Simulated proof of Central Limit Theorem as applied to Exponential Distributions

Rohit Padebettu

September 4, 2016

Overview

The goal of this exercise is to verify via simulations, the truth of predictions of Central Limit Theorem as applied to a raw Exponential Distribution. The mean of the exponential distribution is (1/lambda) and its standard deviation is also (1/lambda). The CLT states that, the mean of a distribution of sample means from ANY underlying distribution should approximate the distribution mean and that such a distribution of sample means is approximately normally distributed. The variance of such a distribution of means should also approximate the variance of the distribution divided by the sample size.

Simulations

In this analysis we run 1000 simulations to generate a distribution of means. Each mean is computed on a sample size of 40 drawn from the underlying exponential distribution with lambda set at 0.2. We proceed to analyze the mean and variance of this distribution of means and compare it to the theoretically expected result. We plot such a distribution of means to show it approximates a normal distribution and compare it to the plot of random samples drawn from the underying Exponential distribution

1 . We begin by loading the required libraries and initializing the rate parameter lambda, the numbers of simulations sims and sample size pop. We then proceed to calculate the mean, standard deviation and variance of the raw Exponential Distribution

```
#load libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
# Exponential Distribution Simulation
lambda <- 0.2
sims<-1000
pop <-40
#Distribution Stats
mean_dist <- (1/lambda)
sd_dist <- (1/lambda)
var_dist <- sd_dist^2
```

2. We begin our simulations by drawing samples of length pop for each simulation using the `rexp()`function in R. We draw `sims` such samples and store it in a 1000 x 40 matrix named `samp_exp_m`.

```
# sample sims*pop random numbers from the distribution
exp_data<-rexp(sims*pop,lambda)
#Store it in a matrix with dimensions 1000 x 40 with each of 1000 rows of 40
elements each representing a sample
samp_exp_m <- matrix(exp_data,sims,pop)
```

3. We calculate the mean of each row of the matrix and store it in a `means_samp_vect` array, which gives us the distribution of means. We have 1000 such means each corresponding to a simulation. We then proceed to calculate the Mean,standard deviation and variance of this distribution of means

```
# Generate Distribution of sample means - 1000 such means computed from 40
element samples
means_samp_vect <- apply(samp_exp_m,MARGIN = 1,FUN = mean)
# Stats for the distribution of sample means
mean_simulate <- mean(means_samp_vect)
sd_simulate  <- sd(means_samp_vect)
var_simulate <-sd_simulate^2
```

Predictions by CLT

We proceed to calculate the means, standard deviation and variance for such a distribution as predicted by CLT.

```
mean_clt_theory <-1/lambda              # Expected Value of mean of samples
approximates mean of distribution
sd_clt_theory <- 1/(lambda*sqrt(pop))   # Standard Deviation of sample is
sigma/sqrt(N) where N is sample size
var_clt_theory <- 1/((lambda^2)*pop)
```

