A Simulation of an Exponential Distribution With an Evaluation Against the Central Limit Theorem (CLT)

barb dornseif - saoirsegirl

February, 2015

1 Overview

The Central Limit Theorem states that the distribution of averages (Means) of Independant and Identically Distributed (IID) variables becomes that of a Standard Normal as the sample size increases. To test this theorem, this paper will explore whether a specifically non-Normal Distribution - the Exponential Distribution - will, when samples are repeatedly simulated, produce sample Means that will result in a distribution that is Normal. We will first simulate several samples of varying sizes and then simulate the average of the Means and Standard Deviation of repeated samples of 40 random observations of an Exponential Distribution.

2 The Simulations

First we must understand the theoretical nature of the Exponential Distribution (ED) to understand how it is not Normal. The description of an Exponential Distribution in wikipedia is as follows:
"In probability theory and statistics, the exponential distribution (a.k.a. negative exponential distribution) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution, and it has the key property of being memoryless."

Further we are told that both the Mean and the Standard Deviation are described as $\frac{1}{\lambda}$. Thus they are equal. For this excercise we are given a $\lambda = 0.2$. Therefore the theoretical Mean as well as the Standard Deviation will be 5.

2.1 Simulation of Theoretical Distributions

Pictorially, graphs of the theoretical densities for an Exponential and Normal distribution where the mean is 5 and the standard deviation = 5 will look like this:

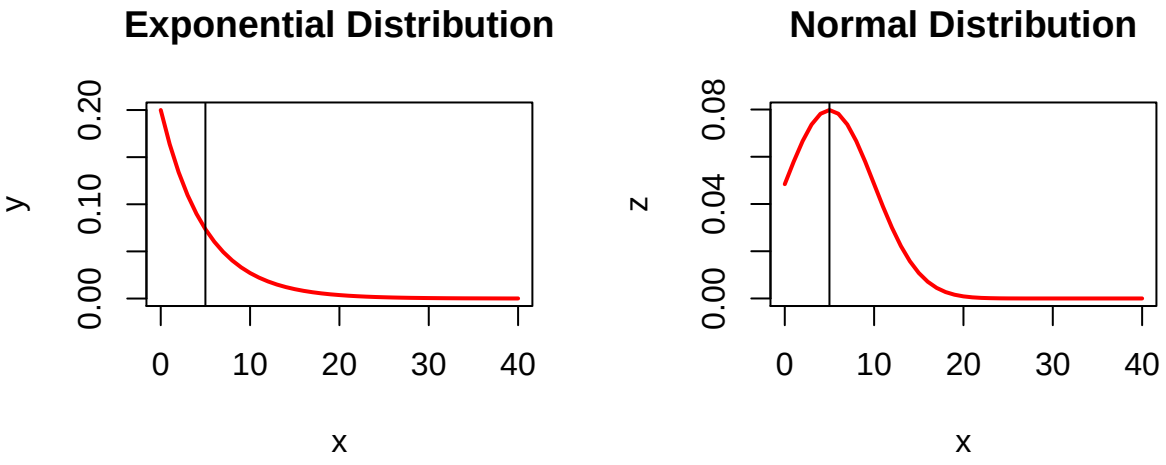

Figure 1 - Code Chunk 1(CC1)

As we can see in Figure 1, these two density plots are quite different, so we know that our simulation excercise of averages - when plotted - will show whether the distribution of sample Means and Standard Deviations will take on the Exponential or Normal distribution.

Knowing that the size of the sample is implicated in the CLT, first we will look at three samples of progressively larger sizes, n = c(40, 400, 4000) to see if the sample size itself alters or smoothes the distribution. We will plot a random sample using our given $\lambda = 0.2$ and the theoretical Mean and first Standard Deviation $\mu \pm 5$ to see the center-point and basic spread of the distribution and overlay our theorectical Exponential Destribution as defined above.

