Simulating the Central Limit Theorem

Kshitij Kulkarni

Monday, October 12, 2015

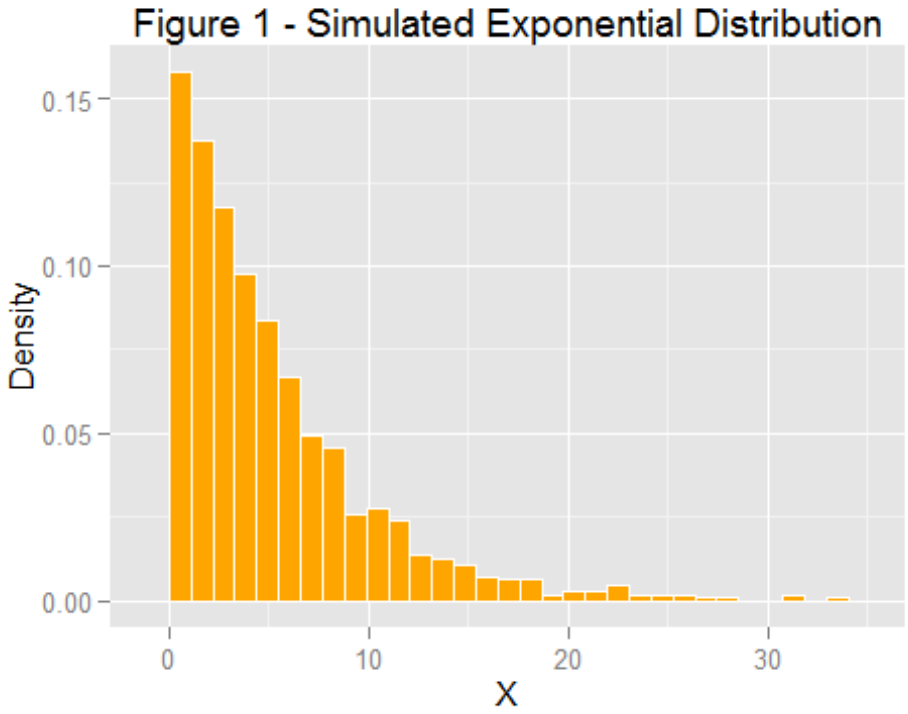
Introduction

This report seeks to investigate the Central Limit Theorem by exploring the relationship between a non-normal (exponential) distribution and a normal distribution. The Central Limit Theorem states that the mean of the any distribution is normally distributed, and this report will investigate this property by simulating 1000 exponential distributions.

The Sample distribution

To take into consideration the worst case scenario, we will take an exponential distribution.An exponential distribution has the shape of a negative exponential function. Often used to model Poisson processes, the distribution is by definition not a normal distribution. The shape of a simulated exponential distribution can be seen in Figure 1 below :

```
library(ggplot2)
set.seed(1)
hist <- data.frame(hist = rexp(1000,0.2))
ggplot(hist, aes(x = hist)) +
  geom_histogram(aes(y = ..density..),colour = "white", fill = "orange") +
  labs(title = "Figure 1 - Simulated Exponential Distribution",
       y = "Density",
       x = "X")
```



While the distribution is not normal, the Central Limit Theorem implies that the mean of exponential distributions are normally distributed. To show this, 1000 exponential distributions are simulated.

Simulations

Investigation of the properties (mean and variance) of the exponential distribution was done using the rexp() function in R. To do this, several parameters were set: - The number of observations to be simulated for each exponential distribution was set to 40 - The rate, Lambda (λ), was set to 0.2 and the number of simulations to be run was set to 1000 times.

```
obs <- 40
lambda <- 0.2
n <- 10000
```

