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

Vineet W. Singh

19 February 2018

Introduction

The law of large numbers and the central limit theorem are part of the foundations on which statistics is based. This brief project demonstrates simulating a set of data and the resultant effects of the law of large numbers and the central limit theorem at work.

Analysis

The law of large numbers states that given enough observations, any statistics generated from those observations will estimate the pouplation statistics with some accuracy.

For example, given a data set with a large number of observations/sampling units, that has been 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)/Variance of the mean is also supposed to be equal to 1/lambda.

In this chunk of R code, we randomly generate 1000 observations from an exponential distribution with a rate parameter of 0.2.

```
mexp<-rexp(1000,0.2)
```

The mean of this particular sample of 1000 observations is:

```
print(mean(mexp))
```

```
## [1] 5.193545
```

The theoretical mean of the original population that follows the exponential distribution is 1/lambda= 1/0.2 = 5.

The SD/Variance of this particular sample of 1000 observations is:

```
print(sd(mexp))
```

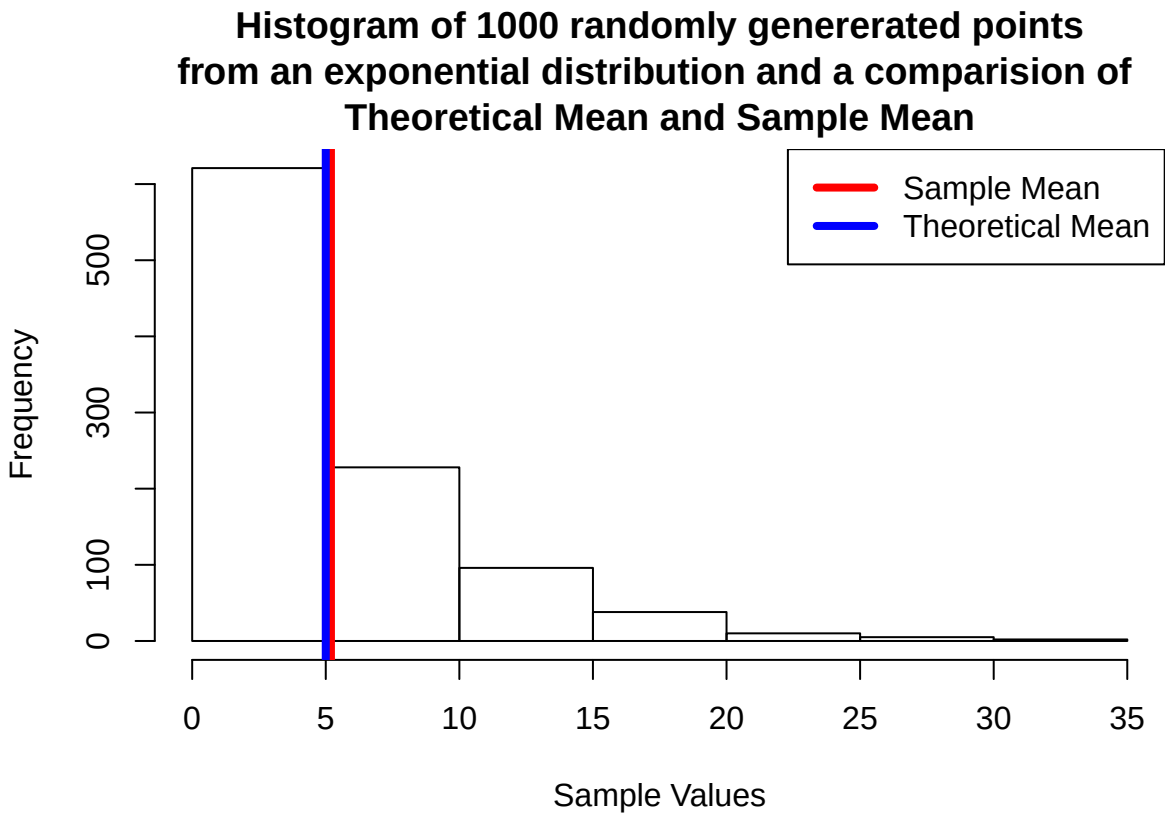
```
## [1] 5.104187
```

The theoretical SD of the original population that follows the exponential distribution is 1/lambda= 1/0.2 = 5.

