Exploration of the Central Limit Theorem applied to the exponential distribution

Alun Meredith

26 September 2015

Overview

In this project I will be investigating the Central Limit Theorem by plotting a distribution of averages measured from an exponential distribution.

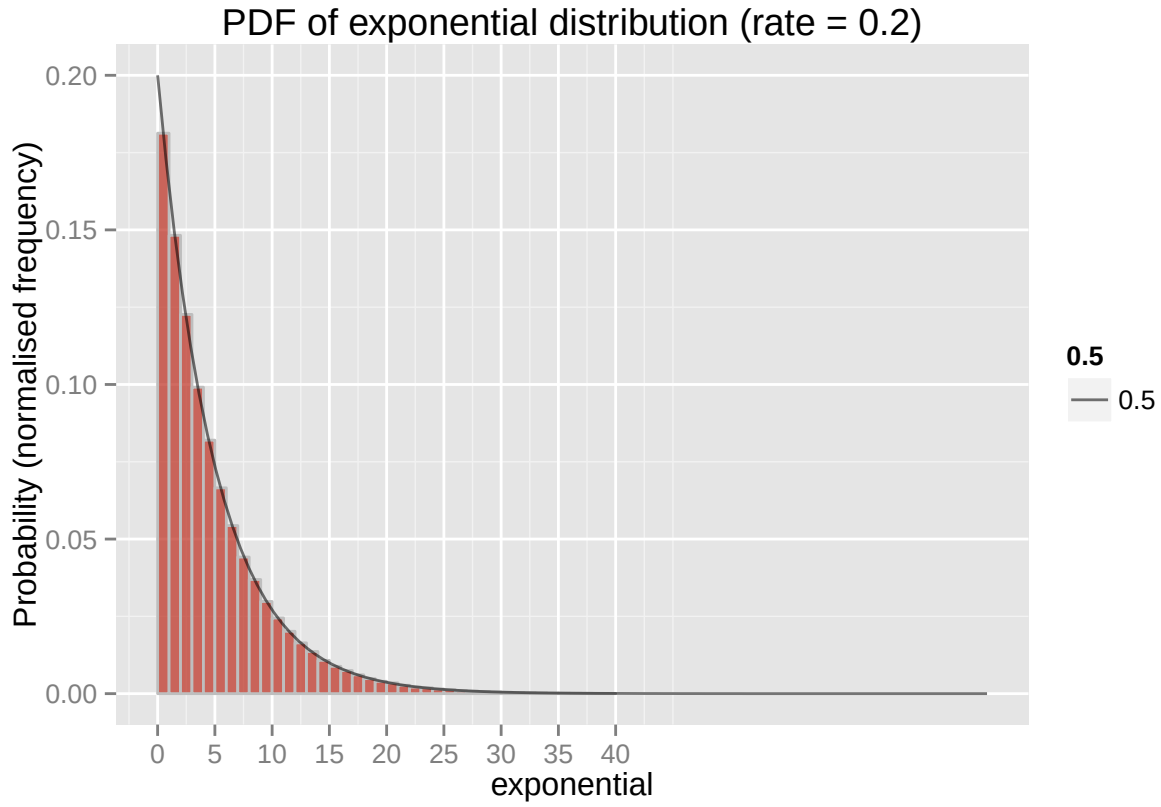
Exponential function

The exponential distribution as described by the equation below is a very bottom heavy distribution with a long tail. Low values are highly likely but the range of the distribution is very large. This is quite different from the gaussian distribution which is symmetric. Due to the strong differences in the distributions it should be easy to distinguish.

$$\lambda e^{-\lambda x}$$

To demonstrate the exponential distribution I have simulated 40000 datapoints and plotted an estimated probability distribution (by normalising the frequency of a histogram) below. The black line shown is the theoretical curve for the exponential probability distribution given a rate parameter 0.2. I will be using rate = 0.2 for the rest of this study.

```
rate = 0.2
exponential = rexp(400000, rate)
ex = as.data.frame(exponential)
```



As you can see in the probability distribution above the simulated data closely follows the theoretical curve.
Central limit theorem

