Statistical Inference Course Project - Part 1

Amit Kohli

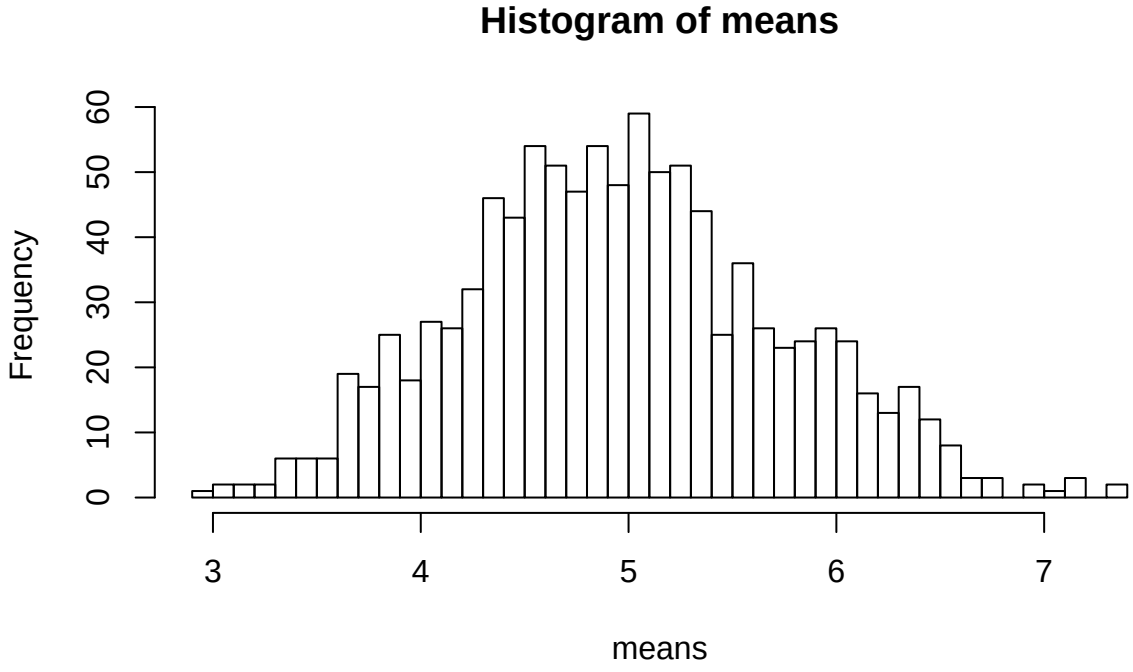
February 3, 2016

Overview:

This reports investigates the exponential distribution in R and compares it with the Central Limit Theorem. The report comprises of a simulation exercises to explore the properties of the distribution of averages of 40 exponentials. The sample and theoritical mean and variance are compared to those of theoretical, and it is concluded that the distribution is approximately normal.

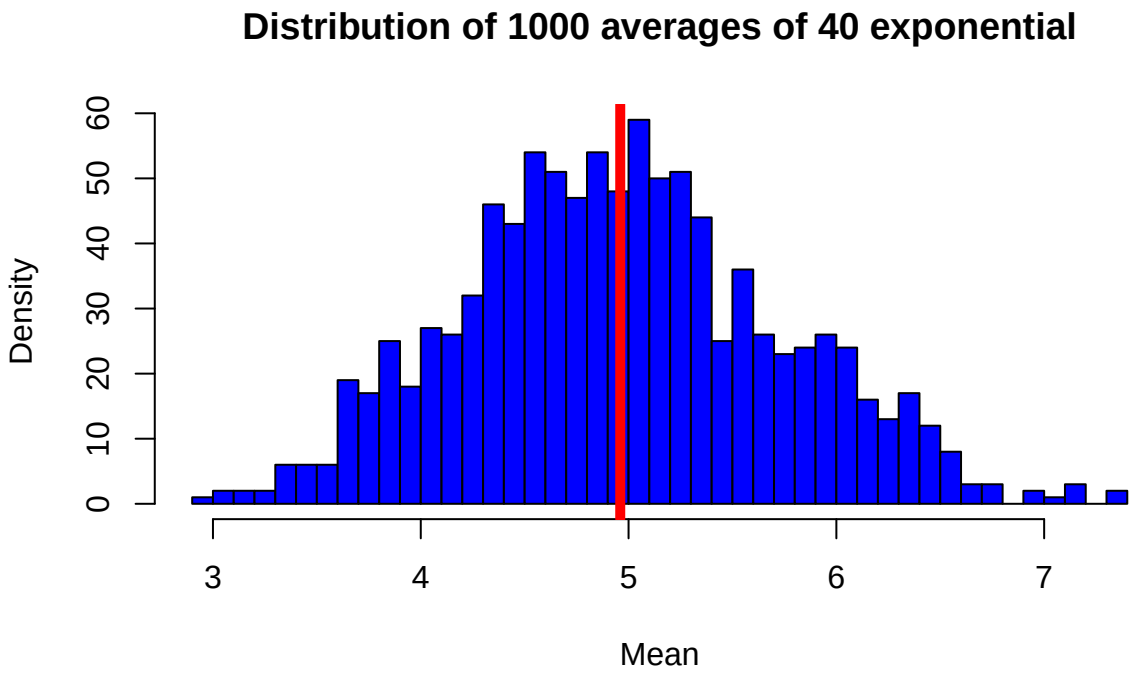
Simulations:

