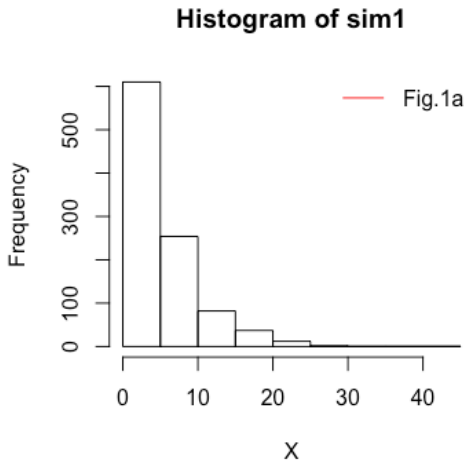
# Demonstration the Central Limit Theorem using simulation of the exponential distribution

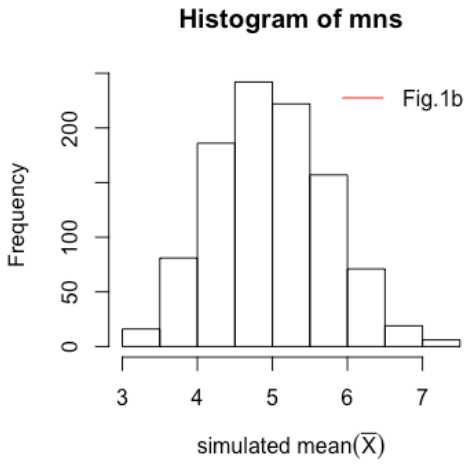
author: "Y.V.Wang"
date: "November 18, 2015"

## Overview:

I investigated the 1000 simulation of exponential distribution in R for random sample size of 40 exponentials. This is an exerise to prove the central liimit theroem by using simulation. It is clear the means of 40 exponentials behave as predicted by the Central limit Theorem and the sample means are approximiately normally distributed. The simulated sample mean of $\mu_{\overline{x}}$ equals to population mean $\mu$. The variance of $\overline{x}$ equals to population variance divided by simple size n.

## Simulation: Testing the central Limit Theorem: Means of 40 Exponentials (please see Appendix 1 for the simulation and plotting code)




**Interpretation of above figures**: *Fig.1a is histogram for 1000 random exponential, we can see it is far from normally distirbuted and it is clearly right skewed. Where Fig.1b is a histogram of sampel means for 1000 samples of 40 random exponentials. It is clear that sample distribution of the sampel mean can be approaximated by a normal distribtuon when sample size is relatively large. Use our simulation, for sample size of 40, the sample mean distribution is close to normal distribution (more discussion later).*

## Sample Mean versus Theoretical Mean

```
# put the simulated mean into a dataframe
sampleMeanDF<-data.frame(mns)
attach(sampleMeanDF) # attach the data frame
```

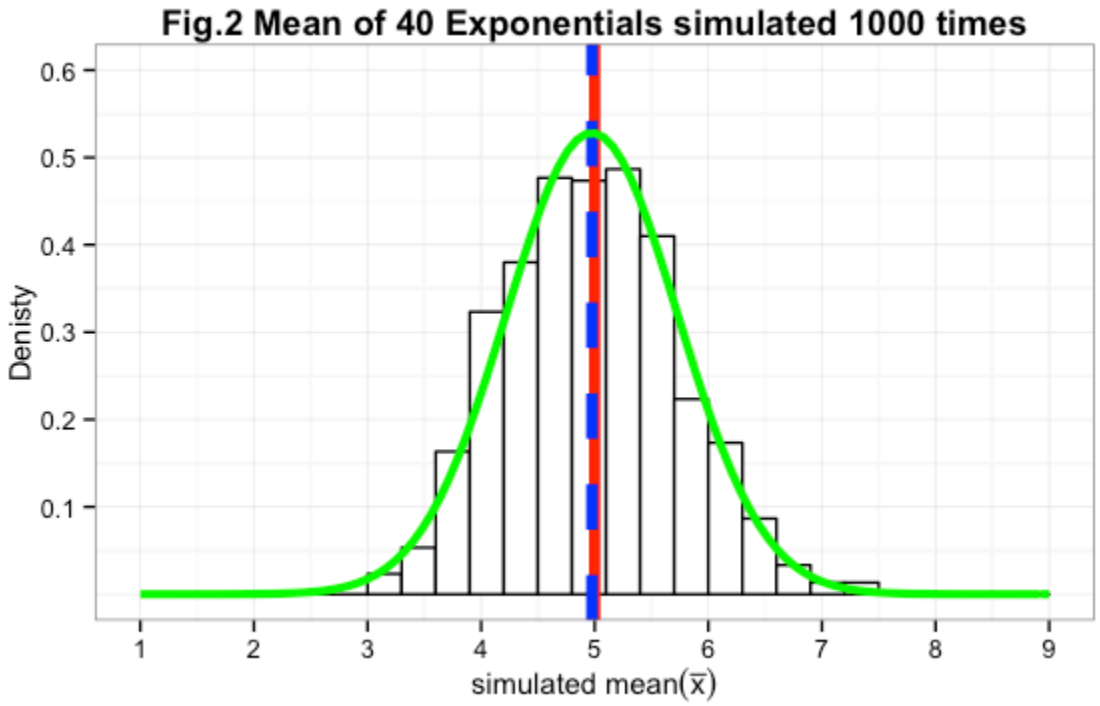
```r
#calucate the theorical mean: because we expect with the increased simu
lation sample sizes, the distribtion will be normaly distributed, and t
he sample mean will center around the population mean. Therefore the th
eoretical mean equal population mean (1/lambda).
realMean=1/0.2

#cacluate the simulated 1000 sample mean
sampleMean<-round(mean(mns), 2)

#compare the theoretical mean and sample mean
meanComparison<-rbind(realMean,sampleMean)
colnames(meanComparison)<-"mean_values"
print(meanComparison) # print out a table to show the theoretical and s
ample means

##            mean_values
## realMean          5.00
## sampleMean        4.97
```

