Inferential Statistics - Simulation Project

Joe Pechacek

July 22, 2015

Overview

The purpose of this report is to show through the use of R Statistics how the Sample Means and Variance for a large number of Exponential Distributions will converge on an Expected Value as more simulations are completed. Furthermore, the distribution of the Sample Means will approximate a Normal Distribution as predicted by the Central Limit Theorm.

Simulations

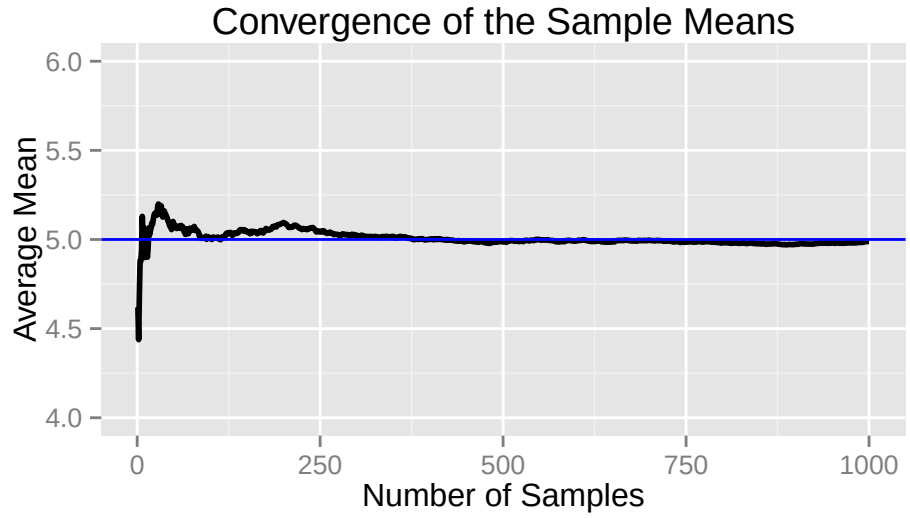
We will be using the Exponential Distribution with a λ equal to 0.2. Each sample will include 40 values for which the Sample Mean will be calculated and stored. As additional interations are completed, a running average of the Sample Means will also be calculated. The same process will be used for the Sample Variances. We will collect 1,000 samples and these samples will then be used to run comparisons to Expected Values for the Mean and Variance. See the Appendix for the code used to generate the 1,000 simulations.

Sample Mean vs. Theoretical Mean

The Theoretical Mean for this simulation as defined by the Central Limit Theorm would be equal to the Mean of the Population that the samples were collected from. In this case, the samples were taken from the Exponential Distribution with a $\lambda = 0.2$. The Mean of this distrbution is $\mu = 1/\lambda$ which would result in an Expected Mean of 5.

The Sample Means collected ranged from 2.97 to 7.85 with a Mean value of 4.99. It was noted that some samples had a Mean that was as much as +/- 2.9 from the Expected Mean, however the central tendency of the data remains very close to the Expected Mean of 5. See Graph 1 in the Appendix.

The following graph shows how over time, the Average of the Sample Means will converge on the Expected Value of 5. As can be seen, the first 100 or so samples show a bit of noise, but after that the line smooths out to be very close to the Expected value of 5.

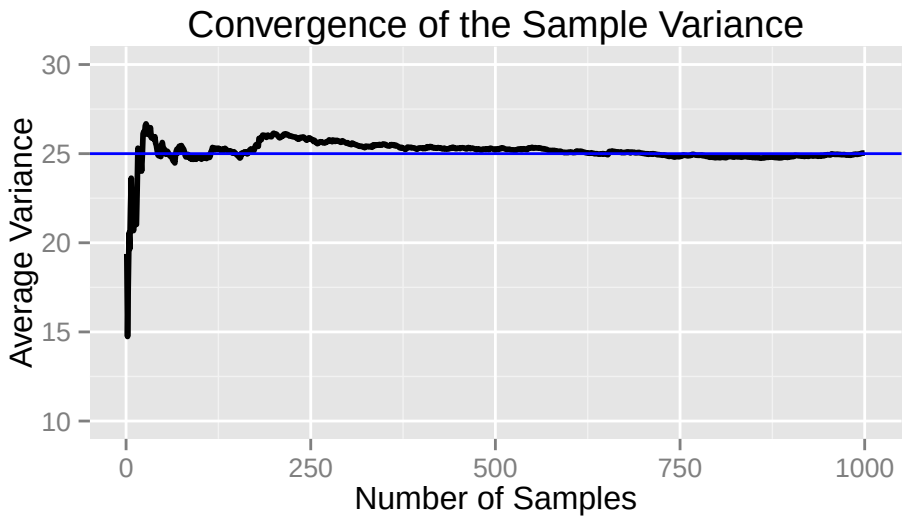


Sample Variance vs. Expected Variance

The Expected Variance for this simulation should again follow the Variance of the Population. In the case of the Exponential Distribution, the Variance is defined as $(1/\lambda)^2$, so we should expect the Variance in each sample to be $(1/\lambda)^2$ or 25.

Using the sample data from the simulations, the Variance of each sample ranged from 6.57 to 122.38 with a Mean value of 25.03. It was observed that the Variance will range by as much as +/- 97.3 from the Expected Variance, however the central tendency was very close to the Expected Variance of 25. This can be seen in Graph 2 of the Appendix.

As we noted with the Sample Means, the Average of the Sample Variance should also converge on the Expected Value as more simulations are completed. The following graph shows that this is true.