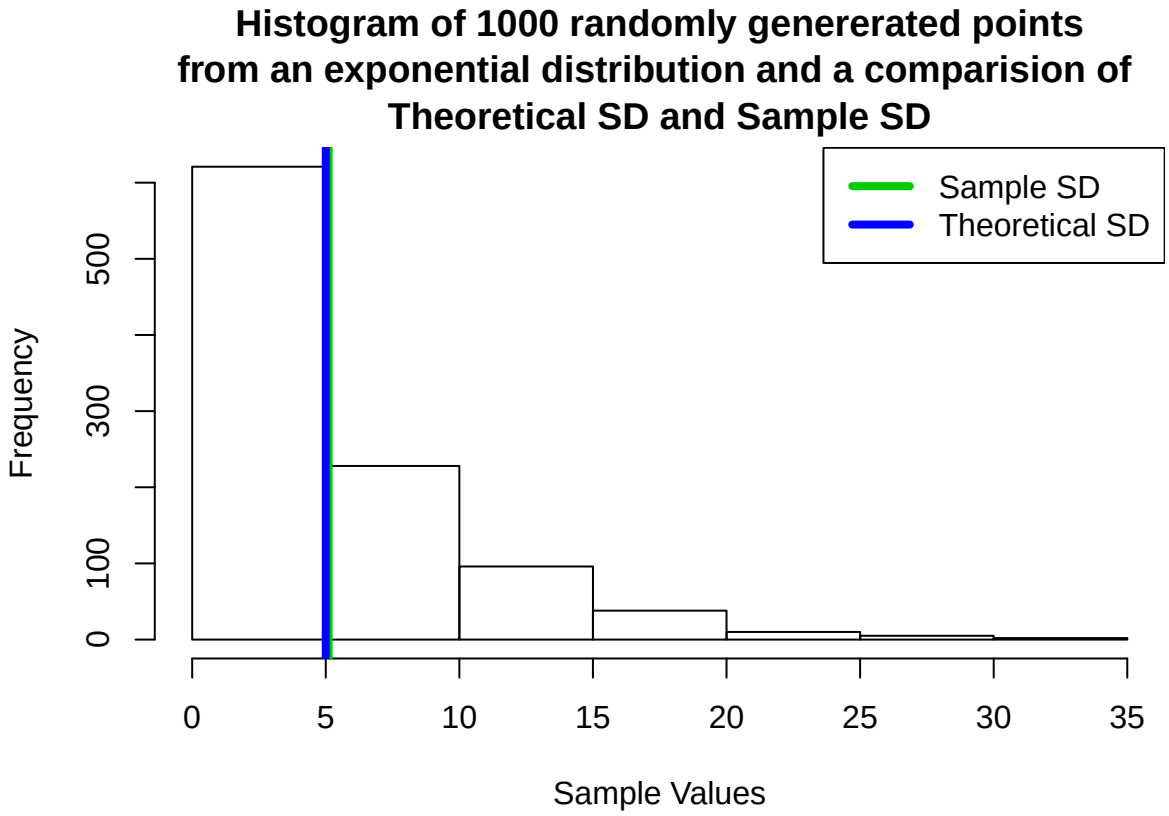
The results above show that the sample mean and SD are quite close to the theoretical mean and SD. This is also demonstrated by the two histograms of randomly generated values (from the exponential distribution) and a comparision between the theoretical mean/SD, sample mean/SD respectively, displayed below:

```
hist(mexp, main=paste0("Histogram of 1000 randomly genererated points\n",
                       "from an exponential distribution and a comparision of \n",
                         "Theoretical Mean and Sample Mean"),
     xlab="Sample Values")
abline(v=mean(mexp),col=2,lwd=4)
```

```r
abline(v=5,col=4,lwd=4)
legend("topright", c("Sample Mean","Theoretical Mean"), col=c(2,4), lwd=4)
```




```r
hist(mexp, main=paste0("Histogram of 1000 randomly genererated points\n",
                       "from an exponential distribution and a comparision of \n",
                          "Theoretical SD and Sample SD"),
     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)
```



The mean & SD of the randomly generated set are quite close to that of the population mean & Standard Deviation(SD)/Variance and can be taken to be fair/close estimates of the population mean and SD thereby confirming the effect of the law of large numbers.

The central limit theorem states that the means of a large number of samples will be normally distributed around the theoretical population mean.

To demonstrate this we first generate 1000 samples of 1000 observations, randomly taken from the exponential distribution with a population rate parameter(lambda) of 0.2.

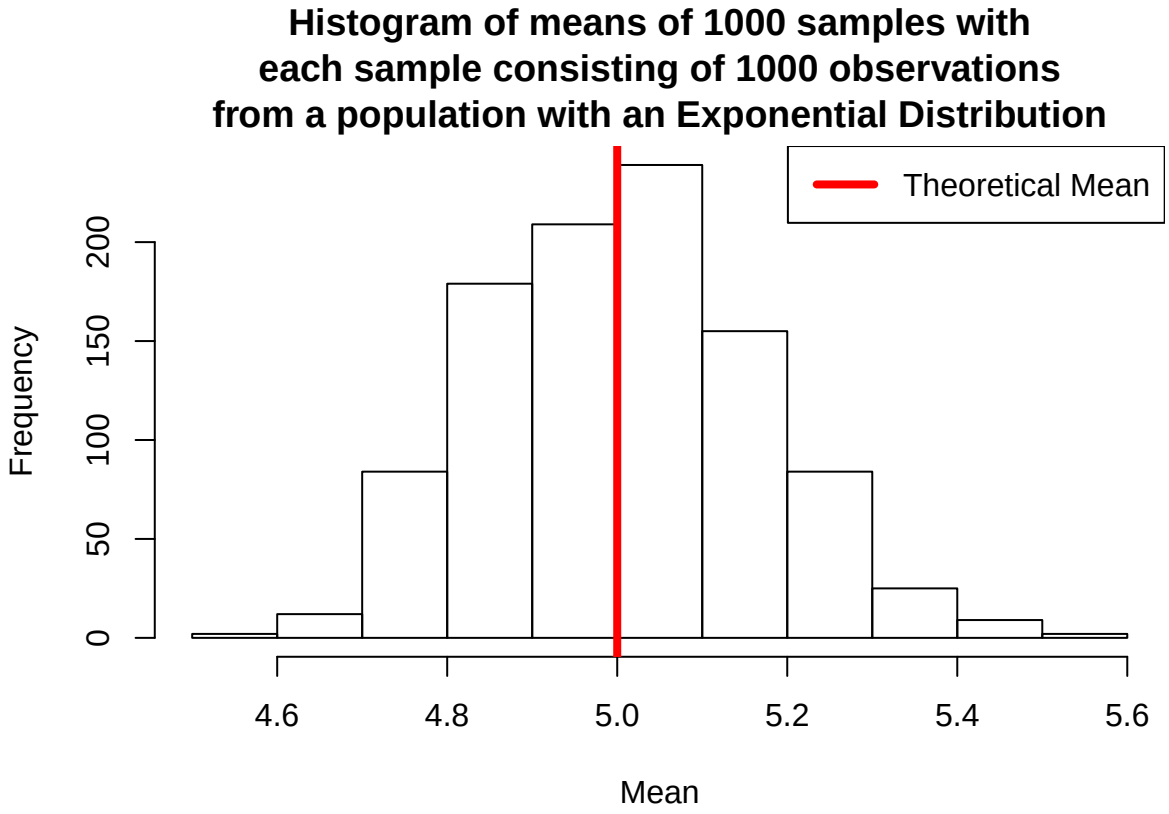
As demonstrated above the population mean in supposed to be 5 and the standard deviation of the population mean is also supposed to be 5.

```
#mns is a list containing 1000 mean values of 1000 observations randomly
#generated from a exponential distribution
mns=NULL
#sds is a list containing 1000 standard deviation of mean of 1000 observations
#randomly generated from the same data as above.
sds=NULL
for (i in 1 : 1000){
  x<-rexp(1000,0.2)
  mns=c(mns,mean(x))
  sds=c(sds,sd(x))
}
```

We can now plot a simple histogram to observe how the actual means from the 1000 samples are distributed. The theoretical population mean is demonstrated as a vertical red line:

```
hist(mns, main=paste0("Histogram of means of 1000 samples with\n",
                      "each sample consisting of 1000 observations\n",
                    "from a population with an Exponential Distribution"),
```

```
     xlab="Mean")
abline(v=5,col=2,lwd=4)
legend("topright", "Theoretical Mean", col=2, lwd=4)
```

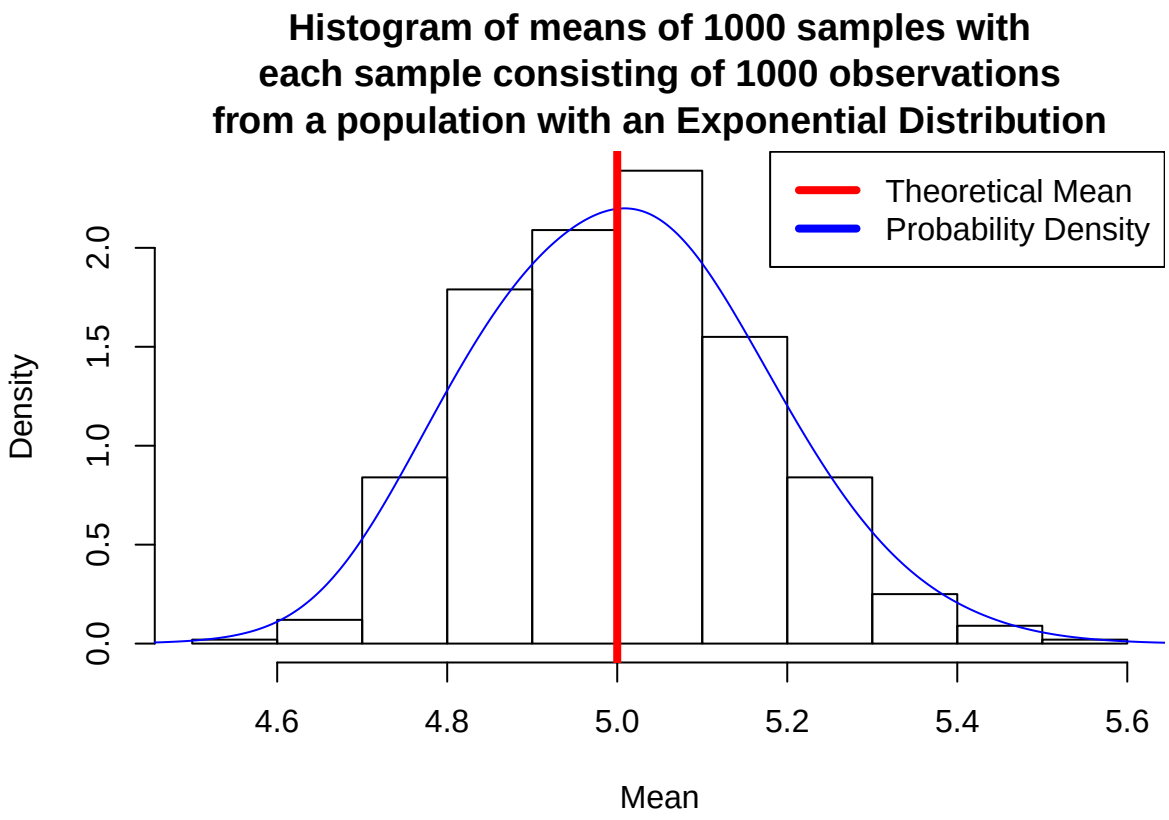


As can be seen from the histogram above, the generated distribution is fairly symmetrical around the theoretical population mean.

Further, there are two ways to demonstrate that the sample means are normally distributed.

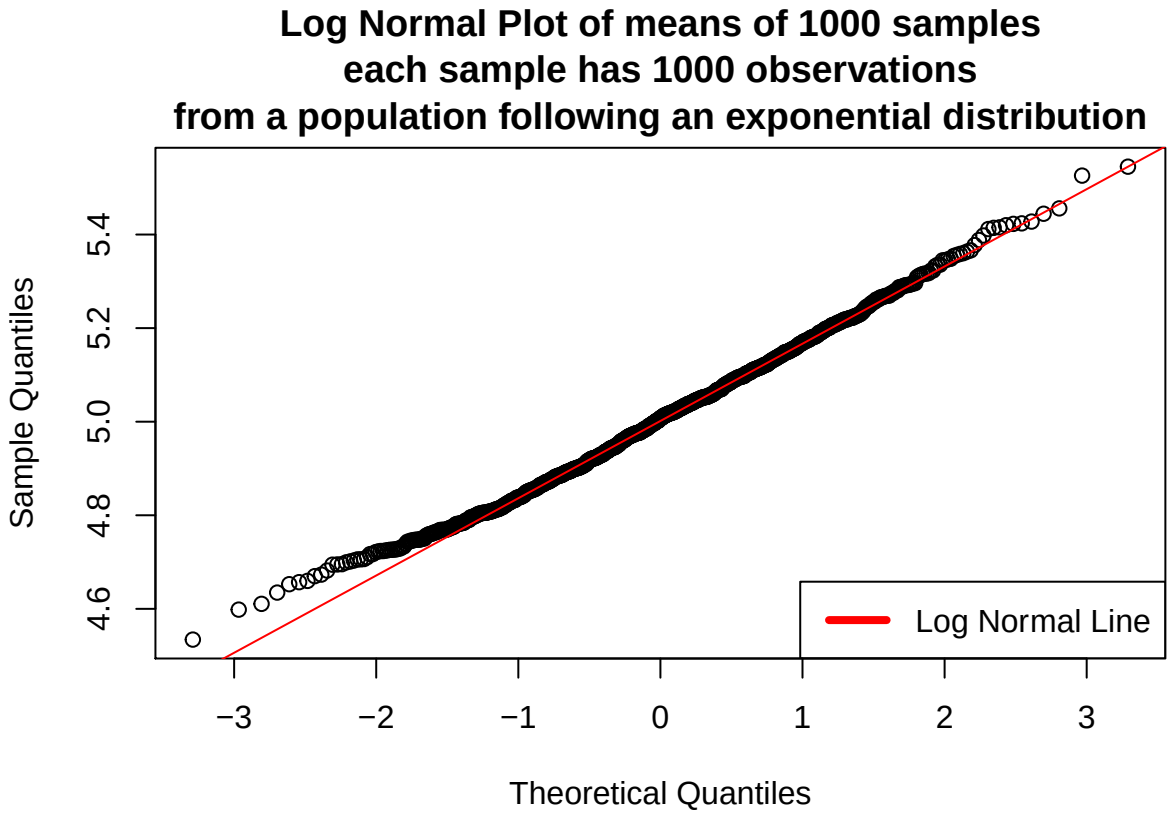
One is to convert the frequencies of the means into probabilities and plot this data as a histogram and superimpose onto this histogram, a density line representing probability density:

```
hist(mns, main=paste0("Histogram of means of 1000 samples with\n",
                      "each sample consisting of 1000 observations\n",
                    "from a population with an Exponential Distribution"),
     xlab="Mean",
     probability = TRUE)
lines(density(mns,adjust=2),col=4)
abline(v=5,col=2,lwd=4)
legend("topright", c("Theoretical Mean","Probability Density"),col=c(2,4),lwd=4)
```



Another way is to use the qqnorm function to generate a log normal plot of the data. If the plot points fall on a straight line the data follows a normal distribution.

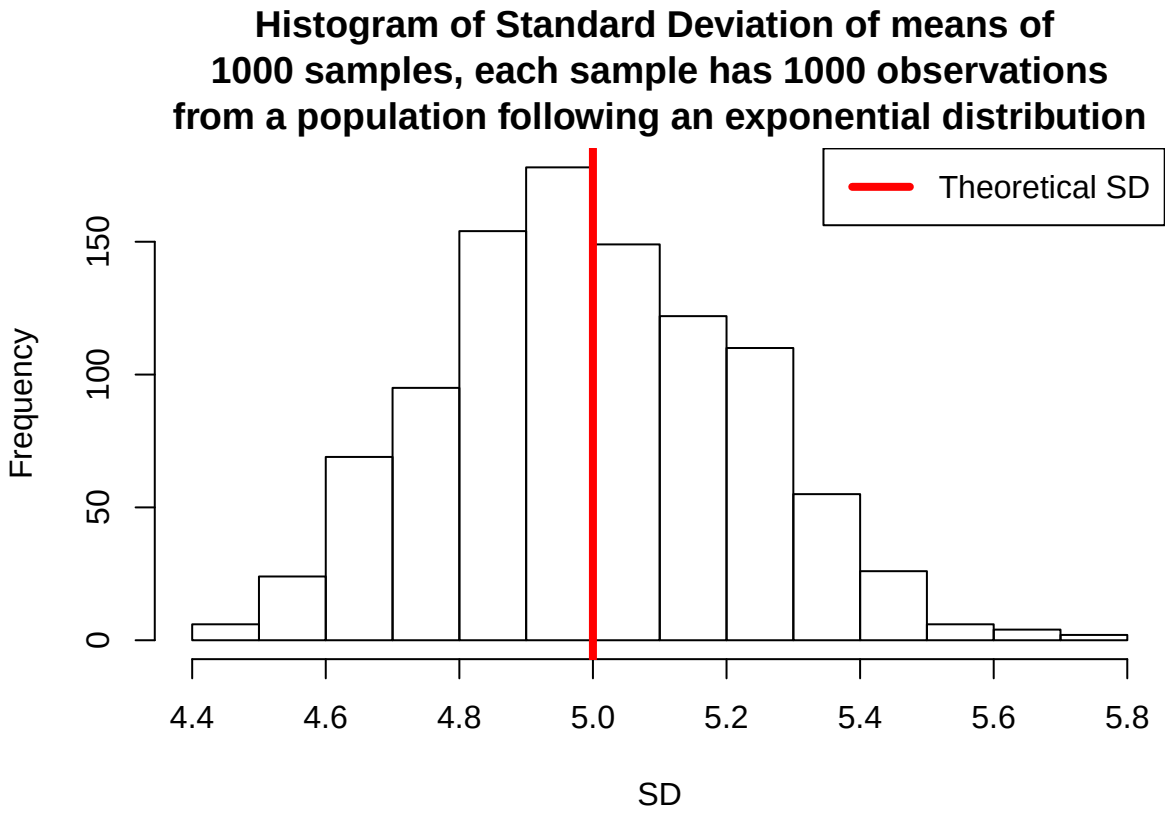
```
qqnorm(mns, main=paste0("Log Normal Plot of means of 1000 samples\n",
                        "each sample has 1000 observations\n",
                    "from a population following an exponential distribution"))
qqline(mns,col=2)
legend("bottomright", "Log Normal Line",col=2,lwd=4)
```



Similarly we can prove that the standard deviations also are normally distributed by

1) Showing that the standard deviations are symmetrerical around the population SD.

