Project: Simulation of data and demonstration of the Central Limit Theorem

Vineet W. Singh

19 February 2018

Introduction

The Central Limit Theorem is one of the foundations on which statistics is based. This brief project demonstrates simulating a set of data and the resultant effects of the Central Limit Theorem at work.

Analysis

Given a data set with observations/sampling units, collected from a population that is exponentially distibuted and which has a certain rate parameter(lambda), then the theoretical population mean is supposed to be equal to 1/lambda and the theoretical standard deviation(SD) of the mean is also supposed to be equal to 1/lambda.
To demonstrate this, we first randomly generate 40 observations from an exponential distribution with a rate parameter of 0.2.

```
mexp<-rexp(40,0.2)
```

The mean of this particular sample of 40 observations is:

```
## [1] 4.45358
```

```
## [1] "Difference between sample mean and theoretical mean is:"
```

```
## [1] -0.5464201
```

The theoretical mean & SD of the exponential distribution from which the data has been randomly sampled is 1/lambda = 1/0.2 = 5.

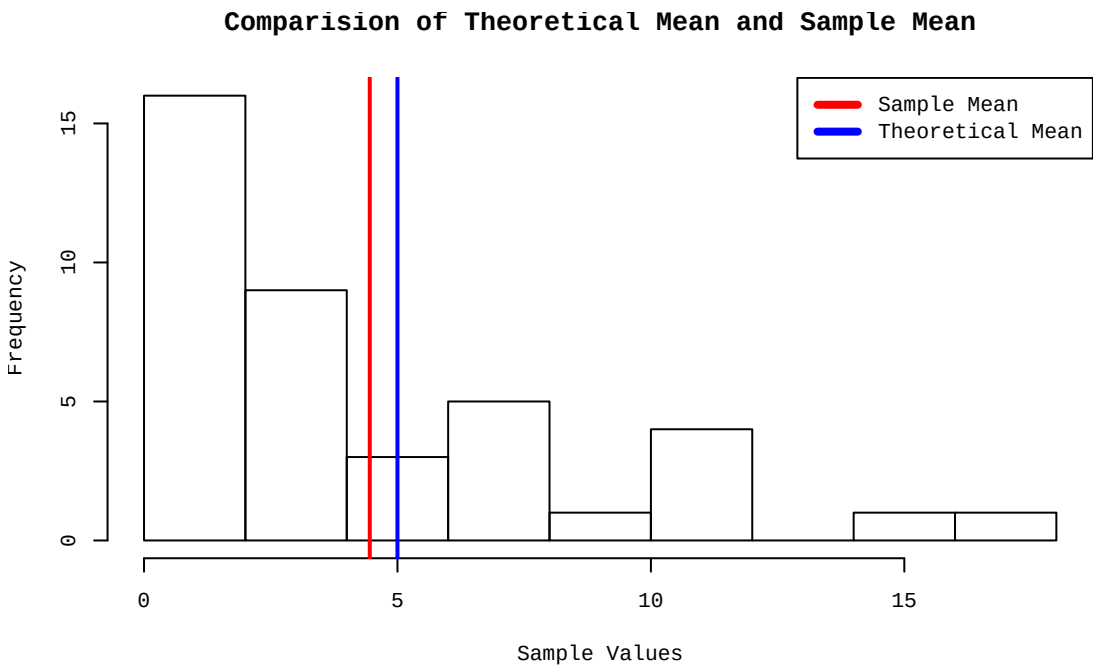
The standard deviation is a better measure of variance as it is in the same units as the mean. The SD of this particular sample of 40 observations is:

