Analysis of the Exponential distribution using simulation

Chris Shaw

23 April 2016

Overview

This paper analyses a simulation of the Exponential distribution in order to investigate how the sample mean and variance compare to the theorietical values. The sample mean distribution is also analysed to see how well it conforms to what is predicted by the Central Limit Theorem (CLT).

We conclude that the distribution of simulated sample means is indeed in accordance with the CLT.

Simulations

The exponential distribution is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. This rate is λ. The mean of this distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$.

In this paper we will explore how a sample of 40 random values from this distribution, denoted by $\bar{X}_n$, behaves over a large number of simulations. The central limit theorem predicts that distribution of sample means generated by the simulation should converge to the following normal distribution:

$$(\bar{X}_n)_i \sim N(\mu, \frac{\sigma^2}{n})$$

as the number of simulations i increases. The μ and σ^2 are the mean and variance of the underlying population, and n is the sample size.

The simulation code to generate 1,000 samples of size 40 is shown in the appendix. The rate λ in the exponential distribution is set to 0.2. A matrix of random values from this distribution is created with 40 columns and 1,000 rows. The name of this matrix is **sim_data**.

Sample Mean versus Theoretical Mean

The mean of each row of **sim_data** is computed and stored in the **sample_means** variable. It is a simple matter to take the mean of these 1,000 values and compare with the theorietical mean $1/\lambda$:

Table 1: Comparison of sample mean to population mean

Statistic	Value
Population mean	5.000
Mean of Sample means	4.972

The sample mean is very close to the theorietical mean after 1,000 simulations.

Sample Variance versus Theoretical Variance

The variance of the exponential distribution is $1/\lambda^2$. In order to calculate the pooled variance across the 1,000 simulations, we use the following formula:

$$\frac{\sum_{i=1}^{1000}(n_i - 1)s_i^2}{\sum_{i=1}^{1000}(n_i - 1)}$$

where s_i^2 is the variance of each sample and n_i is the sample size (in this case 40). The code for this calculation is shown in the appendix.

The following table shows the theorietical variance compared to this pooled variance.

Table 2: Comparison of sample variance to population variance

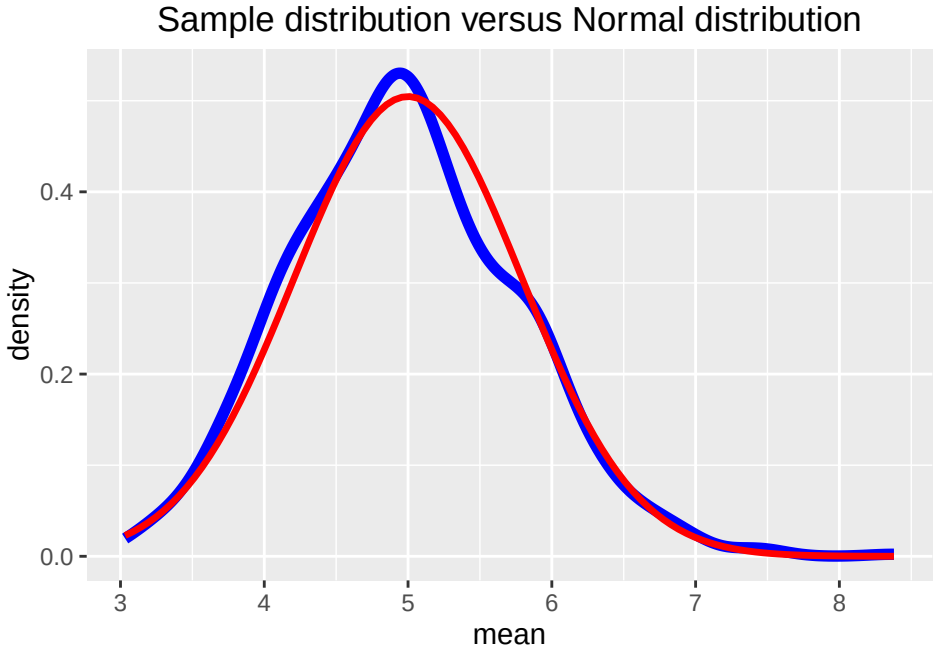
Statistic	Value
Population variance	25.000
Pooled sample variance	24.568

It can be seen that the pooled variation from the simulation is very close again to the theorietical variance after 1,000 simulations.