To simulate the exponential distribution and obtain the simulated parameters (mean and standard deviation),the following was performed: - a vector of length of 1000 was initiated - a loop was used to simulate the following 1000 times: -— 40 numbers are (quasi) randomly generated from the exponential distribution —- the mean/standard deviation of each set of 40 exponentially distributed numbers is calculated —- each mean/standard deviation is assigned to a row in the previously initiated vector.

```
## Set seed to enable reproducibility of the report and code
set.seed(1)
expMeans <- vector(length = n)
for(i in 1:n) {
```

```
expMeans[i] <- mean(rexp(obs, lambda))
}
expMeans <- data.frame(expMeans)
## Reset seed to generate the same random numbers to calculate the standard
deviation
set.seed(1)
expSD <- vector(length = n)
for(i in 1:n) {
  expSD[i] <- sd(rexp(obs, lambda))
  }
expSD <- data.frame(expSD)
```

Sample Mean versus Theoretical Mean

Here we calculate the sample mean and compare it to the theoretical mean of the exponential distribution with $\lambda = 0.2$. The theoretical mean of an exponential distribution is $1/\lambda$. With $\lambda = 0.2$, the mean is equal to $1/0.2 = 5$.

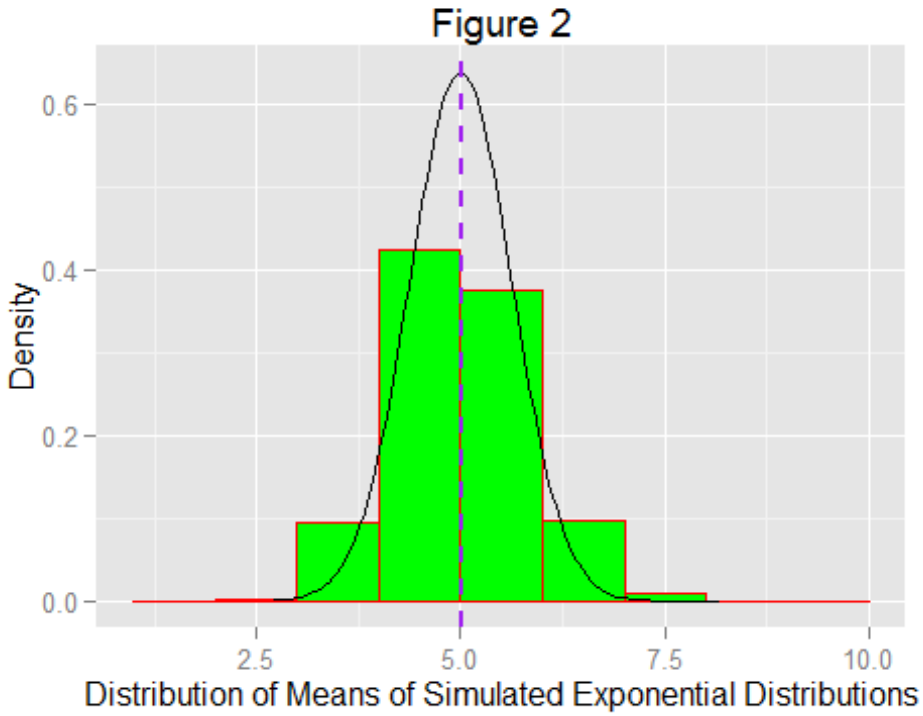
```
theoreticalMean <- 1/0.2
theoreticalMean

## [1] 5
```

The mean of the simulated means is equal to 5.002873, almost equivalent to the theoretical mean of 5.

```
sampleMean <- mean(expMeans$expMeans)
sampleMean

## [1] 5.002873
```

This can be seen in Figure 2 below. A normal curve with mean=5 has been overlayed to allow comparison with the distribution of means. A dashed line also shows the mean of simulated means. Considering a visual perspective, we can observe that the simulated mean of means is the same as the theoretical mean.

```
ggplot(expMeans, aes(x = expMeans)) +
  geom_histogram(aes(y = ..density..), colour = "red", fill = "green",
binwidth = 1) +
  labs(title = "Figure 2",
       y = "Density",
       x = "Distribution of Means of Simulated Exponential Distributions") +
  stat_function(fun = dnorm, args = list(mean = theoreticalMean, sd =
theoreticalMean^2/obs)) +
  geom_vline(aes(xintercept = sampleMean),
             colour = "purple",
             linetype = "dashed",
             size = 1)
```