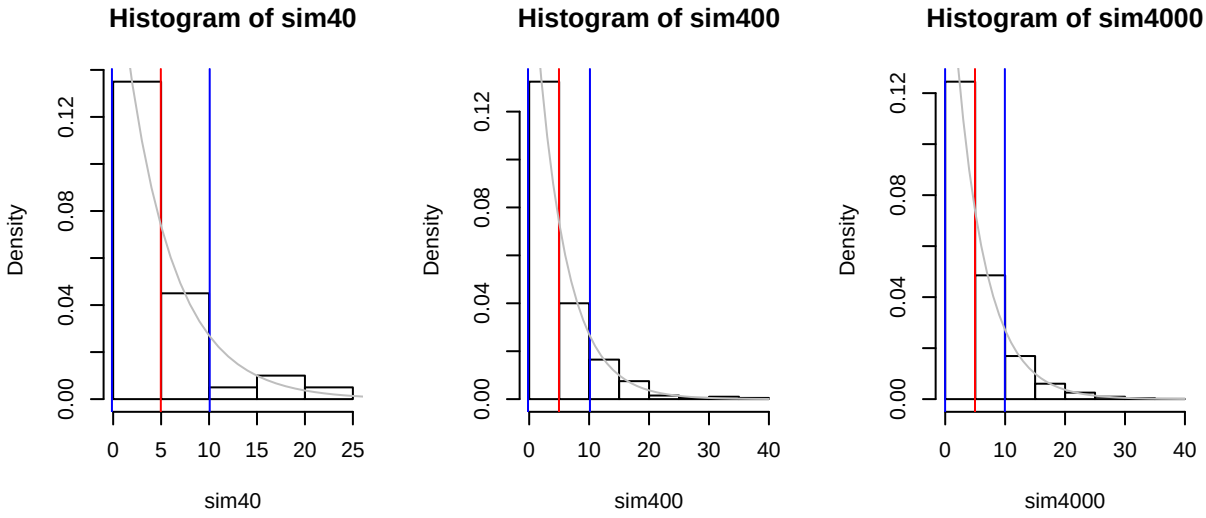

Figure 2 - Code Chunk 2(CC2)

In the figure above, we see that the Mean and Standard Deviation remain roughly independant of the number of observations in each plot. This might be better expressed in table form.

% latex table generated in R 3.1.2 by xtable 1.7-4 package % Sun Feb 22 00:10:24 2015

	n=40	n=400	n=4,000
Theoretical Means	5.00	5.00	5.00
Sample Means	4.97	4.97	4.97
Standard Deviation	5.10	5.17	4.99

Table 1 - Code Chunk 3(CC3)
What we observe is that whether the sample size is 40, 400 or 4,000, the distribution is clear and follows the theoretical distribution nicely. We also see that the Mean is constant as we used a seed value to fix the sampling results for replicability. Thus we should expect that the required n = 40 parameter for the simulation excercise should produce a reliable test. So let's proceed to our simulation of the required 1,000 samples and see if the resulting distribution of the Means from each of our 1,000 samples are Normally Distributed.

2.2 Simulation of Mulitple Samples

As we are given that our samples shall have a $\lambda = 0.2$, and sample size of 40, our next step is to run 1,000 such samples, calculate their Means, save those Means into a vector and plot their distribution to detirmine if they 1) distribute Normally and 2) whether the Mean of Means and the variance of those Means meets the theoretical expectation given by the Central Limit Theorem. Mathematically represented as:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

If we remember that replacing the standard error by its estimated value doesn't change the CLT a useful way to think about the CLT is that $\bar{X}_n$ is approximately $N(\mu, \sigma/\sqrt{n})$. So our expectation is that the Mean will be 5 and the Standard Deviation will be $5/\sqrt{n}$. Given n = 40 we expect 0.79. For simplicity of plotting we will subtract our Mean from 5 and plot the results against a Standard Normal Distribution with Mean = 0 to prove the Theorem.

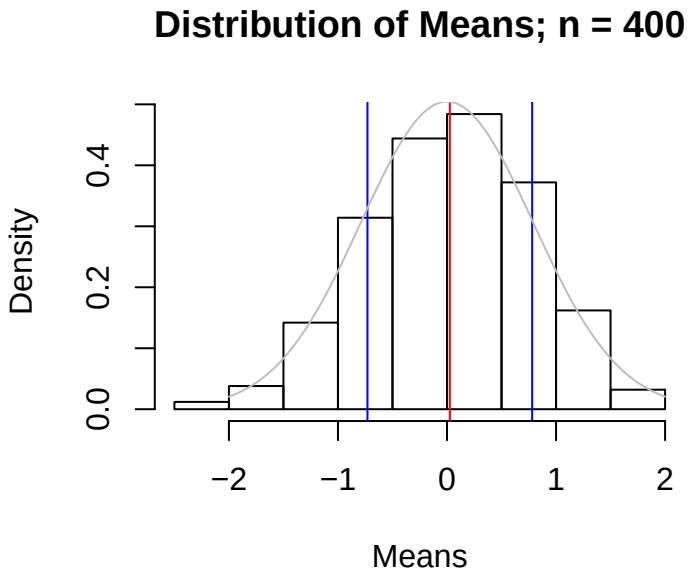

Figure 3 - Code Chunk 4(CC4)

Our simulation shows that the Mean of Means of 1,000 samples is is Normally Distributed and imperically the Mean is 0.026 with a Standard Deviation of 0.755 both of which are very close to our expectation. Further if we were to increase our number of observations to 400 and rerun the same code, we would see that the Mean is closer still and the Standard Deviation even closer to the theoretical expectation as our last plot shows.