Distribution

The next task is to compare the distribution density of the sample means to the theorietical distribution under the Central Limit Theorem of:

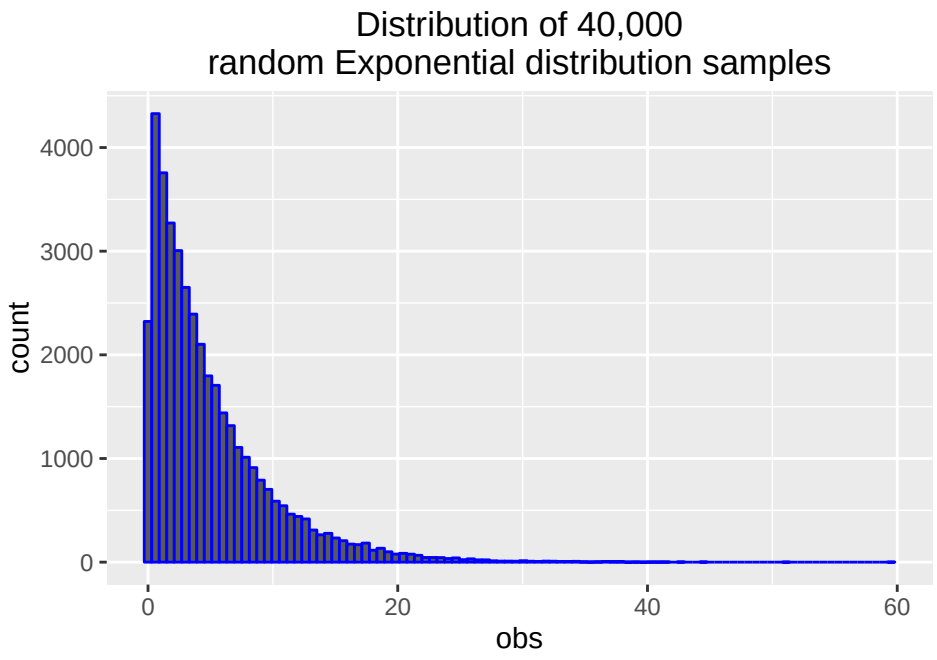
$$X_n \sim N(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$$



The graph shows the distribution of sample means in blue and the Normal distribution above in red for λ=0.2 and n=40.

It can be seen that after 1,000 simulations, the distribution of the sample means itself forms a distribution which is very close to that predicted by the CLT.

Finally, we plot the histogram of all the 40,000 random observations to see how they are distributed. The code for this is in the appendix.



It can clearly be seen that the underlying data does not form a normal distribution. However, as we have seen, thanks to the central limit theorem, we are able to make inferences about the distribution of sample means, even though the underlying data is not normally distributed.

Appendix

This is the code used to generate the matrix of simulation data, and calculate the sample mean and standard deviation:

```
# set up parameters of the exponential distribution
lambda <- 0.2
sample_size <- 40
num_sims <- 1000

mean <- 1/lambda
std <- 1/lambda
var <- std^2

# Ensure simulations are reproducible
set.seed(12345)

# create a matrix of simulation data (rows = simulations)
sim_data <- matrix(rexp(sample_size*num_sims, lambda), nrow = num_sims)

sample_means <- as.data.frame(apply(sim_data, 1, mean))
names(sample_means)[1]="mean"

sample_sds <- as.data.frame(apply(sim_data, 1, sd))
names(sample_sds)[1]="sd"

# Calculate sample mean and variances
sample_mean <- mean(sample_means$mean)
sample_std <- sum(sample_sds^2*(sample_size-1))/
                      (sample_size*num_sims-num_sims)
```

The code below was used to generate the plots:

```
# Plot the density of the simulation sample means against the normal distribution
# with mean of the 1/lambda and standard deviation of 1/lambda divided by
# square root of sample size
ggplot(data=sample_means,aes(x=mean)) +
        geom_line(col="blue", lwd=2, stat="density") +
        stat_function(fun = dnorm,
                      args = list(mean=mean, sd=std/sqrt(sample_size)),
                      color="red", lwd=1.2, lty=1) +
        ggtitle("Sample distribution versus Normal distribution")
```

```
# Plot the histogram of the random observations in sim_data which
# are flattened into a single vector of 40,000 values.
ggplot(data=data.frame(obs=as.vector(sim_data)), aes(x=obs)) +
        geom_histogram(col="blue", bins=100) +
        ggtitle("Distribution of 40,000\nrandom Exponential distribution samples")
```