Statisticall Learning Project (Part I)

By Bin Jia

Overview

This project amis to investigate the distribution of averages of 40 exponentials in R. It is shown that the distribution is approximately normal, which is consistent to the Central Limit Theorem.

Preparing the data

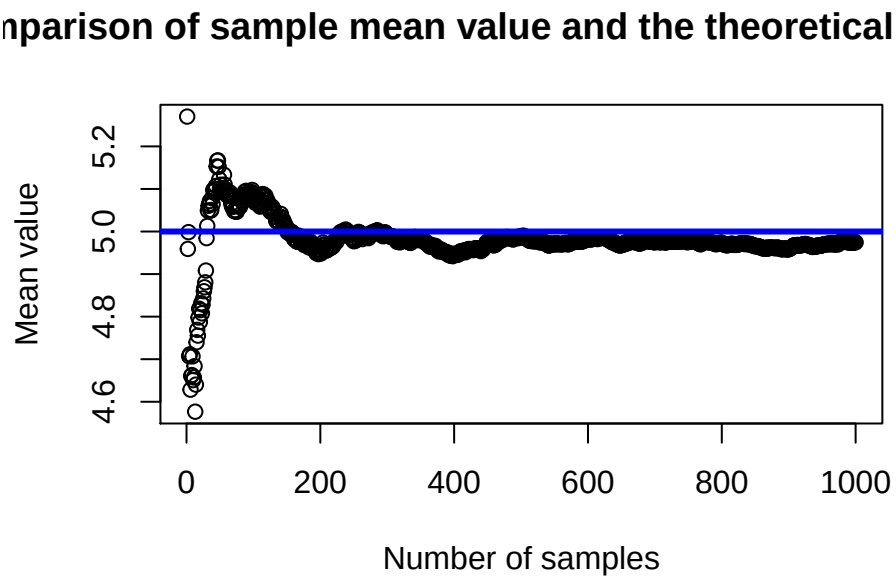
Generate samples of averages of 40 exponentials

```
lambda <- 0.2
mns = NULL
n <- 1000
for (i in 1 : 1000) mns = c(mns, mean(rexp(40, lambda)))
```

Sample Mean versus Theoretical Mean

Compute the difference between the sample mean value and the theoretical mean of the distribution

```
means <- cumsum(mns)/(1:n)
plot(seq(1:n),means,main="Comparison of sample mean value and the theoretical mean value",
  xlab="Number of samples", ylab="Mean value")
abline(5,0, col = "blue", lwd = 3)
```



In the above figure, the blue line denotes the theoretical mean value while the black dot denotes the sample mean corresponding to different number of samples. It can be seen that the sample mean approaches the theoretical mean with the increasing of the number of samples, which is consistent to the Central Limit Theorem.

```
theortical_mean <- 1/lambda
mean(mns) - theortical_mean
```

```
## [1] -0.02544971
```

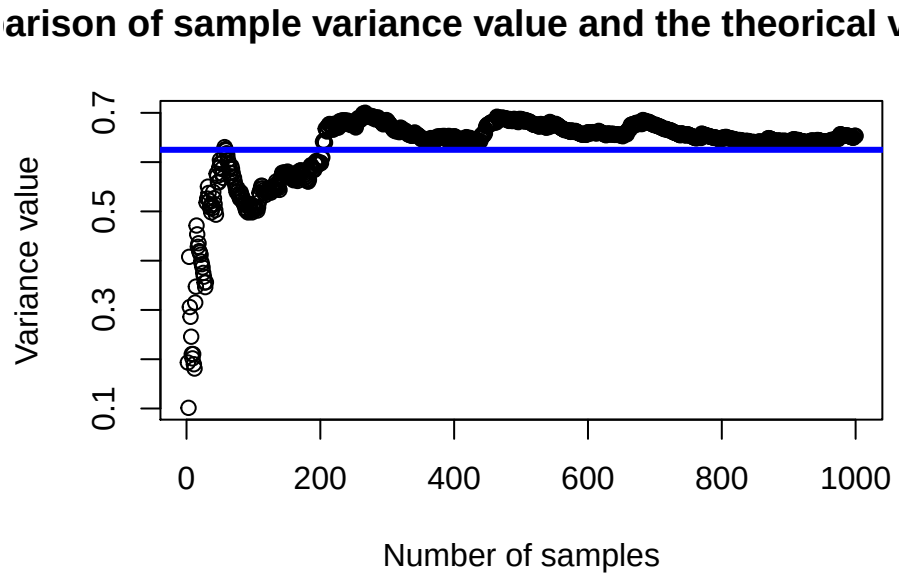
It can be seen that the theoretical mean is very close to the mean of the samples.

