Statistical Inference: Simulation Project

Overview

This project will use simulation to accomplish the following:

1) Demonstrate that the means and variances from a sample of random variables produced from the exponential distribution, approximate the corresponding theoretical values.

2) Show that the distribution of sample means is approximately normally distributed with a mean and variance in line with the central limit theorem.

Simulations

In this section I have generated 1000 samples each of size 40 and kept the mean and variance for each sample.

```
# declaring sample size, number of samples and exponential distribution rate
ss <- 40
n <- 1000
lambda = 0.2
# calculating the theoretical and mean and variance for this value of lambda
tmean = 1/lambda
tvar = (1/lambda)^2
set.seed(1)
# collecting 1000 sample means and 1000 sample variances for the
# exponential distribution
smean <- numeric()
svar <- numeric()
for (i in 1 : n) {
  svariables = rexp(ss, rate = lambda)
  smean = c(smean, mean(svariables))
  svar = c(svar,var(svariables))
}
# Creating a data frame with these values for plotting etc
sim_number <-seq(1:n)
simulations <- data.frame(sim_number,smean,svar)
```

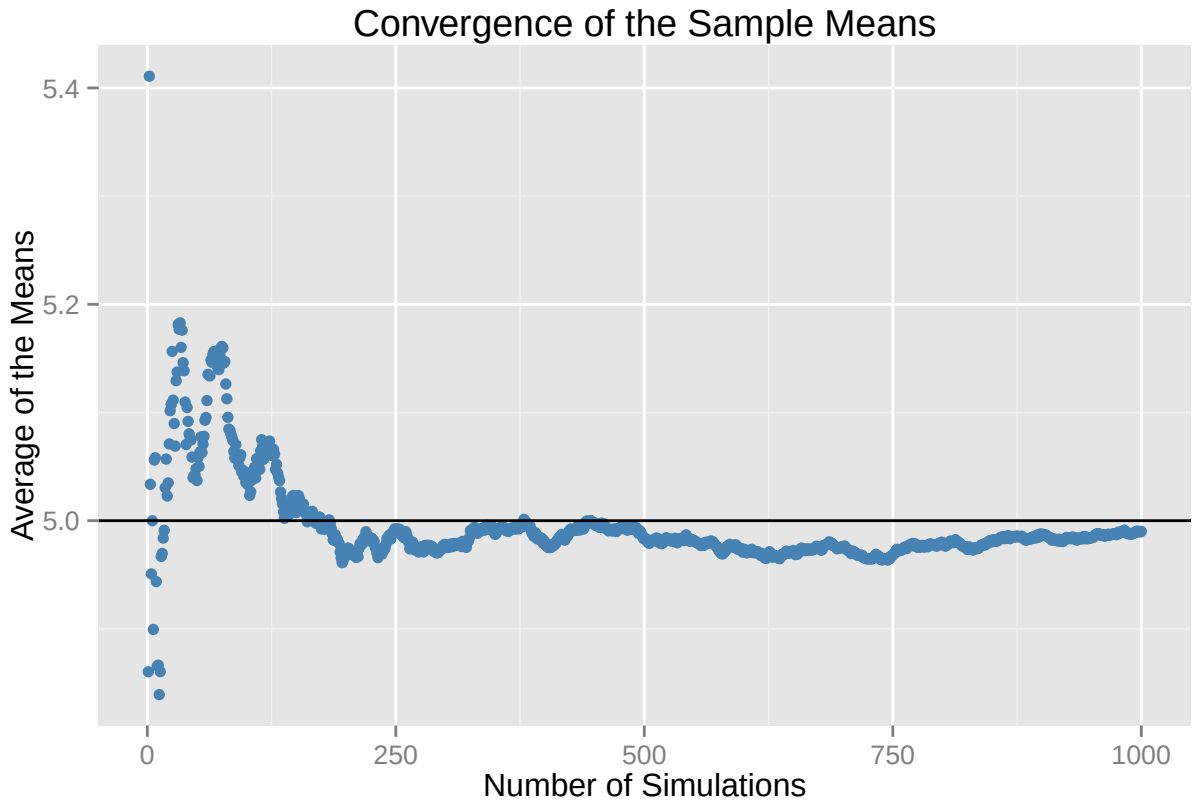
According the law of large numbers consistent estimators should converge to their corresponding population values for large samples. So the sample mean should converge to the population mean and the sample variance should converge to the population variance.

