Part 1: Simulation

Keh-Harng Feng

March 6, 2017

Overview

A computer simulation of random variables following an exponetial distribution is carried out. The means of the simulated variable are computed and plotted. The result follows the prediction from the Central Limit Theorem.

Introduction

Exponential distribution is a probability distribution that has a pdf described by the function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

where λ is the only parameter of the distribution. It is often referred to as the *rate*.

The theoretical expected value or mean of the distribution can be found by evaluating the improper integral of the pdf from x = 0 to x = ∞ using integration by parts:

$$E[X] = \frac{1}{\lambda}$$

Similarly, the theoretical variance and standard deviation are found to be:

$$\sigma^2[X] = \frac{1}{\lambda^2} \sigma[X] = \frac{1}{\lambda}$$

This part of the report uses a computer simulation in R to generate a set of random data following the exponential distribution in order to test the **central limit theorem** (CLT).

Central Limit Theorem (CLT)

From Investopedia:

The central limit theorem (CLT) is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all of the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size.

Simulations

For the purpose of this report, all simulations for the exponential distribution are computed with the rate, λ, set to 0.2.

```
lambda <- 0.2
```

Each sample is set to have n = 5, 25 and 125 measurements. Three matrices with sizes n x 1000 are created by filling each column with a random sample that follows the exponential distribution for 1000 samples:

```
n <- 5

# Setting fixed seed first for reproducibility.
set.seed(123)

samples_5 <- matrix(data = rexp(n*1000, rate = lambda), nrow = n, ncol = 1000)

set.seed(321)
samples_25 <- matrix(data = rexp(n^2*1000, rate = lambda), nrow = n^2, ncol = 1000)

set.seed(132)
samples_125 <- matrix(data = rexp(n^3*1000, rate = lambda), nrow = n^3, ncol = 1000)
```

The sample means and sample variances are computed as follows:

```
sample_means_5 <- apply(samples_5, 2, mean)
sample_vars_5 <- apply(samples_5, 2, var)

sample_means_25 <- apply(samples_25, 2, mean)
sample_vars_25 <- apply(samples_25, 2, var)

sample_means_125 <- apply(samples_125, 2, mean)
sample_vars_125 <- apply(samples_125, 2, var)
```