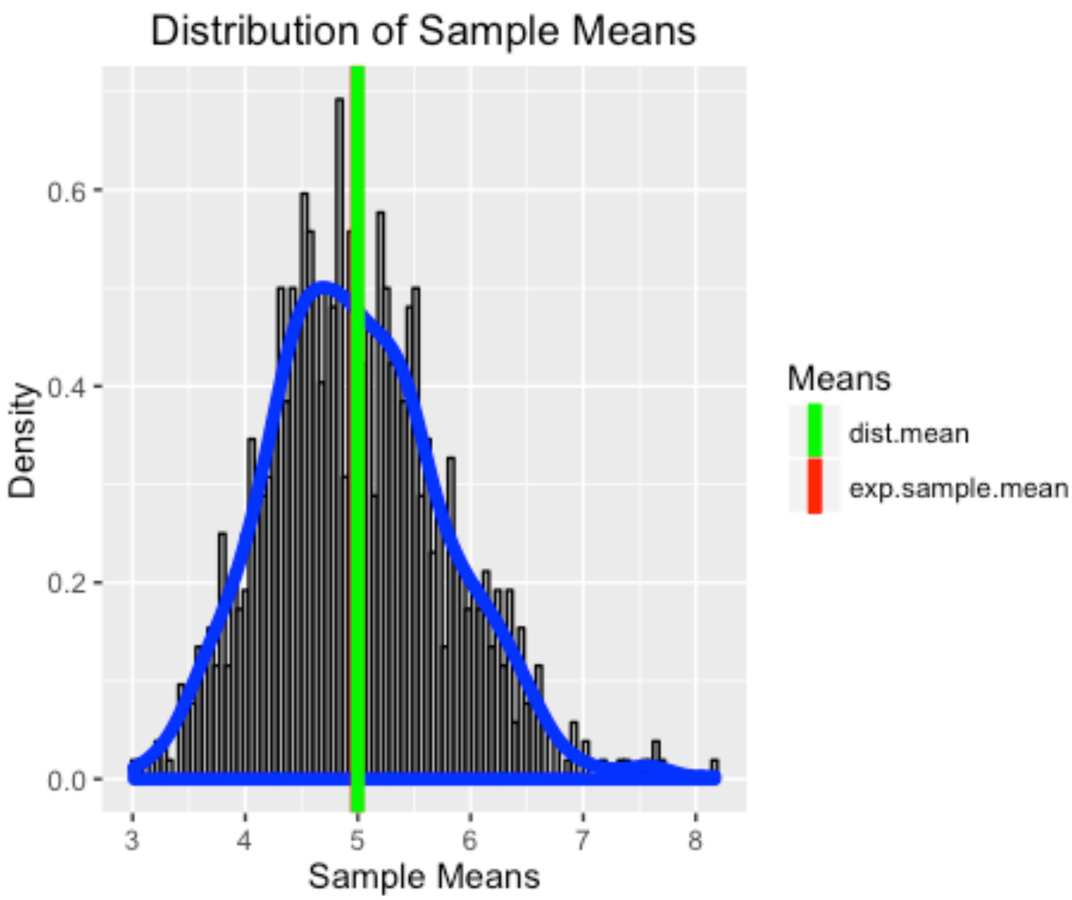
Sample Mean versus Theoretical Mean

CLT predicts that the expected value of a distribution of sample means drawn from an exponential distribution, should approximate the mean of the exponential distribution given by (`1/lambda`). We use ggplot2 to plot the distribution of means as a density histogram. We overlay the red vertical line to show the expected value of the mean of this distribution as calculated and stored in the variable `mean_simulate` above. We also plot the theoretically expected value of the mean in green given by `mean_clt_theory` above

```
MeanDist <- ggplot(data = as.data.frame(means_samp_vect),mapping = aes(x =
means_samp_vect, y=..density..))+
    geom_histogram(fill = 'grey',bins = 100, color= 'black') +
    geom_density(size =2, color="blue")+
    labs(title = "Distribution of Sample Means",x = 'Sample Means',
y='Density')+
```

```r
    geom_vline(aes(xintercept = mean_simulate, colour= "exp.sample.mean"),
size = 2,show.legend = TRUE)+
    geom_vline(aes(xintercept = mean_clt_theory,color= "dist.mean"), size =
2,show.legend = TRUE)+
    scale_color_manual(name="Means",
values=c(exp.sample.mean='red',dist.mean='green'))

print(MeanDist)         # plot of distribution
```



The individual means are as follows

```
## [1] "Expected Value of Distribution of Sample Means"

## [1] 4.992317

## [1] "Expected Value of Distribution of Sample Means as predicted by CLT"

## [1] 5

## [1] "Mean of underlying Exponential Distribution"

## [1] 5
```

Sample Variance versus Theoretical Variance

The variance of the exponential distribution is given by `1/lambda^2`. However CLT states that the variance of a distribution of sample means is given by sigma^2/N where sigma is the standard deviation of the population and N is the sample size. In our case the sample size if 40 as given by `pop`. We compare the variance computed on the distribution of sample means given by `var_simulate` against the CLT predicted variance given by `var_clt_theory`. We print this along with the variance for the underlying distribution given by `var_dist` below

```
## [1] "Variance of distribution of sample means"

## [1] 0.618825

## [1] "Variance of distribution of sample means as predicted by CLT"

## [1] 0.625

## [1] "Variance of underlying Exponential Distribution"

## [1] 25
```