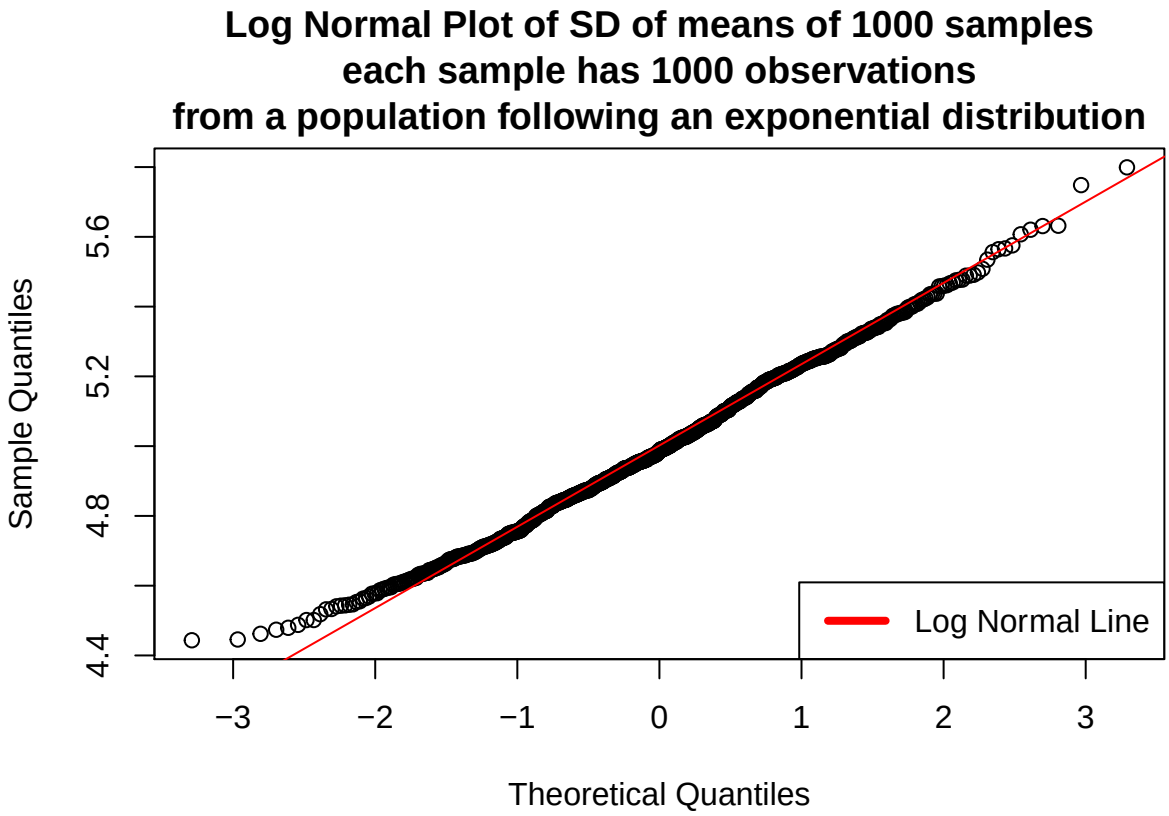
```
hist(sds, main=paste0("Histogram of Standard Deviation of means of \n",
                      "1000 samples, each sample has 1000 observations\n",
                    "from a population following an exponential distribution"),
     xlab="SD")
abline(v=5,col=2,lwd=4)
legend("topright", "Theoretical SD", col=2, lwd=4)
```



and

2) By plotting the log normal plot of the SD values.

```
qqnorm(sds, main=paste0("Log Normal Plot of SD of means of 1000 samples\n",
                        "each sample has 1000 observations\n",
                      "from a population following an exponential distribution"))
qqline(sds,col=2)
legend("bottomright", "Log Normal Line",col=2,lwd=4)
```



Conclusion

We have shown that the law of large is followed for a sample set of 1000 observations.
Furthermore we have also showed as to how the central limit theorem is satisfied for both the mean and the standard deviation of simulated 1000 samples. The original population is exponentially distributed but the means and SD's of all samples generated from this exponential distribution follow a normal distribution centered at the population mean and SD respectively.