Statistical Inference Course Project – Part 1

Author: Venkat Ram Rao

Overview: Analyze the exponential distribution in R (*rexp(n, lambda) with lambda = 0.2*) and compare it with the Central Limit Theorem.

The properties of the expression are as follows: Lambda = **0.2**; Expected Mean = 1/Lambda = **5**; Expected Standard Deviation = 1/Lambda = **5**; Expected Variance = **25**

Analysis: Analysis was done by running 1000 simulations for a size 40 each. The Mean and Variance of each run was calculated. The following code does this(full R code in Appendix Item 1):

```
mns = NULL
vars = NULL
for (i in 1 : 1000)
  {
  mns = c(mns, mean(rexp(40,.2)))
  vars = c(vars, sd(rexp(40,.2))^2)
}
```

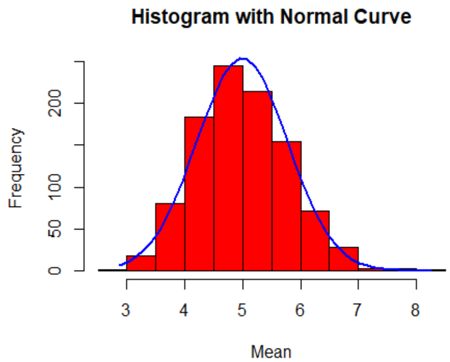
After running the simulations, the resulting Mean and Variance were plotted. I overlayed the Mean and variance with a Normal distribution. As you can see, both follow the Normal distribution closely. The following code does this (full R code in Appendix Item 2):

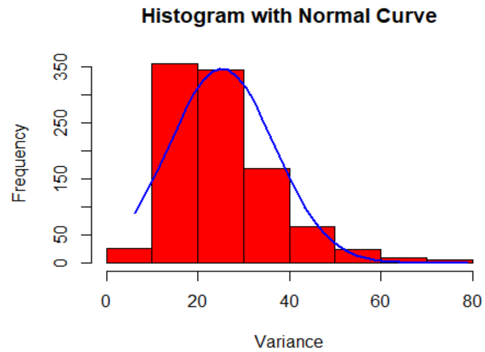
Mean:

```
h<-hist(mns, breaks=10, col="red", xlab="Mean",
   main="Histogram with Normal Curve")
xfit<-seq(min(mns),max(mns),length=40)
yfit<-dnorm(xfit,mean=mean(mns),sd=sd(mns))
yfit <- yfit*diff(h$mids[1:2])*length(mns)
lines(xfit, yfit, col="blue", lwd=2)
```

Variance:

```
h<-hist(vars, breaks=10, col="red", xlab="Variance",
   main="Histogram with Normal Curve")
xfit<-seq(min(vars),max(vars),length=40)
yfit<-dnorm(xfit,mean=mean(vars),sd=sd(vars))
yfit <- yfit*diff(h$mids[1:2])*length(vars)
lines(xfit, yfit, col="blue", lwd=2)
```

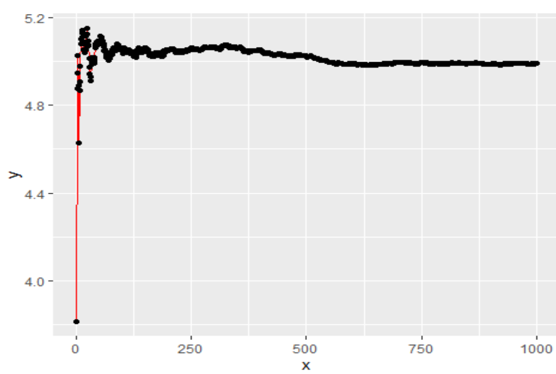




Sample Mean vs Theoretical: Mean of Means: **4.991714**(Theoretical **5**) with a Standard Deviation of **0.7856894**

Sample Variance vs Theoretical Mean of the Variance: **25.23343**(theoretical **25**) with a Standard Deviation of **11.50802**

Finally, tracking the Cumulative means as the number of samples increased showed that the Mean converges to the value of 5: (R code in Appendix Item 3)



Summary: Running the simulation multiple times resulted in similar end results each time. This provides great confidence that the overall Mean of the expression - rexp(n, lambda) - is indeed **5** for lambda = 0.2 and the Variance is **25**.

APPENDIX

Part1:

1) **Generate Sample data:**

```
mns = NULL
vars = NULL
for (i in 1 : 1000)
  {
  mns = c(mns, mean(rexp(40,.2)))
  vars = c(vars, sd(rexp(40,.2))^2)
}

mean(mns)

## [1] 4.991714

sd(mns)

## [1] 0.7856894

mean(vars)

## [1] 25.23343

sd(vars)

## [1] 11.50802
```

2) **Plots:**

a) Plot of Distribution of Means (overlayed with Normal distribution)

```
h<-hist(mns, breaks=10, col="red", xlab="Mean",
   main="Histogram with Normal Curve")
xfit<-seq(min(mns),max(mns),length=40)
yfit<-dnorm(xfit,mean=mean(mns),sd=sd(mns))
yfit <- yfit*diff(h$mids[1:2])*length(mns)
lines(xfit, yfit, col="blue", lwd=2)
```

b) Plot of Distribution of Variances (overlayed with Normal distribution)

```
h<-hist(vars, breaks=10, col="red", xlab="Variance",
   main="Histogram with Normal Curve")
xfit<-seq(min(vars),max(vars),length=40)
yfit<-dnorm(xfit,mean=mean(vars),sd=sd(vars))
yfit <- yfit*diff(h$mids[1:2])*length(vars)
lines(xfit, yfit, col="blue", lwd=2)
```

3) ***Plotting the* cumulative *mean as the number of samples increase.***

```r
library(ggplot2)
avg=NULL
n <- 1000
avg <- cumsum(mns)/(1:n)
 df <- data.frame(x=1:n,y=avg[1:n])

ggplot(data=df, aes(x=x, y=y)) +
  geom_line(color="red")+
  geom_point()
```