Illustration of the Central Limit Theorem using Simulation in R

Krzysztof Jarszak

Sunday, March 22, 2015

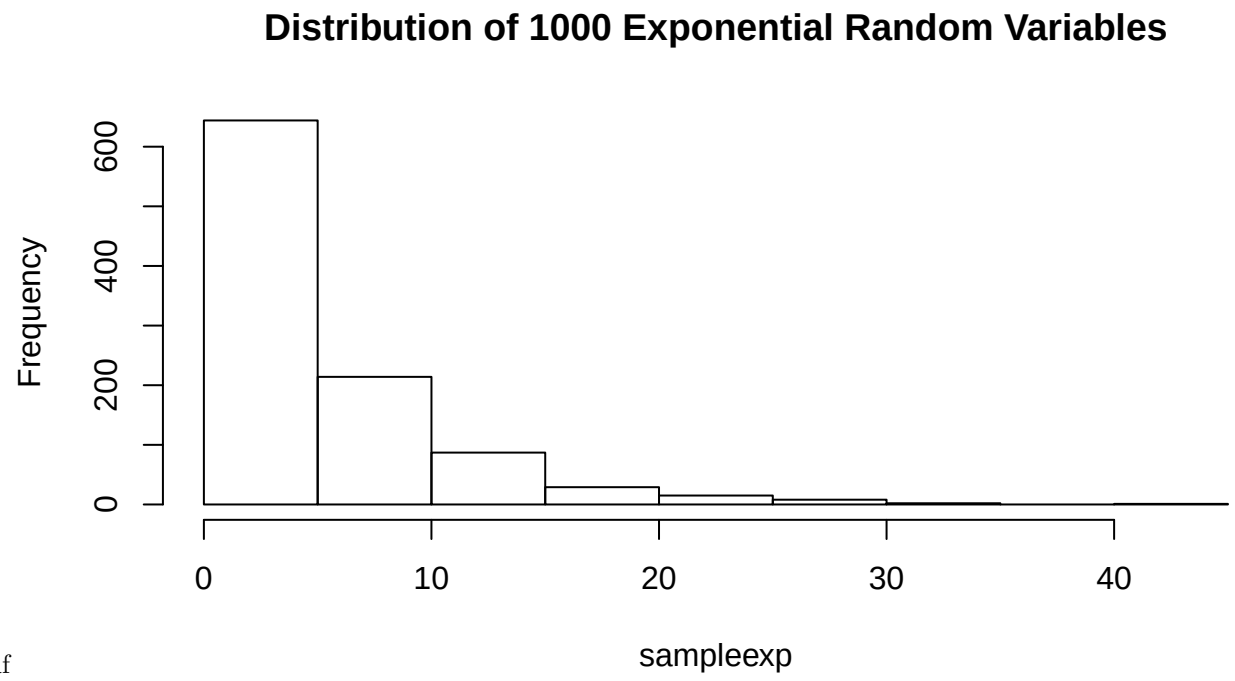
Overview

This project report will illustrate the concept of the popular Central Limit Theorem (CLT) using simulated data in R . Specifically, we will investigate the Exponential distribution and use it to illustrate the Central Limit Theorem. We will compare the sample mean to theoretical mean, the sample variance to the theoretical variance. We will also compare the distribution of a large collection of exponential random variables to that of a large collection of 40 random exponential variables.