In Fig2. I marked the theoretical mean and sample mean using vertical lines in the fo llowing figure. In addition, the normal distribution curve is superimposed on the his togram of 1000 simulations of 40 sample distribution (**The code for making the pl ot is in Appendix 2**)




**Interpretation**: *it is clear that the theoretical mean and the sample mean is almost identical. The theoretical mean is 5, and the sample mean is also 5 (almost). This is according to Central Limit Theorem, if the sample size is large, then* $\bar{x}$ *is also*

*approximately normally distributed. The sample mean of $\overline{x}$ equal to the population mean $\mu$.*

### Distribution

*From the Fig.2 and Fig.1b, we can see the 1000 simulations of 40 sample is approximately normal (see the superimposed normal distribution). This is exactly how the central limit Theorem behave when we increase the sample sizes. Because the Central Limite Theorem tells us that if we take a bunch of samples of size n, and compute the mean of those n sample, the sample means have predictable distribution and they will be normally distributed, with their mean equal to the population mean, and their variance equal to the populaiton variance divided by n.*

### Sample Variance versus Theoretical Variance

```
# calculate the population variance, the population variance is defined
by 1/(lambda^2)
varPop<-1/(0.2^2)
# calculate the theoretical variance for the sample mean. According to
the Central Limite Theorem, the theoretical variance equal to the popul
aiton variance divided by n.
varTheory<-varPop/40
# calucate the actual variance of the 1000 sample means using R funcito
n var().
varSample<-var(mns)

varComparison<-rbind(varTheory,varSample)
colnames(varComparison)<-"variance_value"
print.table(varComparison)

##           variance_value
## varTheory      0.6250000
## varSample      0.5706551
```

**In conclusion**: *The theoretical variance of 40 samples and the 1000 sample mean variance are very close, which this suggests the the Central Limit Theorem is working. Because if we suppose a variable of x of a population is normally distributed with population mean, then for samples of size n, the variable $\overline{x}$ is also normally distirbuted and has mean $\mu$, and variance $\sigma^2/n$. With increasing sample sizes.*

~end of report

## Appendices

1. Simulation code for testing the central Limit Theorem: Means of 40 Exponentials

```
# simulation of the distribution of 1000 random exponentials, n=1000, a
nd rate=0.2.
lambda=0.2
sim1<-rexp(1000,lambda)

# simulation of the distribution of 1000 averages of 40 random exponent
ials. That is run 40 random variable, n=40, and rate=0.2, grab the mean
, and repeat for 1000 times.
set.seed(1234)
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(40,lambda)))

# make a graph for above simulations
par(mfcol=c(1,2))

#histogram for 1000 random exponential
hist(sim1,ylab="Frequency", xlab="1/X")
legend("topright", lty=1,col ="red",legend="Fig.1a",bty="n")

# histogram of sampel means for 1000 sampels of 40 random exponentials
hist(mns,ylab="Frequency",xlab=bquote(paste("simulated mean", (bar(X)),
"")))
legend("topright", lty=1,col ="red",legend="Fig.1b",bty="n")
```