Comparing the underlying distribution to the distribution of sample means

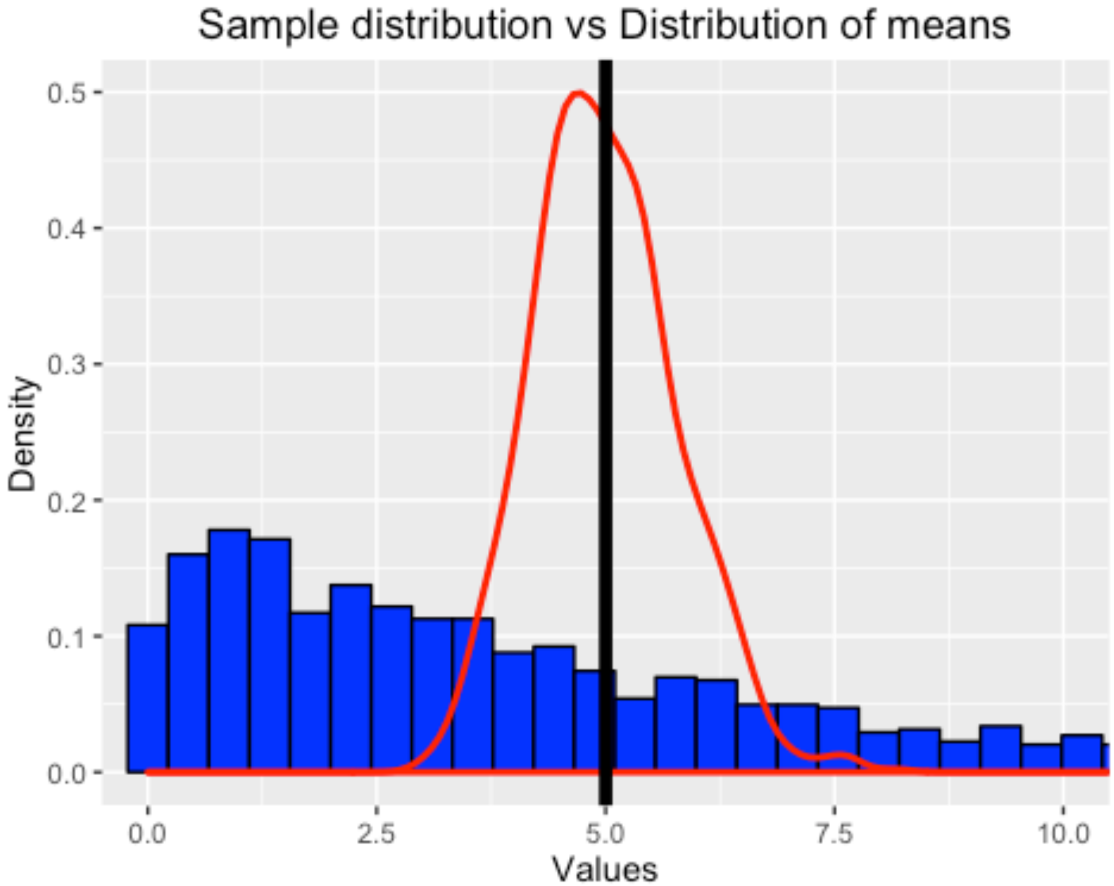
1. Plotting the underlying exponential distribution

In order to plot the exponential distribution, we randomly sample 1000 numbers from the exponential distribution and plot it as a histogram. We also compute and overlay the mean of this distribution in red as compared to the theoretical mean given by `1/lambda` in `green`

2. Overlaying the distribution of sample means to show Normality

We finally overlay a red density line showing the approximate Gaussian shape of the distribution of sample means as compared to the raw exponential distribution itself. The black vertical line centered at value 5 is the theoretical mean as predicted by CLT.

This plot shows that the distribution of means is approximately normal, centered at the population mean of `1/lambda`



Conclusion

From the simulations we ran and the distributions we plotted, we can clearly see that the distribution of sample means from an exponential distribution is normally distributed. We have also shown that the expected value of such a distribution of means approximates the underlying distribution mean as predicted by the CLT.