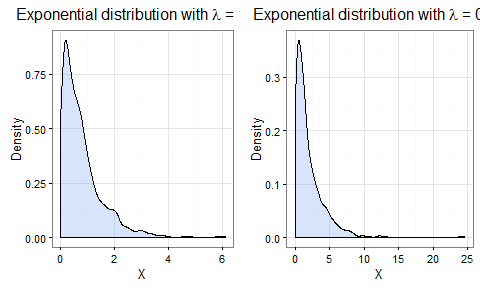
Before proceeding, add in the libraries we will be using in the report and set up some formatting.

suppressMessages(library(extrafont))  
suppressMessages(library(ggplot2))  
suppressMessages(library(gridExtra))  
suppressMessages(library(dplyr))  
  
windowsFonts(Avenir = windowsFont("Avenir Medium"))

### Exponential Distribution

The exponential distribution is defined as .

Two exponential distributions are plotted below with . Note that while they both have the same basic shape, the shape does not resemble a normal distribution. Notice, also that the higher the the more likely it is that a random variable X will have a small value, which makes sense, given that the mean and standard deviation of the exponential distribution is .



## Simulations

For this report we will perform 1000 simulations of 40 random exponentials with .

To generate our simulation data, we use the rexp function to generate a single random exponential distribution, replicate to repeat the generation 1000 times and t to transpose the observation matrix's columns into rows. Finally, data.frame is used to create the appropriate structure for ggplot. We then use dplyr:mutate() to add a column for each distribution's mean and the fluctuation around the theoretical mean using the following calculation

The following R code generates our simulation data.

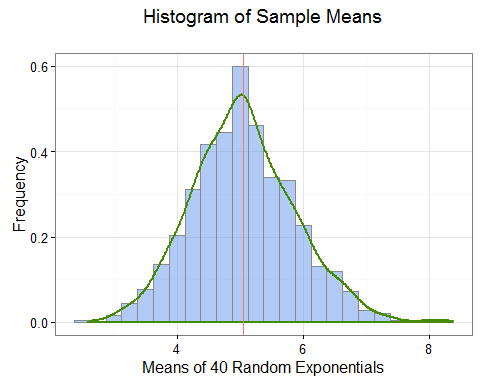
lambda <- .2  
n <- 40  
numSims <- 1000  
mu <- 1 / lambda  
simulations <- data.frame(t(replicate(numSims, rexp(n, lambda)))) %>%  
 mutate(xBar = rowMeans(.), flux = sqrt(n) \* (xBar - mu))

## Comparison of Theoretical and Sample Mean

If the CLT holds true, then what we should see is the distribution of the mean of our simulated 40 exponentials should resemble a normal distribution with mean .

The following plot shows the distribution of sample means along with its density curve. The red vertical line shows the distribution's mean.

ggplot(simulations, aes(x=xBar)) +  
 ggtitle("Histogram of Sample Means\n") +  
 labs(x = "Means of 40 Random Exponentials", y = "Frequency") +  
 geom\_histogram(aes(y = ..density..), col = "azure4", fill = "cornflowerblue", alpha = .5, binwidth = .25) +  
 geom\_density(color = "chartreuse4", size = 1) +  
 geom\_vline(aes(xintercept = mean(xBar), color = "firebrick"), show.legend = FALSE) +  
 theme\_bw(base\_family = "Avenir", base\_size = 12)



By inspection we can see that the plot resemble a normal distribution.

The theoretical mean of the exponential distribution is given by , so in our simulation, the theoritical mean of our distribution is 5. To see how this compares with the sample mean, let us calculate the sample mean with the following R code.

sampleMean <- mean(simulations$xBar)  
data.frame(Mean = c(sampleMean, mu), row.names = c("Sample", "Theoretical"))

## Mean  
## Sample 5.052522  
## Theoretical 5.000000

As you can see, the sample distribution mean of 5.0525224 is very close to the theoretical mean of 5.

## Comparison of Theoretical and Sample Variance

Next we compare the variance of the 1000 sample means with the theoretical variance for our sample. The theoretical variance for our sample is given by the formula

So, lets calculate the variance of our sample means with the following R code and compare it with the theoretical value

sigmaSq <- (1 / (lambda^2)) / n  
sampleVariance <- var(simulations$xBar)  
data.frame(Variance = c(sampleVariance, sigmaSq), row.names = c("Sample", "Theoretical"))

## Variance  
## Sample 0.6719342  
## Theoretical 0.6250000

As you can see, the theoretical variance of 0.625 is very close to the sample variance 0.6719342.