The exponential distribution is simulated in R with rexp(n, lambda) where lambda is the rate parameter, where both the mean and standard deviation of the distribution are 1/lambda. Lambda is set to 0.2 for all of the simulations in which the distribution of averages of 40 exponentials with 1000 simulations are investigated. In this section, we first set the seed (Februray 3, 2016), lambda (0.2), and exponential size (40). Iterating results over 1000 times, the rexp function is used to find averages. The Histogram below shows mean of 1000 random exponentials.



Sample Mean versus Theoretical Mean:

In this section, the report presents the distribution of the sample means and then address the questions regarding the differences between the simulation distribution and theoretical normal distribution. The sample and theoretical means are reported for the comparison.

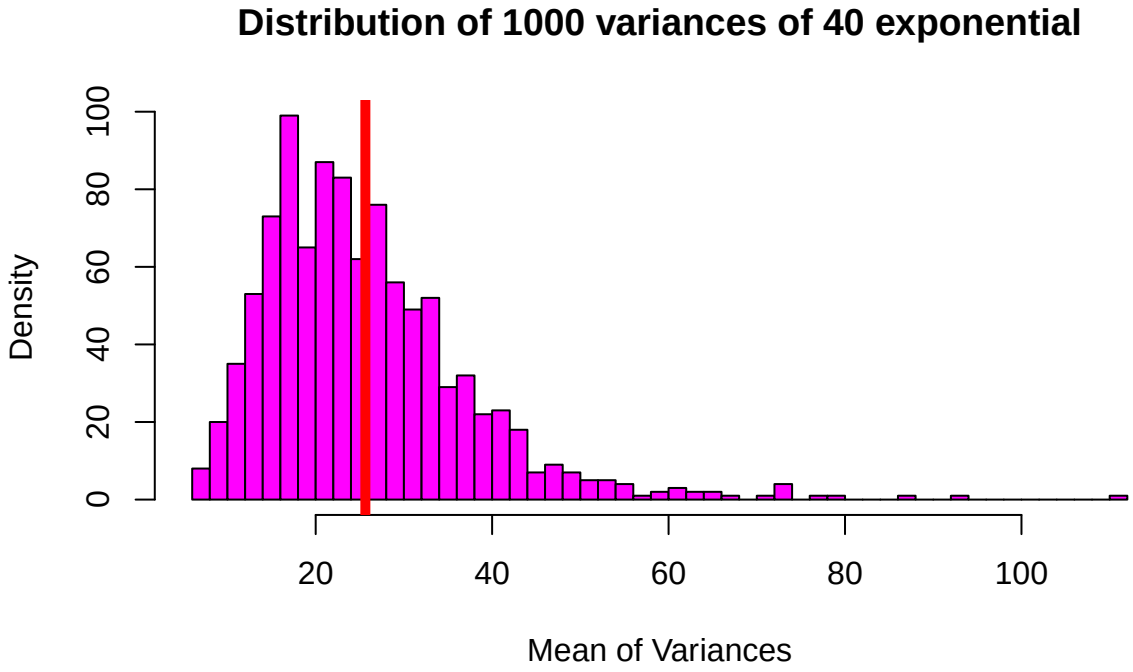


Conclusion: Referring to the theoretical and sample mean values, the two values very close.

```
##      Theoretical Sample
## Mean           5   4.96
```

Sample Variance versus Theoretical Variance:

The mean of variances of 40 random exponential distribution for 1000 simulation runs is plotted as follows.