2. Figure 2 is made by ggplot with following code:

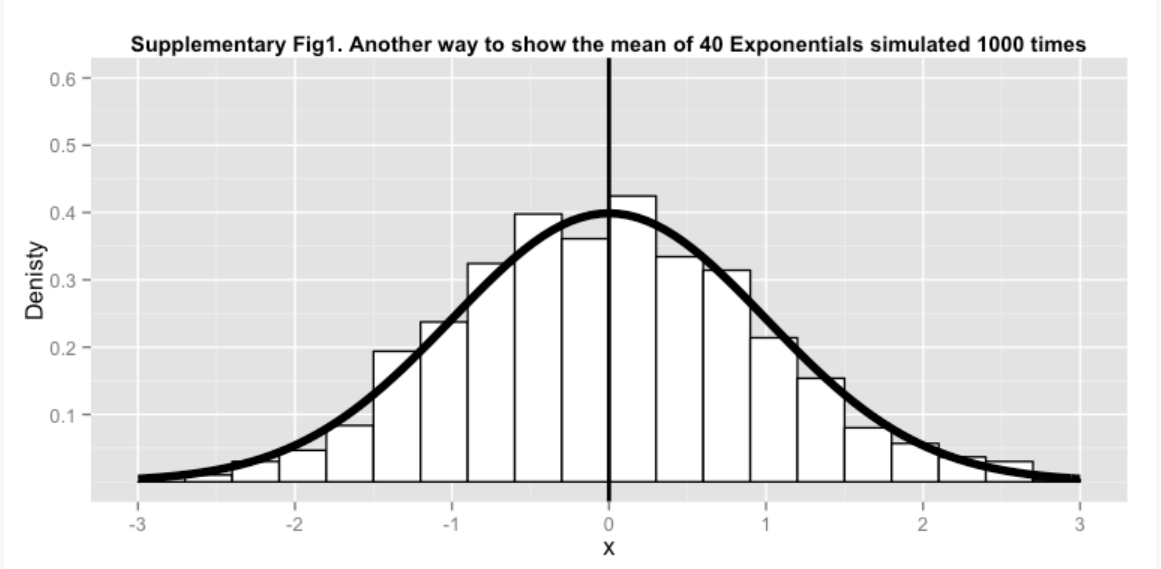
```
require(ggplot2)
ggplot(sampleMeanDF,aes(x=mns))+
        geom_histogram(binwidth=0.3, color="black",fill="white",aes(y =
..density..))+
        xlab(bquote(paste("simulated mean", (bar(x)), "")))+
        ylab("Denisty")+
        scale_x_continuous(limits=c(1, 9),breaks=c(1,2,3,4,5,6,7,8,9))+
        scale_y_continuous(limits=c(0,0.8),breaks=seq(0.1,0.8,by=0.1))+
        ggtitle("Fig.2 Mean of 40 Exponentials simulated 1000 times")+
        theme_bw()+
        theme(axis.title=element_text(size=11))+
        theme(plot.title = element_text(size = rel(1.1),face="bold"))+
        geom_vline(x =1/0.2, size = 2, color="red")+
        geom_vline(x=mean(mns),size=2,linetype="dashed",color="blue")+
        stat_function(fun=dnorm, size = 1.5, color="green",args=list(me
an =mean(mns),sd =sd(mns)))
```

3. *Please note another way to simulate the a standard normal random variable of exponenital is to apply CLT using

$$\frac{\text{Estimate} - \text{MeanofEstimate}}{\text{Std. Err. ofEstimate}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \to N(0,1)$$

```
nosim <- 1000
# convert to standard normal, since mean=5, and sample size n=40, sampl
e variance=5/sqrt(40)
cfunc <- function(x, n) 0.2 * sqrt(40) * (mean(x) - 5)
# simulate data for sample size of 40
lambda=0.2
data <- data.frame(x = apply(matrix(rexp(nosim*40,lambda), nosim), 1, c
func))
ggplot(data, aes(x = x)) +
        geom_histogram(binwidth=.3,colour = "black", fill="white", aes(
y = ..density..)) +
        scale_x_continuous(limits=c(-3, 3),breaks=c(-3,-2,-1,0,1,2,3))+
        scale_y_continuous(limits=c(0,0.8),breaks=seq(0.1,0.8,by=0.1))+
        geom_vline(xintercept = 0, size = 1) + stat_function(fun = dnor
m, size = 2)+
        xlab("x")+
        ylab("Denisty")+
        ggtitle("Supplementary Fig1. Another way to show the mean of 40
Exponentials simulated 1000 times")+
        theme(plot.title = element_text(size = rel(0.9),face="bold"))
```

```
# where x is the difference between a randam varible and mean
```



4. The information of the working platform and system can be found below.

```
sessionInfo()
```