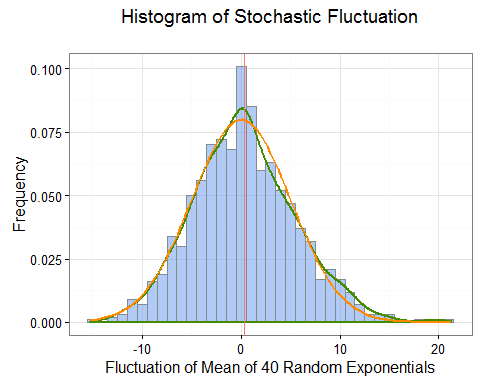
## Distribution vs. Normal Distribution

We have already seen in the above plots that the distribution of averages of 40 random exponentials resembles a normal distribution. The distribution of averages is also quite different from a distribution of even a large set of random exponentials, as we saw in the introduction. Lets take a look at some further comparisons of this distribution of means as it relates to the normal distribution.

According to Wikipedia, Classical CLT states that as n gets larger, the distribution of the difference between the sample average and its limit , when multiplied by the factor approximates the normal distribution with mean and variance .

The simulation data contains this random fluctuation in the variable flux so we can examine its distribution with the following R code, which shows the histogram of these fluctuations, along with its associated density curve. A normal distribution is also overlaid for comparison.

ggplot(simulations, aes(x=flux)) +  
 ggtitle("Histogram of Stochastic Fluctuation\n") +  
 labs(x = "Fluctuation of Mean of 40 Random Exponentials", y = "Frequency") +  
 geom\_histogram(aes(y = ..density..), col = "azure4", fill = "cornflowerblue", alpha = .5, binwidth = 1) +  
 geom\_density(color = "chartreuse4", size = 1) +  
 stat\_function(fun = dnorm, color = "darkorange", size = 1, args = list(mean = 0, sd = 5)) +  
 geom\_vline(aes(xintercept = mean(flux), color = "firebrick"), show.legend = FALSE) +  
 theme\_bw(base\_family = "Avenir", base\_size = 12)



As expected, this histogram resembles a normal distribution. Lets take a look at the mean and standard deviation of this distribution to see if it approaches what the CLT would assert.

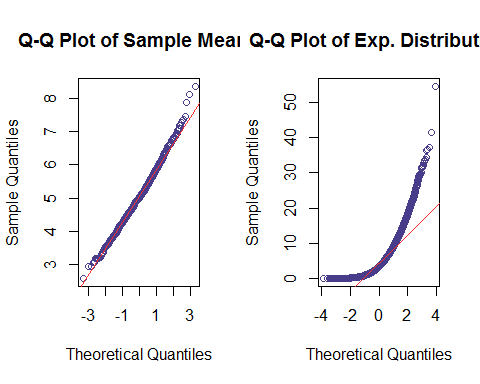
fluxMean <- mean(simulations$flux)  
fluxSd <- sd(simulations$flux)  
data.frame(Value = c(fluxMean, fluxSd), row.names = c("Mean", "Standard Deviation"))

## Value  
## Mean 0.3321809  
## Standard Deviation 5.1843389

From this we can see that the CLT's assertion about the distribution being normal with mean 0 is true, since 0.3321809 is close to zero.

One other method of comparing this distribution of means with a normal distribution is by using a Quantile-Quantile Plot, or q-q plot. A q-q plot plots the quantiles of the first data set against the quantiles of the second dataset and can be useful in determining if two data sets come from a population with a common distribution. In our case, instead of plotting with two different datasets, we can plot our sample distibution agains the the normal distribution by using the qqnorm and qqline in R. By way of comparison, we also display a q-q plot of a large random exponential distribution with and .

bigSim <- rexp(10000, .2)  
par(mfrow=c(1,2))  
qqnorm(y = simulations$xBar, col = "darkslateblue", main = "Q-Q Plot of Sample Means")  
qqline(y = simulations$xBar, col = "firebrick1")  
qqnorm(y = bigSim, col = "darkslateblue", main = "Q-Q Plot of Exp. Distribution")  
qqline(y = bigSim, col = "firebrick1")



We can see from these q-q plot that the quantiles of our sample distribution follow fairly closely with the theoretical quantiles from the normal distribution, indicating that it approximates the normal distribution. By way of comparison, the q-q plot of the large random sample exponential distribution does not follow nearly as closely.