So each sample mean and sample variance is approximating the population variance. Therefore if we average the sample means and variances, the averages should get closer to the population means and variances the more samples we take.

The plots below show the averages for 1 sample up to 1000 samples.

Sample Means Versus Theoretical Mean

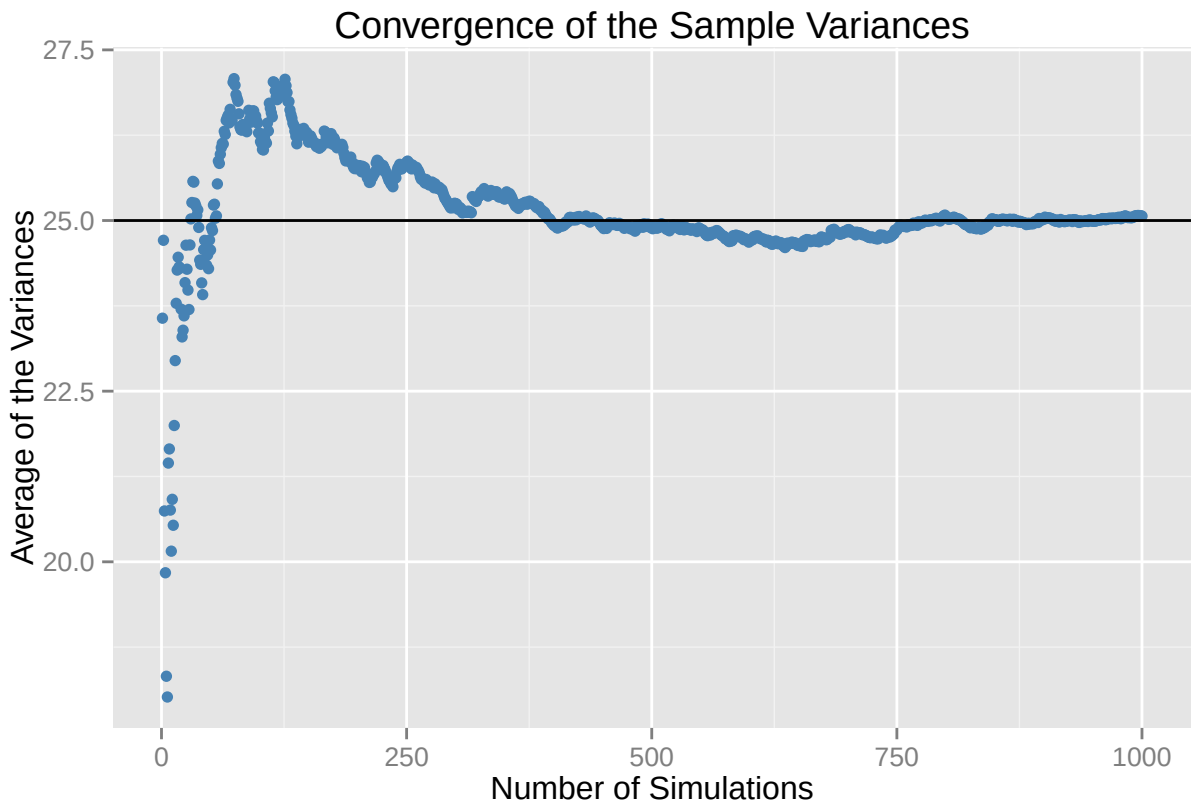
```r
# Calculating the sample averages
simulations$average_means <-cumsum(simulations$smean[1:n])/(1:n)
# plotting and including a horizontal line for the population mean
library(ggplot2)
# Creating the convergence plot
g <- ggplot(simulations,aes(sim_number,average_means))
g <- g + geom_point(color = "steelblue")
g <- g + geom_hline(yintercept = tmean)
g <- g + labs(x="Number of Simulations",
              y="Average of the Means",
              title="Convergence of the Sample Means")
g
```



Sample Variance Versus Theoretical Variance

```r
# Calculating the sample averages
simulations$average_vars <-cumsum(simulations$svar[1:n])/(1:n)
# plotting and including a horizontal line for the population mean
library(ggplot2)
# Creating the convergence plot
g <- ggplot(simulations,aes(sim_number,average_vars))
g <- g + geom_point(color = "steelblue")
g <- g + geom_hline(yintercept = tvar)
g <- g + labs(x="Number of Simulations",
              y="Average of the Variances",
```

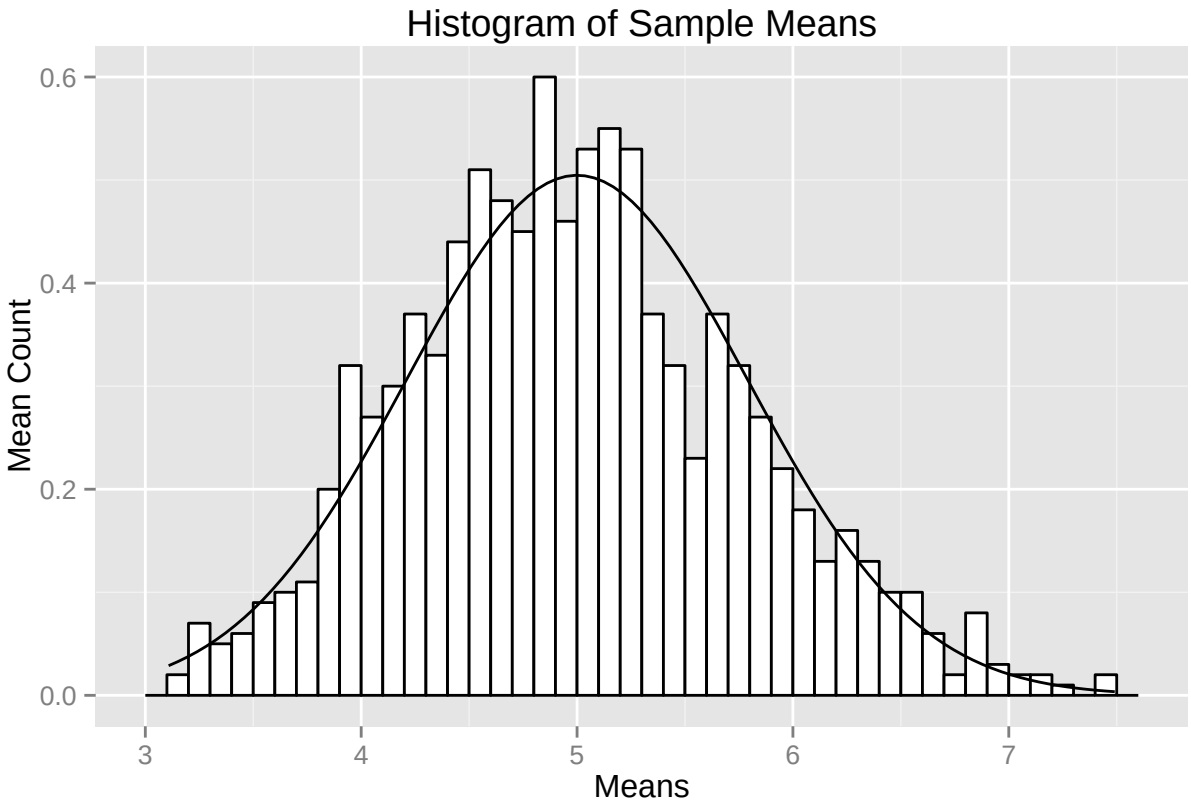
```
             title="Convergence of the Sample Variances")
g
```