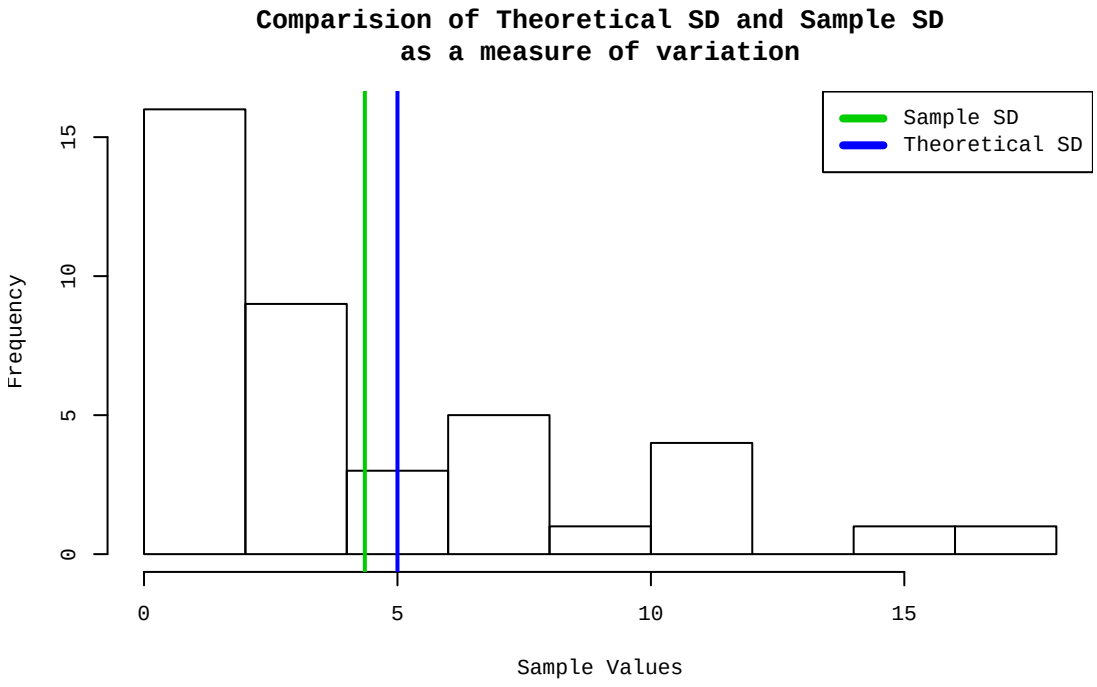
```
## [1] 4.356729
```

```
## [1] "Difference between theoretical SD and sample SD is:"
```

```
## [1] -0.6432707
```

It would be difficult to plot the variance (due to difference of units) on the histogram so it is probably a better idea to plot the SD.





Histogram of 40 randomly genererated points and Mean and SD

The mean & SD of the randomly generated set are close to that of the population mean & Standard Deviation(SD) but given the limited sample size there is bound to error in the estimates. The solution to minimising this error is to use the law of large numbers to collect a large number of samples with the same number of observations. An alternative to collecting a large number of samples is to randomly generate them using a computer.

Simulation and demonstration of the Central Limit Theorem

The central limit theorem states that the means of a large number of samples will be normally distributed around the theoretical population mean.
To begin this demonstration we need to generate 1000 samples of 40 observations, randomly taken from the exponential distribution with a rate parameter(lambda) of 0.2.