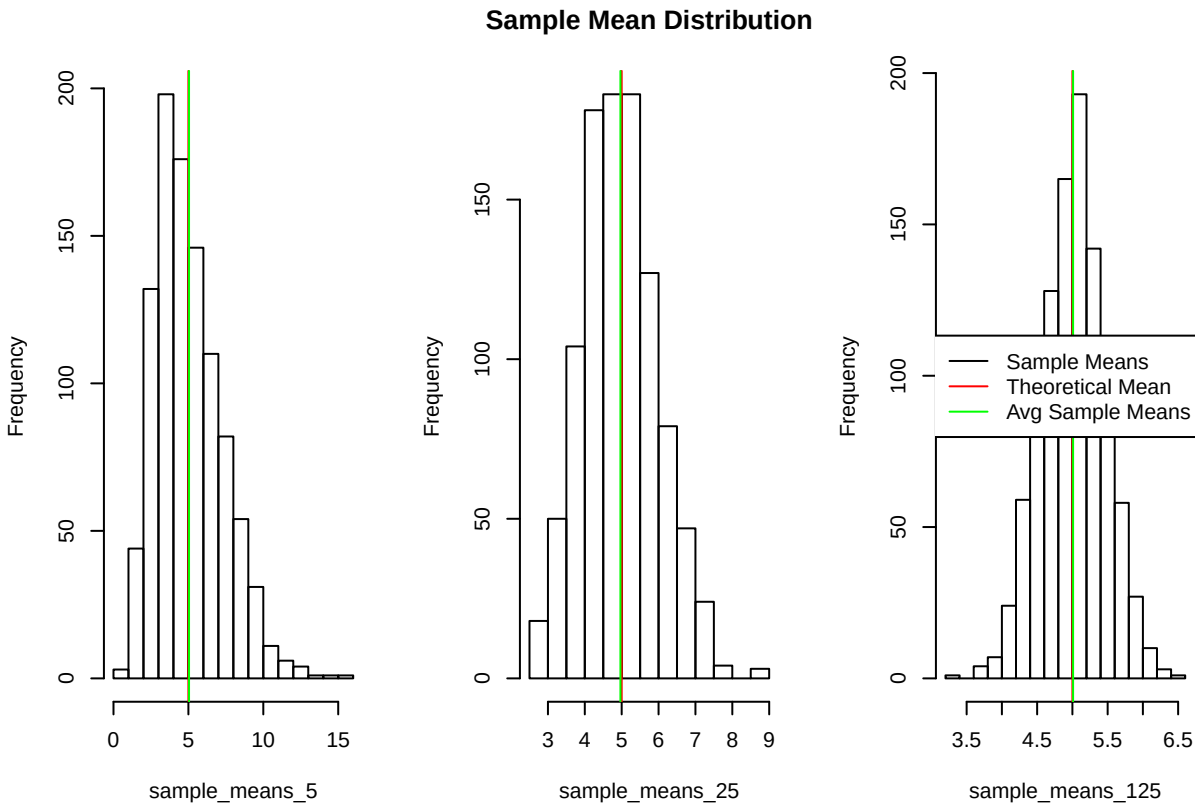
Sample Mean versus Theoretical Mean

Since lambda is known we can compute the theoretical mean and variance.

```
mean_t <- 1/lambda
var_t <- 1/lambda^2
```

The numerical value of the theoretical mean is **5**. This is superimposed on the distribution of sample means below:

```
par(mfrow = c(1,3))
hist(sample_means_5, main = '')
abline(v = mean_t, col = 'red')
mean_5 <- mean(sample_means_5)
abline(v = mean_5, col = 'green')

hist(sample_means_25, main = 'Sample Mean Distribution')
abline(v = mean_t, col = 'red')
mean_25 <- mean(sample_means_25)
abline(v = mean_25, col = 'green')

hist(sample_means_125, main = '')
abline(v = mean_t, col = 'red')
mean_125 <- mean(sample_means_125)
abline(v = mean_125, col = 'green')
legend('center', c('Sample Means', 'Theoretical Mean', 'Avg Sample Means'),
       lty = c(1,1,1), col = c('black', 'red', 'green'), bg = 'white')
```

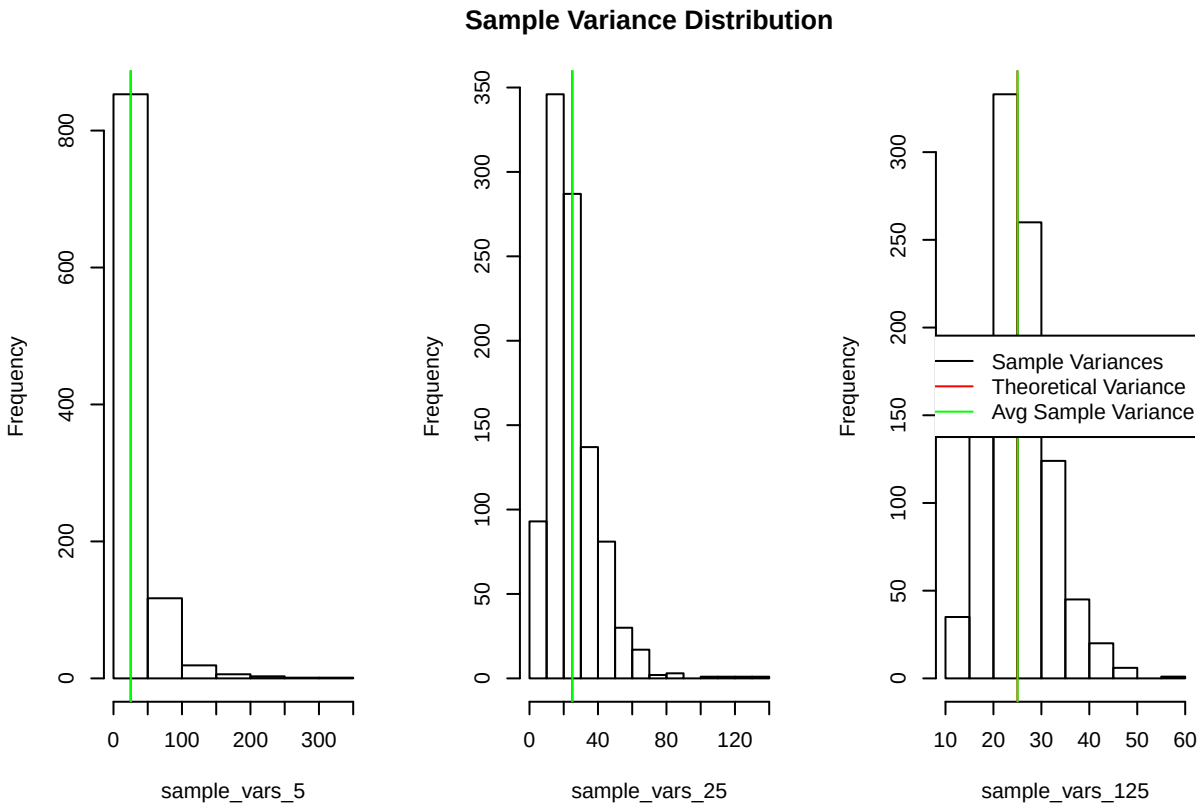


Qualitatively, it is clear from the range of the x-axis that as the sample size goes up, the distribution of sample means starts to group around the theoretical mean more tightly (average sample mean = 5.0097381).