Sample Varaince versus Theoretical Varaince

Compute the difference between the sample variance value and the theorical variance of the distribution

```
sample_variance <- seq(1,n)
for(i in seq(1,n)){
  sample_variance[i] <- sd(mns[seq(1,i)])^2
}


theortical_variance <- (1/lambda/sqrt(40))^2
plot(seq(1:n),sample_variance,
     main="Comparison of sample variance value and the theorical variance value",
 xlab="Number of samples", ylab="Variance value")
abline(theortical_variance,0, col = "blue", lwd = 3)
```

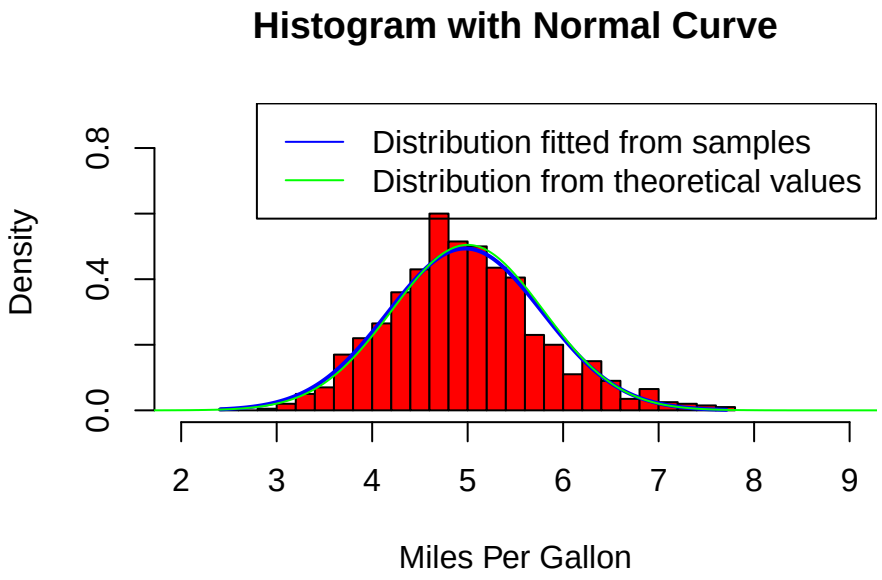


In the above figure, the blue line denotes the theoretical variance value while the black dot denotes the sample variance corresponding to different number of samples. It can be seen that the sample variance approaches the theoretical variance with the increasing of the number of samples, which is consistent to the Central Limit Theorem.

Distribution

Draw the fitted distribution with sample mean and sample variance and the distribution with the theoretical mean and variance

```
h<-hist(mns, prob=TRUE, breaks=20, col="red", xlab="Miles Per Gallon",
   main="Histogram with Normal Curve", xlim = c(2,9), ylim = c(0,0.9))
xfit<-seq(min(mns),max(mns),length=400)
yfit<-dnorm(xfit,mean=mean(mns),sd=sd(mns))
lines(xfit, yfit, col="blue", lwd=2)
x <- seq(0,20,length.out=1000)
lines(x,dnorm(x,theortical_mean,sqrt(theortical_variance)),col="green")
legend("topright",col = c("blue",'green'),
       c("Distribution fitted from samples", "Distribution from theoretical values"),
       lwd = 1)
```



As shwon in above figure, it can be seen that the shape of distribution with sample mean and sample variance is very CLOSE to the shape of the distribution with theoretical mean and variance. The result is consistent to the Central Limit Theorem.