The mean and variance of the distribution of means of samples is given by

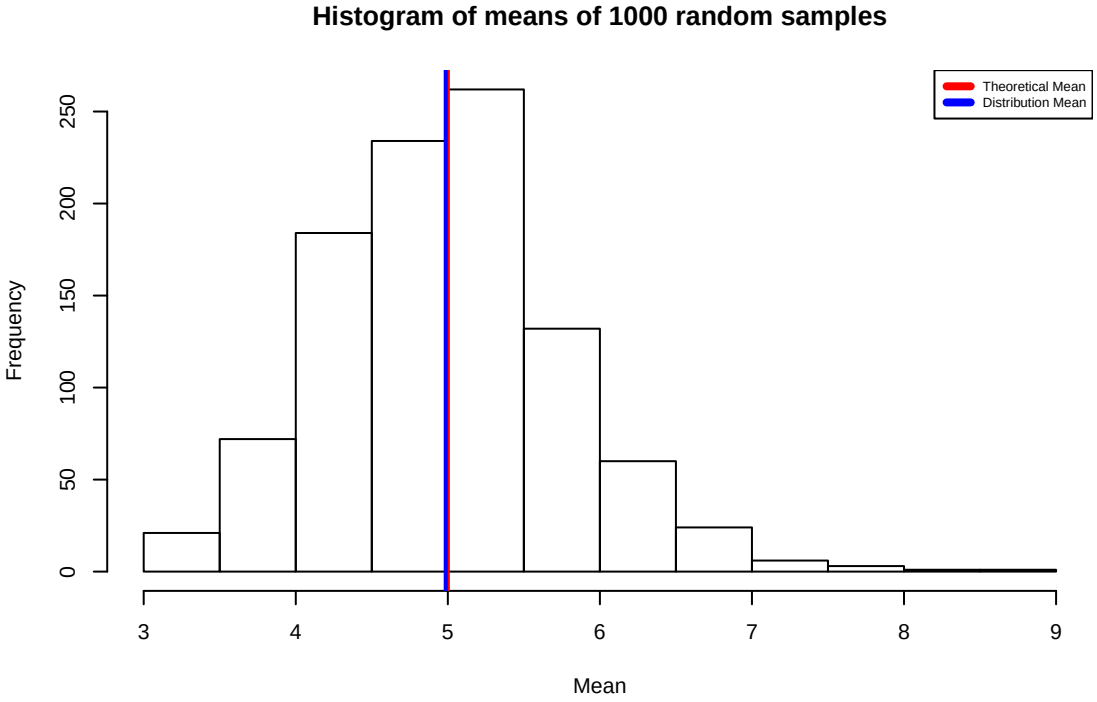
```
## [1] "Distribution Mean is: 4.98595343542716"
```

```
## [1] "Distribution Variance is: 0.60530632394495"
```

```
## [1] "Theoretical Variances: 0.625"
```

```
## [1] "Difference in Variances: 0.0196936760550496"
```

After ploting a simple histogram to observe how the 1000 sample means are arranged about the theoretical mean (red line in the plot), we find that the generated sample means cluster around the theoretical mean and the distribution of means is fairly symmetrical signifying a normal distribution.



Variances in the means are well under the theoretical variance i.e. ((1/lambda)^2*(1/sample size)).

The normality of the data (means) can also be confirmed by the use of the qqnorm function to generate a log normal plot of the data. If the data points on the plot lie almost on or about a straight line, then the data follows a normal distribution. This can be observed on a log-normal plot confirming the normal distribution of the sample means. (PLOT 1 in the Appendix)