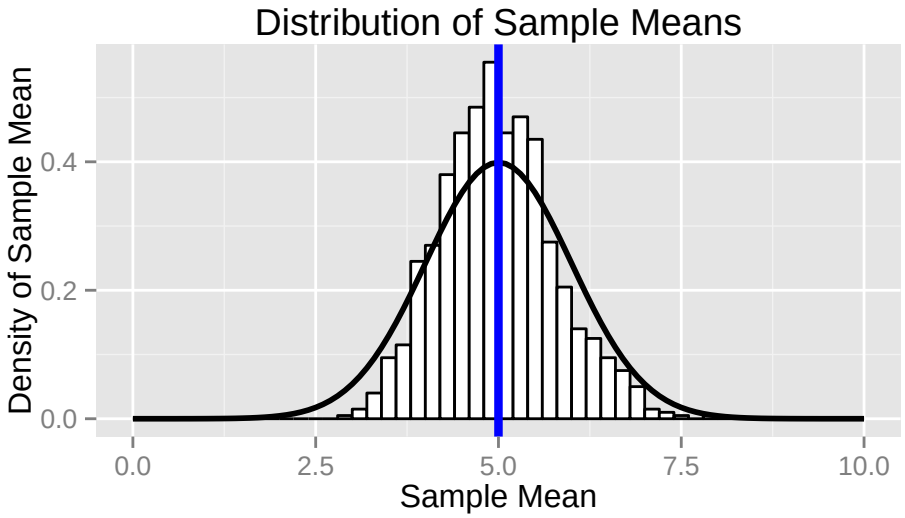
Distribution

According to the Central Limit Theorm, we should expect that the distribution of the Sample Means should approximate the Standard Normal distribution or 'bell curve'. The Standard Normal Distribution can be characterized as having values equally distributed about the Mean and should approximate a bell shape. To see if this is true for our simulation, we can use a couple different methods. First, if we call `summary()` on our simulated data, we can see that the Median and Mean are very close to each other indicating that 50% of the data is greater than and less than the Mean value. We can also look at the relationship between the 1st and 3rd quartiles as they should be about equal distant from the Mean.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.972   4.445   4.949   4.988   5.475   7.850
```

We can see that the Mean and Median are close in value. The difference between the Mean and the 1st Quartile is 0.543 and the difference between the Mean and the 3rd Quartile is 0.487 so the fact that these values are nearly the same indicates symetry.

However, since a picture is worth a 1,000 words, the histogram on the following page shows that the distribution of the Sample Means follows the Standard Normal and is approximately "bell shaped."



It can be seen in the histogram presented that the data is centered on the Expected Mean of 5 and there is symetry about the mean. To further empahsize this relationship, a Standard Normal Distribution has been layered over the histogram.

Summary

The simulations and analysis performed in this report do support the Central Limit Theorm in that both the average of the Sample Means and the average of the Sample Variances converged on the expected values as more simulations were added to the analysis. Additionally, as shown in the preceeding graph, the close relationship between the histogram and the overlay Standard Normal distribution supports the Central Limit Theorm predictions that the average of Sample Means will approximate the Standard Normal.

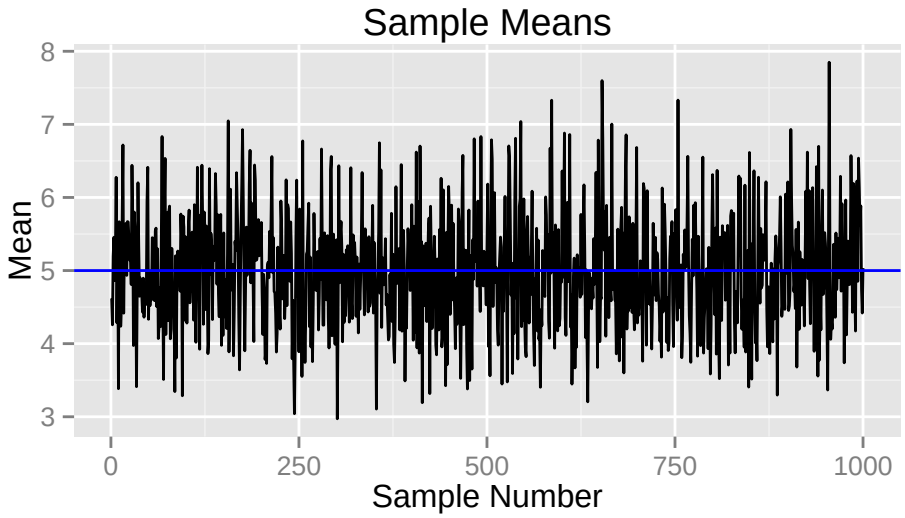
Appendix

The following R code was used to generate the 1,000 Sample Means and Variances values. A `for()` loop was used to repeat a function to calculate the Mean for 40 random values that follow the Exponential Distribution with a λ of 0.2. Each iteration will add another set of values until we have 1,000 samples. The same is repeated for the Variance of each sample by creating a `datavar` object. Note the use of `set.seed()` to ensure the same random sample is used for both the Mean and Variance simulations.

```
set.seed(135)
datamean <- data.frame(sample = 0, mean.value = 0, avg.mean = 0)
for (i in 1:1000){
     datamean[i, 1] <- i
     datamean[i, 2] <- mean(rexp(40, 0.2))
     datamean[i, 3] <- mean(datamean$mean)
}

set.seed(135)
datavar <- data.frame(sample = 0, variance = 0, avg.var = 0)
for (i in 1:1000){
     datavar[i, 1] <- i
     datavar[i, 2] <- var(rexp(40, 0.2))
     datavar[i, 3] <- mean(datavar$var)
}
```

Graph 1 showing the actual Sample Means collected during the simulations:



Graph 2 showing the actual Sample Variances collected during the simulations: