Statistical Inference: Project - Simulation Exercise

Nagesh Madhwal

October 19, 2015

OVERVIEW

The purpose of this report is to investigate the exponential distribution in R and compare it with the Central Limit Theorem. We are going to simulate an exponential distribution and analyze its mean & variance. Comparison will be done between the throretical & sample values and we will show that for a thousand simulations the distribution of means approximates the normal distribution.

Data, Output and Simulation

The exponential distribution which can be simulated in R with rexp(n, lambda) has a mean of lambda and standard distribution of lambda. We have been asked to simulate the distribution averages with lambda = 0.2, n = 40 and a thousand simulations need to be done.Theoretical values for the exponential distribution: Mean = 5, Standard Deviation = 5, Variance = 25. The R output from mean, var & sd by simulating an exponential distribution with lambda = 0.2 & n = 40 is shown below:

```
## Mean = 5.335801 SD = 5.663525 Variance = 32.07552
```

*All R code is provided in the Appendix

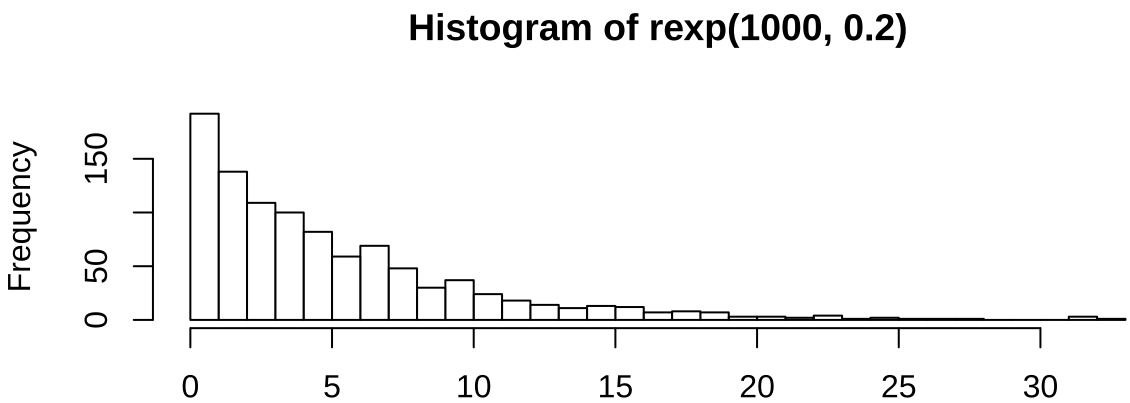
For a sampling distribution of sample means with n = 40 the theoretical values will be: Mean = 5, Standard Deviation = Population SD / sqrt(n) = 0.79, Variance = Population Variance / n = 0.625. The R output for mean, var & sd by simulation a set of 40 distribution means for the exponential distribution with lambda = 0.2 & n = 40 is shown below

```
## Mean = 5.070883 SD = 0.9569752 Variance = 0.9158015
```

For the purpose of our analysis we will do a thousand simulations to generate the distribution of means. This is equivalent to taking random samples of 40 values from an exponential distribution and tabulating their mean, repeating the process a thousand times to obtain a thousand values of the mean (of a sampling distribution of size 40).

LLN and CLT

Our analysis we will be relying on the " Law of Large numbers (LLN)" and the "Central Limit Theorem (CLT)". Law of Large Numbers states that as the sample size grows the average of sampling distrubtions converge towards the population mean.The Central Limit Theorem states that as the sample size increases the distribution of sample averages approximates the Standard Normal.
LLN & CLT require that the variables are "Independent & Identically Distributed (iid)". The variables should have equal probability of occurance and they are mutually independent. Our purpose here is to see how well our simulated data follows LLN & CLT.
For higly skewed populations like our exponential distribution (see figure below) large sample sizes are required for close approximation to standard normal.



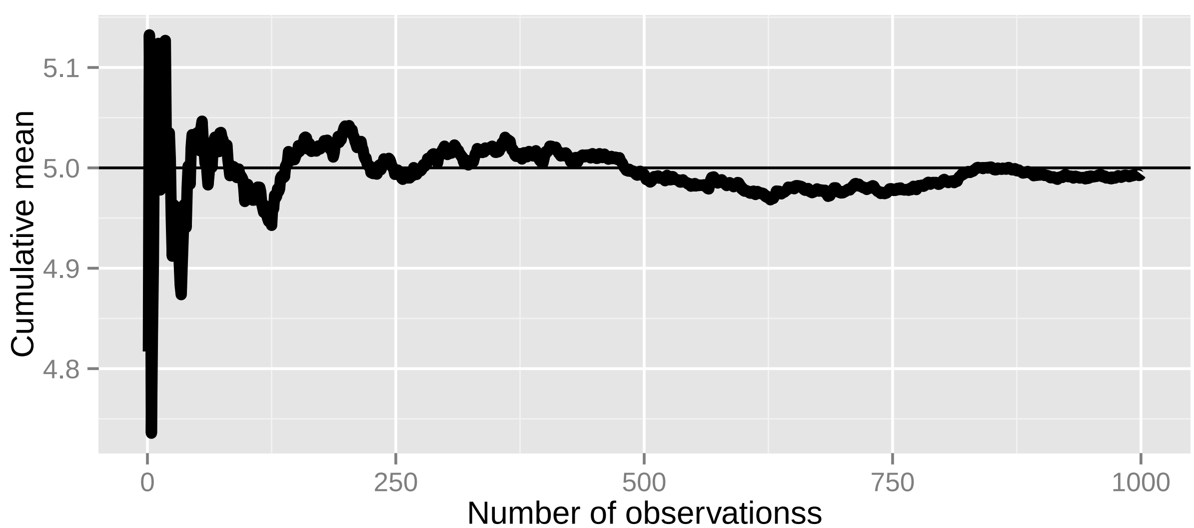
Comparison of Sample Mean & Theoretical Mean

The LLN says that the sample mean of iid sample is converges to the population mean. We have already checked the conditions for iid. Comparision between Theoretical mean & sample mean from 10, 100, 1000 simulations is given below:

```
## Theoretical mean = 5, Simulated Sample Mean(1000 simulations) = 4.979631
```

```
## Mean from 10 Simulations = 5.024898 Mean from 100 Simulations = 5.000532
```

The plot below illustrates how our simulated distribution behaves as the number of obervations increases.



The y-intercept equal to 5 represented by the horizontal line in the plot, shows the Theoretical mean & we can see that as the number of observations increases the sample mean is convergeing towards the theoretical mean.

Comparison of the Sample Variance & Theoretical Variance

The difference between Theoretical Variance & Sample Variance for just one simulation of 40 samples was significant. For a thousand simulations the numbers are much closer as seen below:

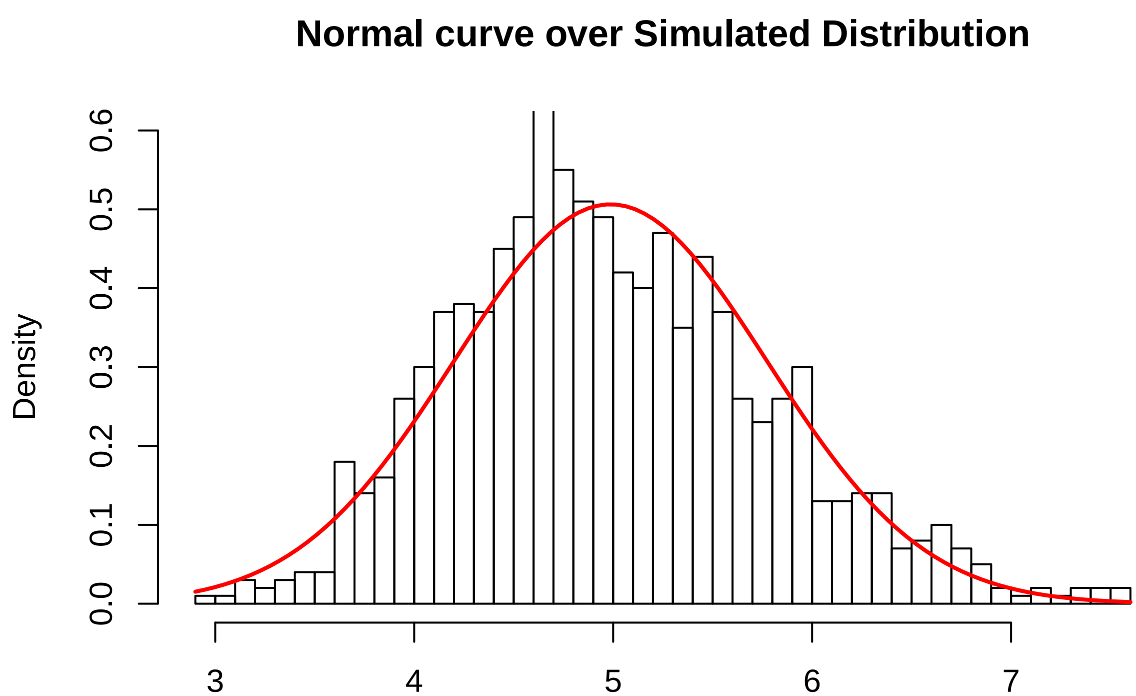
```
## Theoretical Variance = 0.625, Simulated Sample Variance = 0.6216855
```

```
## Variance from 10 Simulations = 0.3289488 Variance from 100 Simulations = 0.6830934
```

With a larger number of simulations the mean starts to converge towards the theoretical mean resulting in the variance converging towards the theoretical variance.

Comparison to Normal Distribution

Figure below shows the normal distribution imposed on the simulated distribution:




```
## Two sigma limit 3.410922 6.562336  vs 95% ci  3.653514 6.652944
```

As we can see from the curve & the matching the 95% limits to calculated two sigma limits the our simulated distribution is closely approximating the Normal distribution.

Appendix 1 - References

1. Course Material on Asymptotics : https://github.com/bcaffo/courses/blob/master/06_StatisticalInference/07_Asymptopia/index.md

2. Vatrious Wiki Pages on definitions

3. OpenIntro Statistics - Third Edition: David M Diez, Christopher D Barr & Mine C. etinkaya-Rundel

4. Various Stack overflow questions for code clarifications etc.

R Code

Code used for simulating the data

```
g <- replciate(1000, mean(rexp(40, 0.2)))
## Calculates the mean of the exponential distribution of length 40, one thousand
## times & stores the values in g
```

Code to plot show how the cumulative means converge to the theoretical mean

```
suppressWarnings(library(ggplot2))
means <- cumsum(replicate(1000, mean(rexp(40, 0.2)))) / (1  : 1000)
g <- ggplot(data.frame(x = 1 : 1000, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 5) + geom_line(size = 2)
g <- g + labs(x = "Number of observationss", y = "Cumulative mean")
g
```

Code to plot the histogram & superimpose the normal curve:

```
g <- replicate(1000, mean(rexp(40, 0.2)))
  m <- mean(g)
  std <- sd(g)
  hist(g, breaks=36, prob=TRUE,
     xlab="", ylim=c(0, 0.6),
     main="Normal curve over Simulated Distribution")
   curve(dnorm(x, mean=m, sd=std),
       col="red", lwd=2, add=TRUE)
ci <- quantile(g, probs = c(0.025, 0.975))
cat("Two sigma limit", m - 2* std,  m + 2* std, " vs 95% ci ", ci, "\n")
```

There is no attempt to set seed as every new simulation will adhere to the same results & the work is completely reproducible even without having to fix the algorithm for simulation.