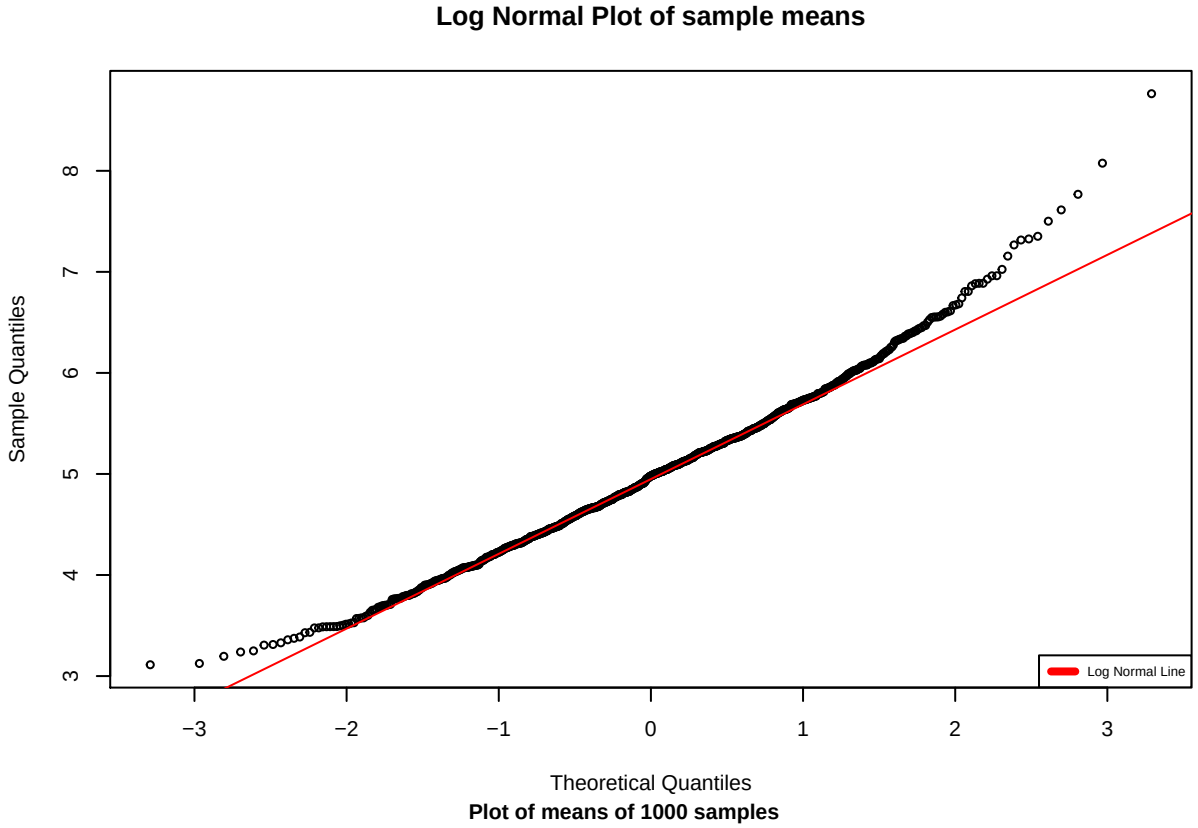
Conclusion

We have simulated a set of 1000 samples of 40 observations (from the exponential distribution) and we have shown that the the Central Limit Theorem is satisfied for the mean of the simulated samples. The original population is exponentially distributed but the means of samples generated from this population
follow a normal distribution which is centered at the population mean. The variance of the generated means are also under the theoretical variation limit.

Appendix

PLOT 1 & CODE for the log normal plot of sample means

```
par(cex=0.7)
qqnorm(mns, main="Log Normal Plot of sample means",cex=0.7)
qqline(mns,col=2)
legend("bottomright", "Log Normal Line",col=2,lwd=4,cex=0.6)
mtext(text=paste0("Plot of means of 1000 samples\n",
                        "Each sample = 40 observations\n",
                      "from an Exponential Distribution"),
      side=1,line=1,outer=TRUE,cex=0.7,font=2)
```



SIMULATION CODE

```
mns=NULL
vars=NULL
for (i in 1 : 1000){
  x<-rexp(40,0.2)
  mns=c(mns,mean(x))
  vars=c(vars,var(x))
}
```

CODE for the histograms of random observations

```
par(mfrow=c(2,1),oma=c(3,0,0,0),family="mono",font=1,cex=0.7)
hist(mexp, main="comparision of Theoretical Mean and Sample Mean",
     xlab="Sample Values")
abline(v=mean(mexp),col=2,lwd=4)
abline(v=5,col=4,lwd=4)
```

```
legend("topright", c("Sample Mean","Theoretical Mean"), col=c(2,4), lwd=4)
hist(mexp, main=paste0("Comparision of Theoretical SD and Sample SD\n",
                       "as a measure of variation"),
     xlab="Sample Values")
abline(v=sd(mexp),col=3,lwd=4)
abline(v=5,col=4,lwd=4)
legend("topright", c("Sample SD",
                     "Theoretical SD"), col=c(3,4), lwd=4)
mtext(text=paste0("Histogram of 1000 randomly genererated points and ",
                  " Mean and SD"),side=1,line=1,outer=TRUE,cex=0.7,font=2)
```

CODE for generating mean/variance of distribtuion

```
meanmns<-mean(mns)
varmns<-var(mns)
tvarmns<-(25/40)
print(paste("Distribution Mean is:",meanmns))
print(paste("Distribution Variance is:",varmns))
print(paste("Theoretical variation should be:",tvarmns))
print(paste("Difference is:",abs(tvarmns-varmns)))
```