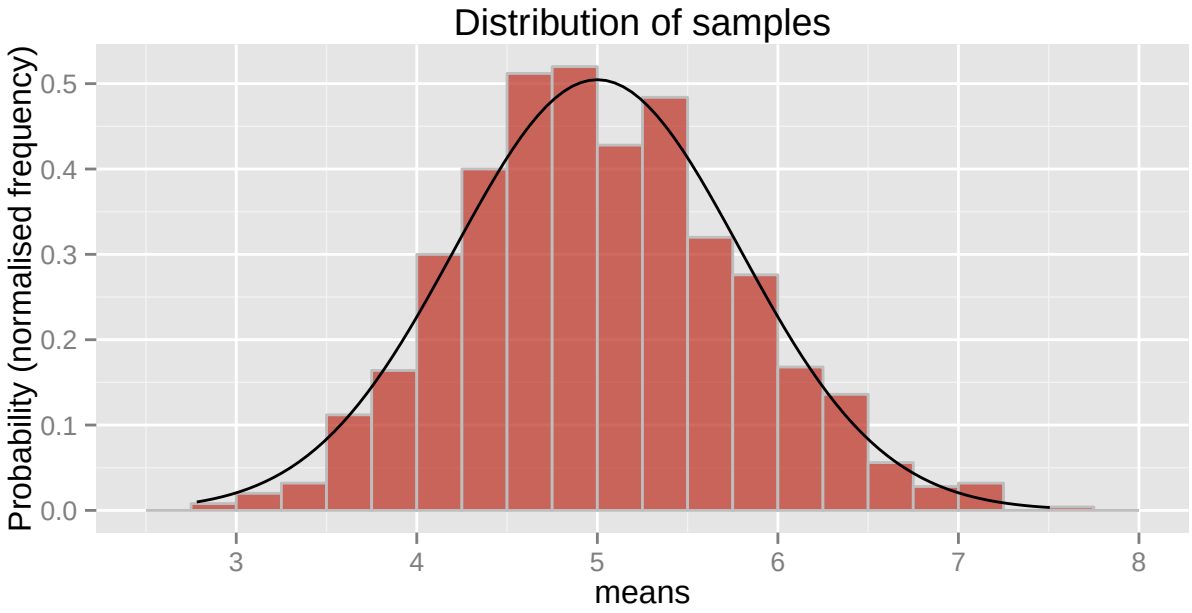
The Central limit theorem describes that the distribution of sample measurements from a population is approximately described by a normal (gaussian) distribution. This distribution is centred at the population mean with standard deviation equal to the standard error of the mean. The set of samples approximates a normal distribution only as the number of samples becomes large and different distribution may trend to the normal distribution at different rates. The central limit theorem gives no guarantee when the sample size is large enough for the approximation to be true.

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

Simulations

To test the central limit theorem (CLT) we will simulate 1000 samples of 40 and compare the distribution of averages.

```
m = 1000
n = 40
set.seed(99)
samples <- sapply(1:m, function(x) rexp(n, rate))
means <- apply(samples, 2, mean)
```



Sample vs Theory

The graph above shows a histogram of the distribution of sample means. The frequency of means has been normalised to give a probability distribution. The black line shows the theoretical normal distribution given by the central limit theorem. This line fits the distribution fairly well but certainly with room for improvement. This suggests that while the central limit theorem holds in this instance we would like to see more samples to fully approximate it to a normal distribution.

Sample mean

The mean of our 1000 samples of 40 measurements is NA. The central limit theorem suggests this should be approximately equal to the mean of the population distribution $\lambda e^{-\lambda x}$ which is given by $\frac{1}{\lambda} = 5$.

This seems to be a fairly good approximation with a percentage error of NA%

Sample variance

The variance of our distribution of sample means is 0.593. The central limit theorem suggests this should approximate to the variance of the population distribution over the number of samples.

$$\frac{\sigma^2}{n} = \frac{1}{\lambda^2 n} = \frac{5^2}{40} = 0.625$$

The percentage error between our sample variance and population variance is 5.13%.

Summary

While our sample mean is a close approximation the sample variance has much bigger error than that of the mean and suggests that a larger number of samples is needed to approximate the central limit theorem fully.