Statistical Inference Project Part 1

Rich Robinson

Monday, May 18, 2015

This part of the project explores the exponential distribution and compares the exact form provided by R functions with the approximated form provided by the Central Limit Theorem. Comparisons are made betweek the theoretical values for the mean and variance, calculated from the details given about our distribution, and those calculated from our simulated sample data.

Defining the attributes of our distribution. These are:

- lambda - the exponential rate
- mu - theoretical mean
- sd - theoretical standard deviation
- n - number of samples, values or distributions
- nosim - number of simulations/distributions

```
lambda <- 0.2
mu <- 1 / lambda
sd <- 1 / lambda
n <- 40
nosim <- 1000
```

Simulations

To create our sample distribution to attain comparisons between the theoretical values of mean and variance against those for our sample, a single large sample of `n` X `nosim` variables was generated using the `rexp()` function in `R` with the parameter λ as defined above. This single distribution was then subdivided into a `nosim` by `n` matrix to create our 1000 individual samples.

To illustrate that our sample is approximately a Normal disribution although it is actually an Exponential distribution required using the Central Limit Theorem which states that the means of small samples from a larger population are distributed like a normal distribution and as the sample size of these groups increases, then the behavior of the means becomes more close to a notmal distribution. This was illustrated by visual comparison against a normal distribution characterised by the same parameters as our exponential distribution (just to make the comparison clearer). The `rnorm()` function with parameters n=40, μ=5 and σ=5 was called 1000 times, with the mean from each distribution stored in a numerical vector array.

To ensure that all these simulations are reproducible, the `set.seed()` function was executed at the beginning.

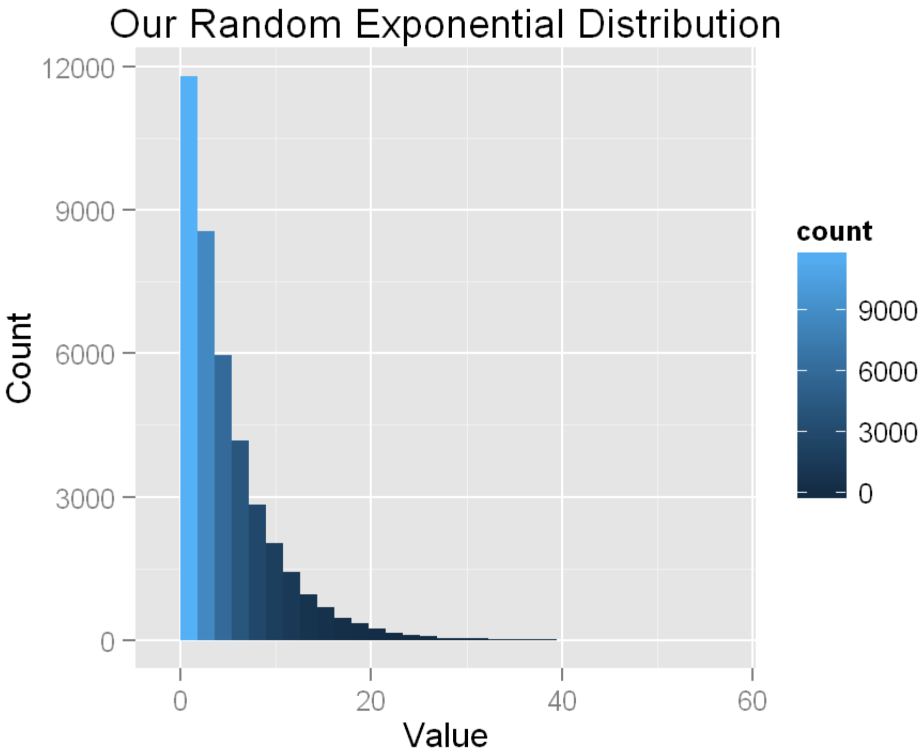
Sample Mean Vs Theoretical Mean

Here we create a 1000 random exponential distributions by firstly creating a single exponential distribution of 410^{4} variables then forming a 1000 by 40 matrix from these values.

```
##       [,1]  [,2]  [,3]  [,4]  [,5]
## [1,] 5.413 4.117 5.135 5.500 5.487
## [2,] 4.416 5.561 4.303 3.215 4.840
## [3,] 6.064 4.780 5.017 4.437 4.031
## [4,] 5.453 3.461 4.876 4.799 6.735
```

```
##   smp_mu mu
## 1  5.013  5
```

The matrix formed in the code block above shows the means of the first 20 random distributions in our sample. It is clear to see that these range from 3.215 to 6.735, centered around our expected mean value of 5. Further more, for quick direct comparison the mean of all our sample means is 5.0126026 which is extremely close to our expected value of 5.



This histogram shows the distribution of values in our example. It is clear to see that it is a negative exponential distribution and although the majority of the values are close to zero and less than our expected mean, the few high outlying values offset these to bring the mean towards the expected value.