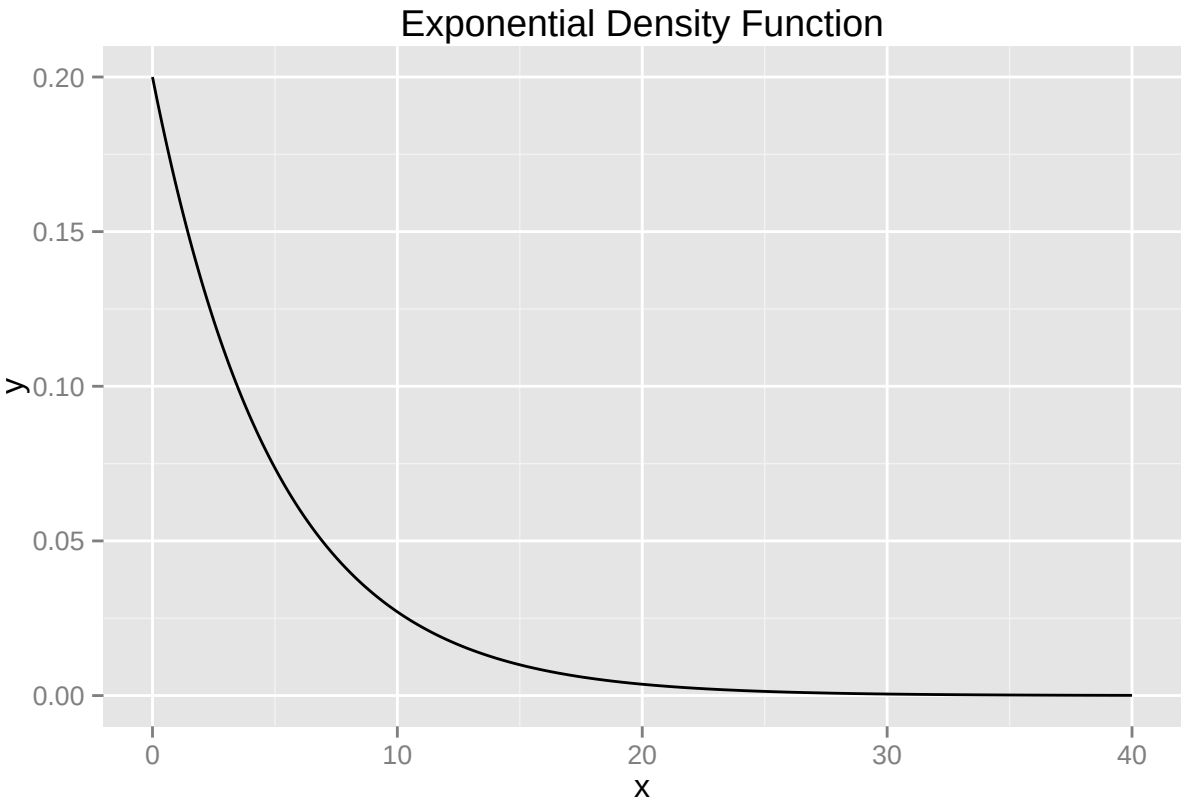
Distribution Comparison

According to the central limit theorem the distribution of sample means should be normal with a mean of 5 (= 1/lambda) and a variance of 25/40. To verify this I will plot a histogram of the sample means and overlay this with a plot of the density function for the normal distribution with the mean and variance stated. For comparison I will produce a further plot of the exponential distribution.

```
# Producing a plot of the sample mean histogram overlayed with corresponding
# normal distribution as predicted by the CLT
g <- ggplot(simulations,aes(x=smean))
g <- g + geom_histogram(color="black",
                        fill="white",
                        binwidth=0.1,
                        aes(y = ..density..))
g <- g + labs(x="Means",
              y="Mean Count",
              title="Histogram of Sample Means")
g <- g + stat_function(fun = dnorm,
                       arg=list(mean = tmean,
                                sd = sqrt(tvar/ss)))
g
```




```
# For comparison here is a plot the exponential density function
x <- seq(0,40,by=0.1)
y <- dexp(x,rate=lambda)
df <- data.frame(x,y)
g <- ggplot(df,aes(x,y))
g <- g + geom_line()
g <- g + labs(title="Exponential Density Function")
g
```



Conclusion

Clearly the distribution of the sampling means is much closer to the normal distribution specified by the CLT than it is to the exponential distribution from which the variables in the samples where drawn. This clearly verifies the central limit theorem.