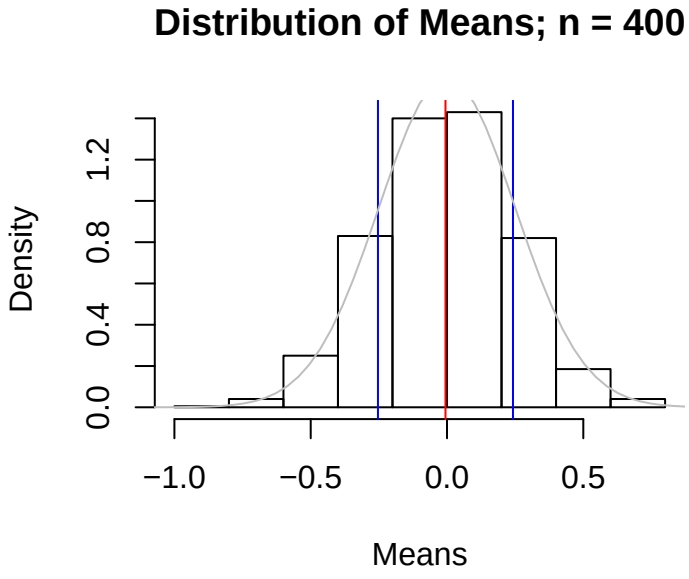

Figure 4 - Code Chunk 5(CC5)

Which is in fact what we observe; the Mean is -0.006 with a Standard Deviation of 0.248. Thus we have proven the CLT works on a distribution of non-Normal samples to provide a Normal Distribution of Means of those sample and that as the sample size grows the Normal Distribution will become more focused on the theoretical Mean of the Population that those samples are pulled from.

3 Appendix 1 – Code chunks for each data simulation, graph and table

CC1 - Load needed packages, set seed and universal variables. Create theretical density distributions.

```
library(ggplot2)
set.seed(1234)
lambda <- 0.2
n <- 40
iter <- 1000
x=seq(0,40)
y=dexp(x, lambda)
par(mfrow = c(1,2))
plot(x,y,type="l",lwd=2,col="red", main="Exponential Distribution")
abline(v=5)
z = dnorm(x,5,5) # normal dist mean = 5 and SD = 5
plot(x,z,type="l",lwd=2,col="red", main="Normal Distribution")
abline(v=5)
```

CC2 - Simulate samples of random Exponentials with n = c(40,400,4000)

```
# simulate n random exponentials & calculating the mean and standard deviation of each simulation
sim40 <- rexp(n, lambda); sim400<- rexp(n*10, lambda); sim4000<- rexp(n*100, lambda)
mean40 <- mean(sim40); mean400 <- mean(sim40); mean4000 <- mean(sim40)
std40 = sqrt(var(sim40)); std400 = sqrt(var(sim400)); std4000 = sqrt(var(sim4000))
# prepare graph layout
par(mfrow = c(1,3))
hist(sim40, freq=FALSE); abline(v=mean40, col = "red")
abline(v=mean40-std40, col="blue"); abline(v=mean40+std40, col="blue")
lines(x,y, col="grey")
hist(sim400, freq=FALSE); abline(v=mean400, col = "red")
abline(v=mean400-std400, col="blue"); abline(v=mean400+std400, col="blue")
lines(x,y, col="grey")
hist(sim4000, freq=FALSE); abline(v=mean4000, col = "red")
abline(v=mean4000-std4000, col="blue"); abline(v=mean4000+std4000, col="blue")
lines(x,y, col="grey")
```

CC3 - Present Table of Means and Standard Deviations

```
library(xtable)
Theoretical <- c(5, 5, 5)
Means <- c(mean40, mean400, mean4000)
SDev <- c(std40, std400, std4000)
table1 <- rbind(Theoretical, Means, SDev)
rownames(table1) <- c("Theoretical Means", "Sample Means", "Standard Deviation"); colnames(table1) <- c
xtable(table1, format = "markdown")
```

CC4 - Mean of Means: n=40

```
set.seed(1234)
Means = NULL
for (i in 1 : iter) {
```

```
    Means = c(Means, 5 - mean(rexp(n,lambda)))}
Mean1000 <- mean(Means)
STD1000 <- sqrt(var(Means))
hist(Means, freq = FALSE, main = "Distribution of Means; n = 400")
abline(v = Mean1000, col = 'red')
abline(v = Mean1000 - STD1000, col = 'blue')
abline(v = Mean1000 + STD1000, col = 'blue')
x=seq(-2,2, length=100)
s = dnorm(x, 0, 1/lambda/sqrt(n)) # normal dist mean = 5 and Stan error = 5 /sqrt(n)
lines(x,s, col='grey')
```

CC5 - Mean of Means: n=400

```
set.seed(1234)
n = 400
Means = NULL
for (i in 1 : iter) {
    Means = c(Means, 5 - mean(rexp(n,lambda)))}
Mean1000 <- mean(Means)
STD1000 <- sqrt(var(Means))
hist(Means, freq = FALSE, main = "Distribution of Means; n = 400")
abline(v = Mean1000, col = 'red')
abline(v = Mean1000 - STD1000, col = 'blue')
abline(v = Mean1000 + STD1000, col = 'blue')
x=seq(-2,2, length=100)
s = dnorm(x, 0, 1/lambda/sqrt(n)) # normal dist mean = 5 and Stan error = 5 /sqrt(n)
lines(x,s, col='grey')
```