Comparison of Sample Mean to the Theorectical Mean of the Distribution

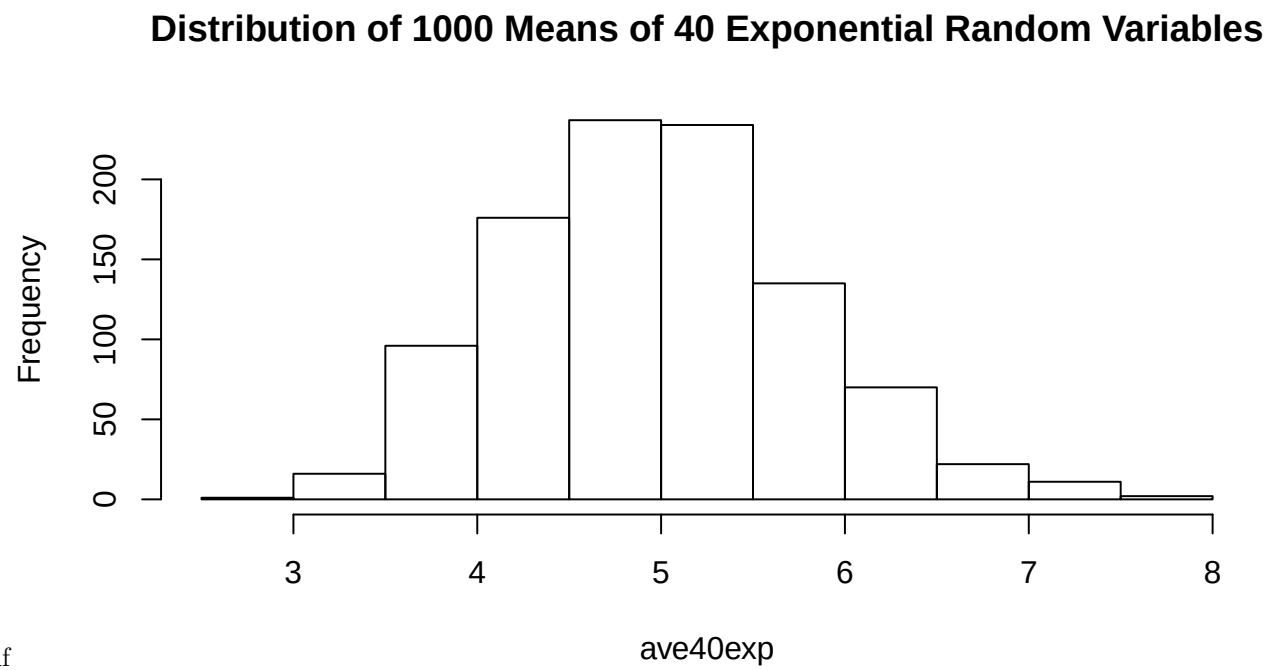
To achieve this aim, we will simulate 1000 exponential variables and compare the distribution of these variates to the distribution of 1000 collection of averages of 40 Exponential Distribution. Note that the lambda parameter of the exponential distribution will be set to 0.2 for the simulation.

```
sampleexp<-rexp(1000, 0.2)  # simulate 1000 exponential RVs
ave40exp<- NULL
for (exp in 1:1000){
     ave40exp<-c(ave40exp, mean(rexp(40, 0.2)))
     # this block will simulate 1000 collection of averages of 40 exponential RVs
}
hist(sampleexp, main = 'Distribution of 1000 Exponential Random Variables')
```



1-1.pdf

```
hist(ave40exp, main = 'Distribution of 1000 Means of 40 Exponential Random Variables')
```



1-2.pdf

In the two figure above, notice how the distribution of 1000 generated exponential random variables is skewed to the right while the distribution of the sample means follows normal distribution. This is as expected because CLT states that the arithmetic mean of a sufficiently large number of iterates of independent random variables will be approximately normally distributed, regardless of the underlying distribution (in this case exponential distribution).

```
meansampleexp <- mean(sampleexp)
meansampleexp
```

```
## [1] 5.111596
```

```
meanave40exp <- mean (ave40exp)
meanave40exp
```

```
## [1] 4.96102
```

From the above output, note how the mean of the 1000 exponential random variables is close to the mean of averages (mean of means)of 40 random variables. This is expected as Central Limit Theorem implies that the mean of the sample averages converge in probability to the expected mean.