Sample Variance versus Theoretical Variance

Similar to the previous section, we calculate the mean of sample variances and compare it to the theoretical variance of the exponential distribution with $\lambda = 0.2$. The theoretical standard deviation of the exponential distribution is equal to the mean(=5).

```
theoreticalSD <- 1/0.2
theoreticalSD

## [1] 5
```

The mean of the simulated standard deviations is 4.90, which is also similar to the theoretical standard deviation.

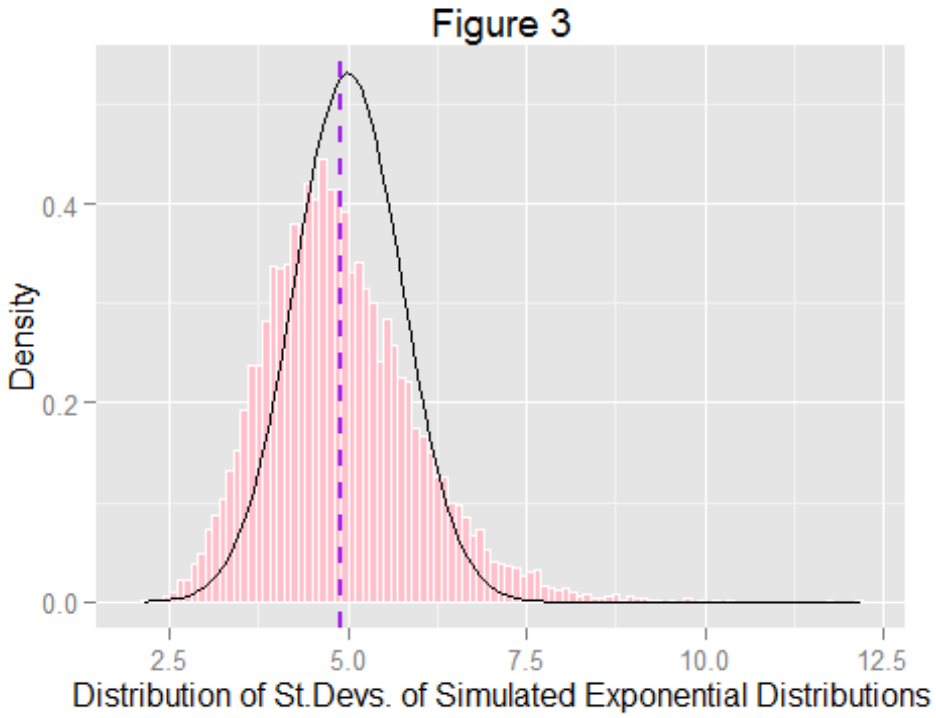
```
sampleSD <- mean(expSD$expSD)
sampleSD

## [1] 4.890633
```

A visual representation of the distribution of the standard deviation in Figure 3 verifies the Central Limit Theorem, the distribution of standard deviations is approximately normal. However, the average of standard deviation is slightly lower than the theoretical value: 4.9 vs 5.

```
ggplot(expSD, aes(x = expSD)) +
  geom_histogram(aes(y = ..density..), colour = "white", fill = "pink",
binwidth = .1) +
  labs(title = "Figure 3",
```

```
       y = "Density",
       x = "Distribution of St.Devs. of Simulated Exponential Distributions")
+
  stat_function(fun = dnorm, args = list(mean = theoreticalSD, sd = 0.75)) +
  geom_vline(aes(xintercept = sampleSD),
             colour = "purple",
             linetype = "dashed",
             size = 1)
```

Figure 3



Conclusion

Hence, we see that irrespective of the distribution of data , the distribution of the means tend to be normal as we simulate more and more , i.e the Central Limit Theorem holds true.