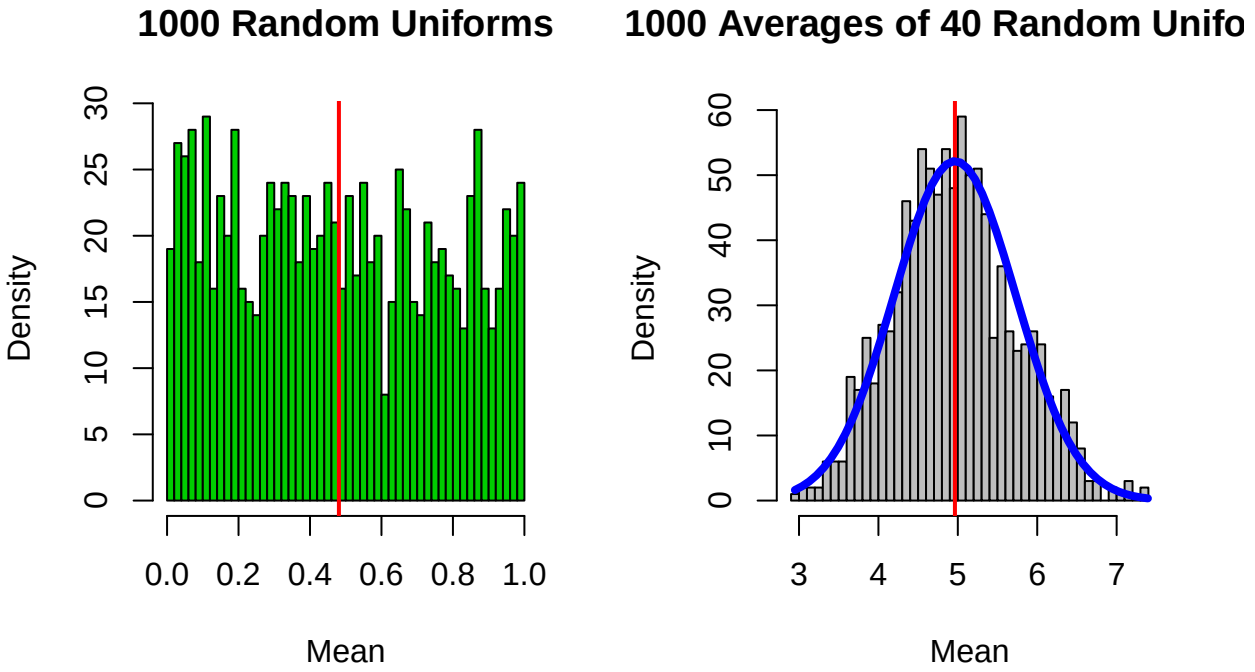


Conclusion: It is observable in the graph that the sample variance is very close to the theoritical variance.

```
##          Theoretical Sample
## Variance          25  25.63
```

Distribution:

To answer the third question for investigating whether the distibution looks approximately like normal distribution, two different simulations are run and corresponding graphs are plotted. The first graph shows the distrbution of 1000 random expnoentilas, and the second graph shows the everage of 1000 simulations of 40 exponentials as follws.



Conclusion: As it can be seen in the last two diagrams, both graphs have similar sample-means around the middle of the plot. However, the average of 40 exponentials for a large number of simulations (corresponding to the right-side plot) looks like normal distribution while the collection of random unifroms (left-side graph) looks different from a normal distribution.

This simulation confirms the Central Limit Theorem (CLT), as the distribution of averages of a distribution behaves like a normal distibution for the sufficiently large smaple sizes.

Code for Simulations:

```
set.seed(20160203)
lambda <- 0.2
size <- 40
means <- NULL
variance <- NULL

for (i in 1:1000) means = c(means, mean(rexp(size, lambda)))
for (i in 1:1000) variance = c(variance, var(rexp(size, lambda)))
hist(means,breaks=50)
```

Code for Sample Mean versus Theoretical Mean:

```
sample_mean <- round(mean(means), 2)
theo_mean <- 1/lambda
hist(means, main="Distribution of 1000 averages of 40 exponential",
     xlab= "Mean", ylab="Density",
     breaks= 50, col="blue")
abline(v= sample_mean, col="red",lwd= 5)

matrix(data=c(theo_mean, sample_mean),
       nrow=1, ncol=2, byrow=TRUE,
       dimnames=list(c("Mean"), c("Theoretical", "Sample")))
```

Code for Sample Variance versus Theoretical Variance:

```
sample_var <- round(mean(variance), 2)
theo_var <- (1/lambda)^2

hist(variance,
     main="Distribution of 1000 variances of 40 exponential",
     xlab= "Mean of Variances", ylab="Density", breaks= 50, col="magenta")
abline(v= sample_var, col="red",lwd= 5)

matrix(data=c(theo_var, sample_var), nrow=1, ncol=2, byrow=TRUE,
       dimnames=list(c("Variance"), c("Theoretical", "Sample")))
```

Code for Distribution:

```r
x <- runif(1000)
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(runif(40)))

par(mfrow=c(1,2))

hist(x, breaks = 50,main="1000 Random Uniforms",
  xlab= "Mean", ylab="Density", col="Green3")
abline(v= mean(x), col="red",lwd= 2)

hg<-hist(means,breaks= 50,col="grey",
         main="1000 Averages of 40 Random Uniforms",
         xlab= "Mean",ylab="Density")
x_fit<-seq(min(means),max(means),length= 50)
y_fit<-dnorm(x_fit,mean=mean(means), sd=sd(means))
y_fit<-y_fit*diff(hg$mids[1:2])*length(means)
abline(v=mean(means),col="red",lwd=2)
lines(x_fit, y_fit,col="blue",lwd=4)
```

GitHub Repo