Comparison of Sample Variance to the Theorectical Variance of the Distribution

To achieve this, we compare the variance of the simulated 1000 exponential random variables to the variance of collection of averages of 40 exponential random variables.

```
varsampleexp <- var(sampleexp)
varsampleexp
```

```
## [1] 28.69676
```

```
varave40exp <- var(ave40exp)
varave40exp
```

```
## [1] 0.6417291
```

The output above shows that the variance of the 1000 random variables and the variance of the collection of averages of 40 random variables. Note that the distribution of the 1000 random variates is more variable than that of the sampling distribution of 40 random variables. Also note that the variance of the sampling distribution of the 40 random variables is approximately the variance of the 1000 random variables divided by the sample size (40). This observation is in agreement with the CLT.

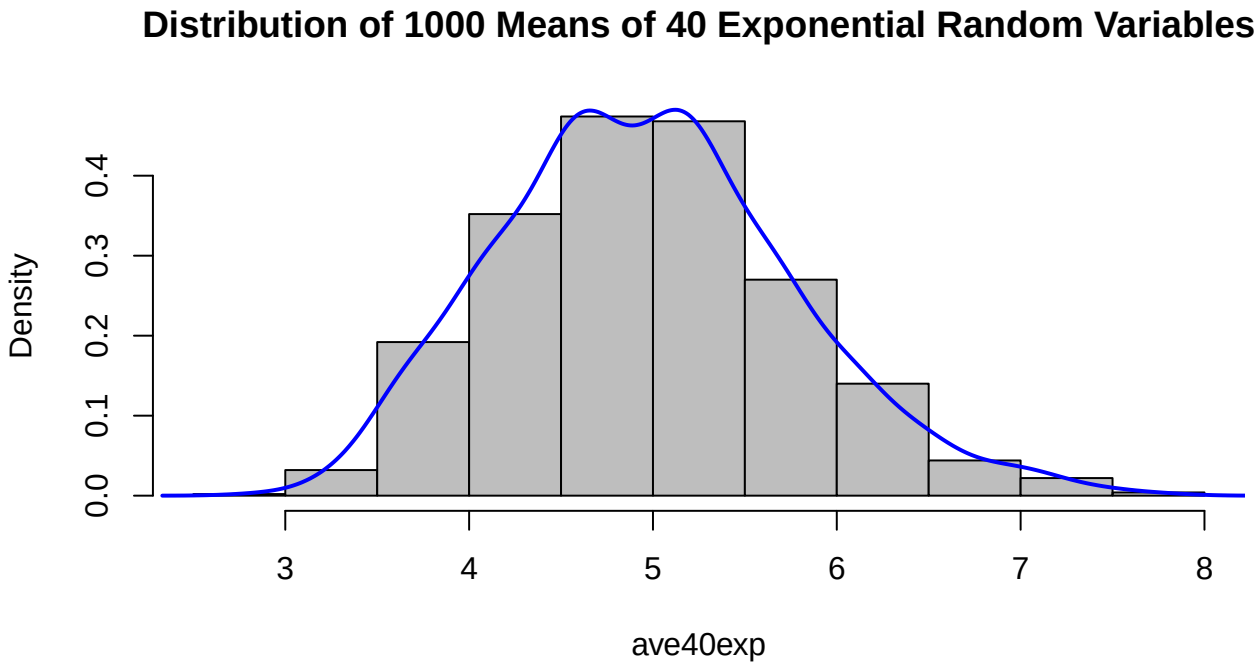
Distribution of the 1000 Collections of Averages of 40 Exponential Random Variables (Sampling Distribution).

In this section, our aim is to show that the sampling distribution always follow Normal distribution regardless of the underlying distribution. To show this, we plot the histogram of the sampling distribution and its density compared with the density of Normal distribution. We can also carry out formal tests for normality. Tests we can use include Kolmogorov - Smirnov test. 'fitdistrplus' package was used to carry out the tests.