Sample Variance versus Theoretical Variance

The numerical value of the theoretical variance is **25**. Once again this is superimposed on the distribution of sample variances below:

```
par(mfrow = c(1,3))
hist(sample_vars_5, main = '')
abline(v = var_t, col = 'red')
var_5 <- mean(sample_vars_5)
abline(v = var_5, col = 'green')

hist(sample_vars_25, main = 'Sample Variance Distribution')
abline(v = var_t, col = 'red')
var_25 <- mean(sample_vars_25)
abline(v = var_25, col = 'green')

hist(sample_vars_125, main = '')
abline(v = var_t, col = 'red')
var_125 <- mean(sample_vars_125)
abline(v = var_125, col = 'green')
legend('center', c('Sample Variances', 'Theoretical Variance', 'Avg Sample Variances'),
       lty = c(1,1,1), col = c('black', 'red', 'green'), bg = 'white')
```



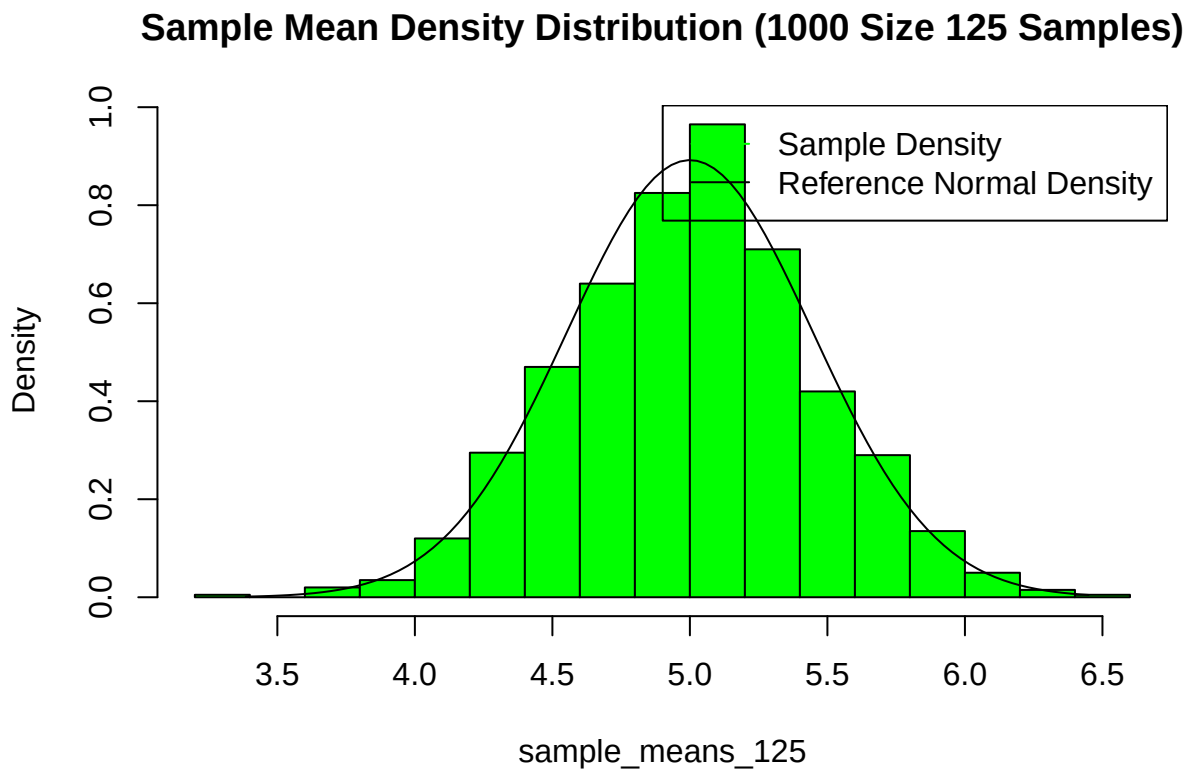
Similar to the sample means, as the sample size increases the sample variances become more clustered around the theoretical value (average sample variance = 25.0474321). The convergence of average sample mean and average sample variance to the theoretical values is expected, since they are both unbiased statistics.

Distribution Identification

This section redoubles effort to provide more qualitative evidence that supports the CLT while also verifying the second part of the theorem. That is, the distribution of the sample mean is a normal distribution with $Var[x] = \frac{\sigma^2(x)}{n}$ where $\sigma^2(x)$ is the true variance of the population and n is the sample size.