Sample Variance Vs Theoretical Variance

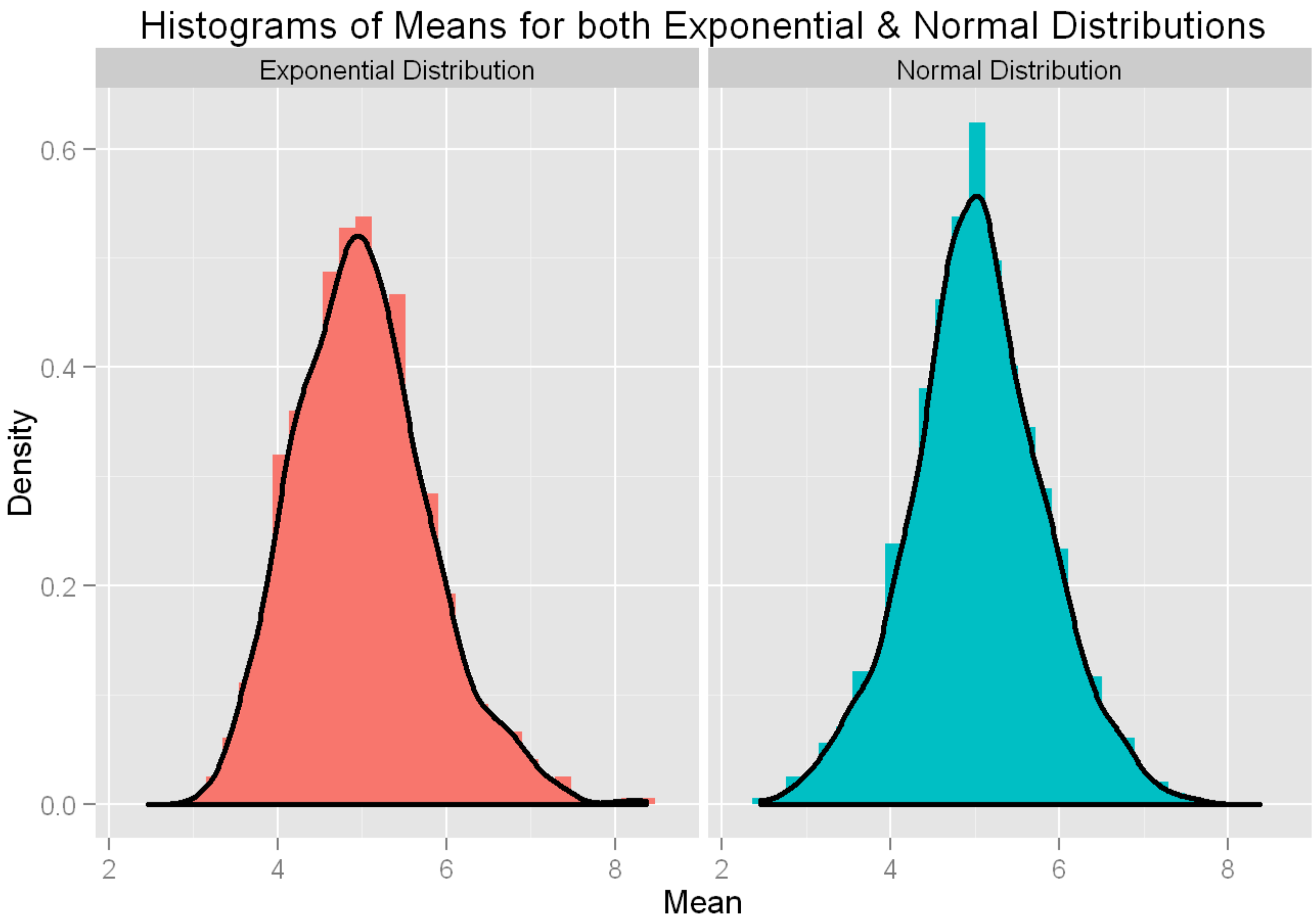
Using the same random exponential distribution as in question 1, we are going to compare how the variance of our distribution is against the theoretical variance calculated from the parameters.

```
##   smp_var var
## 1   24.65  25
```

The output above shows the variance of our distribution (`smp_var`) against the theoretical variance (`var`). Our expected variance for this distribution is 25, our calculated variance of 24.65 is pretty close.

Distribution

The Central Limit Theorem states that although a distribution may not be normal in itself, if it is split into smaller distributions then the means of the sub-distributions will be approximately normally distributed.



The left histogram shows the distribution of means from 1000 exponential distributions of 40 randomly generated variables with the attributes at the start. Although the upper tail is slightly stretched it does look very Gaussian in shape. For comparison the right histogram shows the distribution of means of 1000 normal distributions of 40 randomly generated variables with attributes the same as those used in out exponential distributions. The general shape of both distributions are very similar. Thus we can suggest that the exponential distribution does behave similar to a normal distribution.