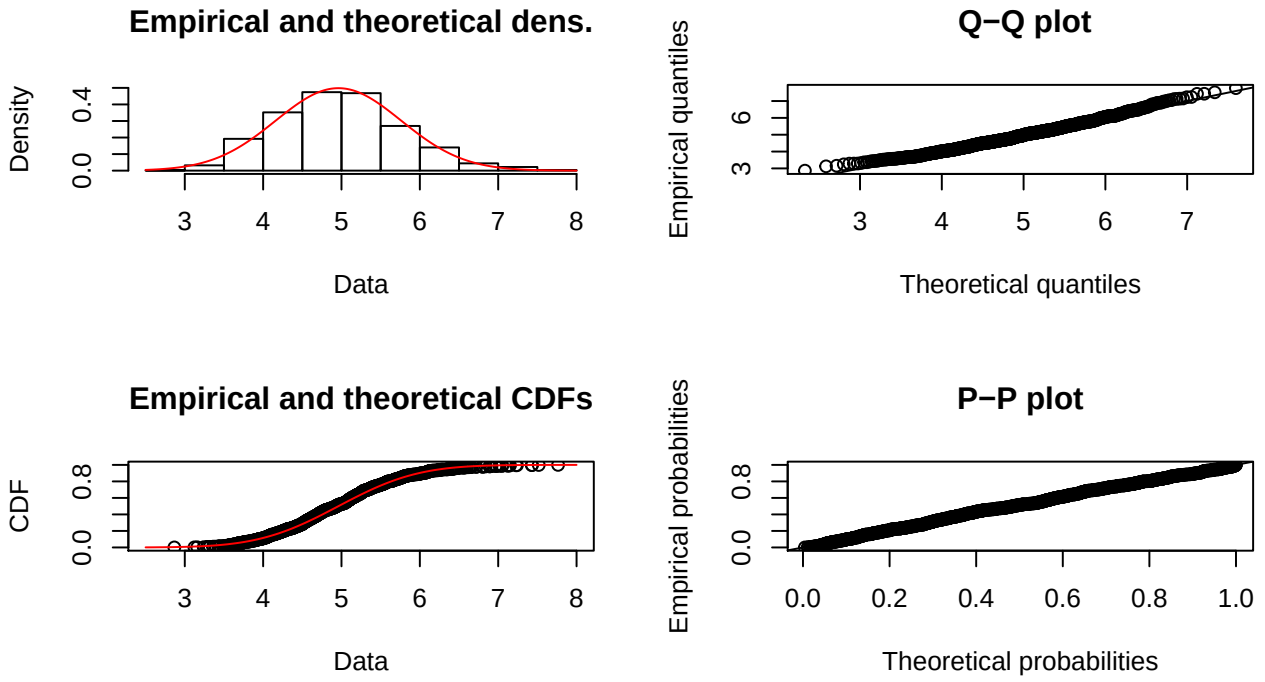
```
require (fitdistrplus)
```

```
## Loading required package: fitdistrplus
```

```
hist(ave40exp, main = 'Distribution of 1000 Means of 40 Exponential Random Variables', prob = TRUE, col=
lines(density(ave40exp), col = 'blue', lwd=2)
```



```
fit<-fitdist(ave40exp, "norm")
fitgofstat<-gofstat(fit)
plot(fit)
```




```
print(fitgofstat$kstest)
```

```
##     1-mle-norm
## "not rejected"
```

The first figure above shows the histogram of the sampling distribution and its density. The histogram and density line looks very much like the bell curve of the Normal Distribution. The second figure shows the fit of the sampling distribution when a Normal distributon is fiited to the data. The Quantile plot and PP plot show a good fit of Normal distribution to the data. Also a formal goodness of fit test was carried out using Kolmogorov-Smirnov test and the Null hypothesis of a good fit was not rejected.

Summary

The purpose of this project report is to try and illustrate the famous Central Limit Theorem which states that the sampling distribution of a large number of independent, identically distributed variables will be approximately normal (with mean equal to expected mean of the underlying distribution and variance equal to expected variance of the underlying distribution divided by sample size), regardless of the underlying distribution.