A histogram of the probability density of the sample means from size 125 samples is shown below. A plot of the normal distribution with mean = 5 and variance = 5 is superimposed on top.

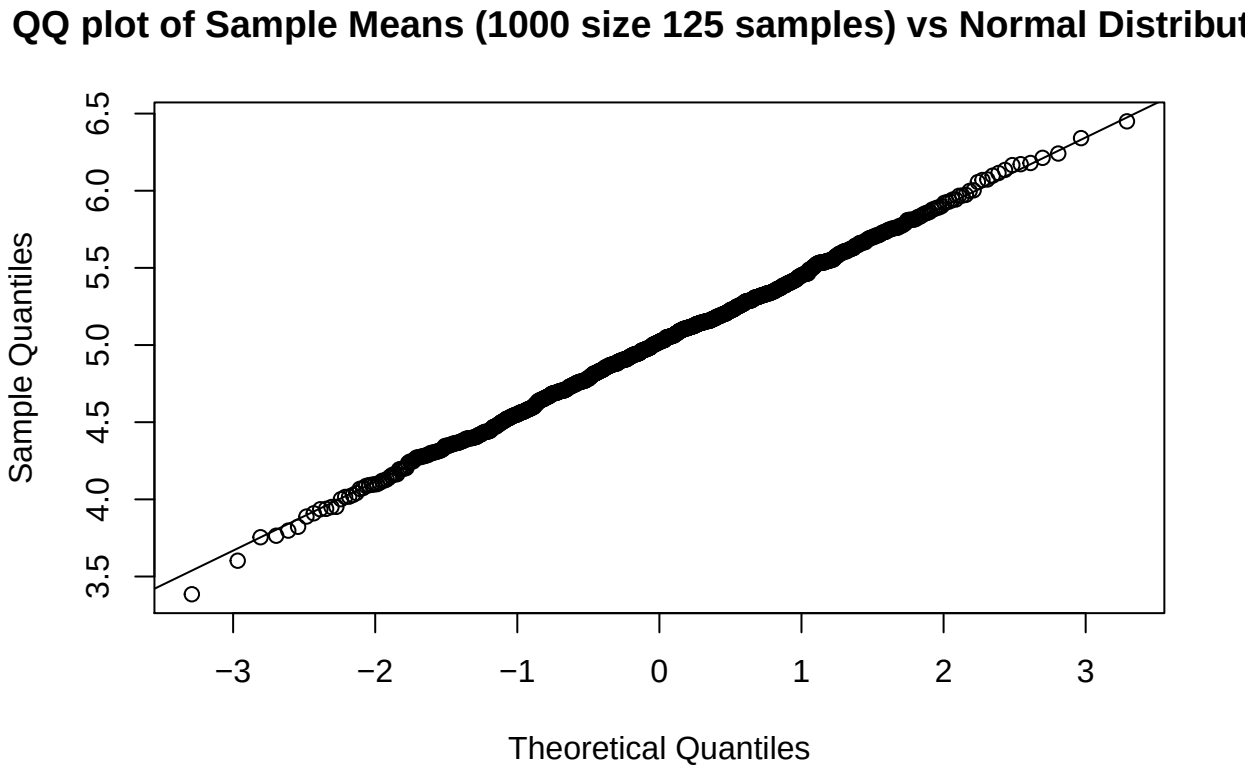
```
# creating reference normal distribution
x = seq(from = min(sample_means_125), to = max(sample_means_125), length.out = 100)
ref_means = dnorm(x, mean = mean_t, sd = sqrt(var_t/125))

par(mfrow = c(1, 1))

hist(sample_means_125, breaks = 20, prob = TRUE,
     main = 'Sample Mean Density Distribution (1000 Size 125 Samples)', col = 'green')
lines(x, ref_means)
legend('topright', c('Sample Density', 'Reference Normal Density'), lty = c(1, 1),
       col = c('green', 'black'))
```



The good agreement in shapes indicates that the variance in the sample means indeed follows $Var[X] = \frac{\sigma^2(x)}{n}$ where n = 125 is the number of measurements per sample as predicted by CLT.

Another tool to check the normality of the distribution is the quantile-quantile (QQ) plot, where the quantile a data point belongs to in the empirical distribution is checked against its quantile in a normal distribution. If the empirical distribution matches up well with the normal distribution, most of its points should lie on the diagonal. A QQ plot showing how the sample mean distribution matches up with a normal distribution is shown below.

```
qqnorm(sample_means_125,
       main = 'QQ plot of Sample Means (1000 size 125 samples) vs Normal Distribution')
qqline(sample_means_125)
```



Since most points are either directly on the diagonal or fairly close to it, the distribution of sample means with 1000 samples is most likely a normal distribution, as predicted by CLT.