Report for simulation exercise

NOTE: I understand that "*Each pdf report should be no more than 3 pages with 3 pages of supporting appendix material if needed (code, figures, etcetera).*". **However**, I find it is easier to follow the report if all the appendix material is integrated with the conclusions, since I'm "kniting" the R markdown file. Thanks for understanding, hope you enjoy :)

Simulation

I started by doing the required simulation. The instructions were has follow:

1. "The exponential distribution can be simulated in R with **rexp(n, lambda)** where lambda is the rate parameter"
2. "Set **lambda = 0.2** for all of the simulations"

As such, lambda was set to 0.2

```
lambda=0.2
```

3. "The mean of exponential distribution is **1/lambda** and the standard deviation is also 1/lambda

The value for the real mean was stored in the R variable real.mean for future reference

```
real.mean<-1/lambda
```

4. "In this simulation, you will investigate the distribution of means of **40 exponential(0.2)s**"

5. "Note that you will need to **do a thousand or so simulated means** of 40 exponentials"

An example on how to simulate according to these instructions was provided:
"As a motivating example, compare the distribution of 1000 random uniforms **hist(runif(1000))** and the distribution of 1000 means of 40 random uniforms **mns = NULL, for (i in 1 : 1000) mns = c(mns, mean(runif(40))), hist(mns)**"

Similarly, the distribution of 40 exponentials was taken

```
set.seed(1357) #for reproducibility
hist(rexp(40,0.2),xlab="Sample values",main="",breaks=15)
```

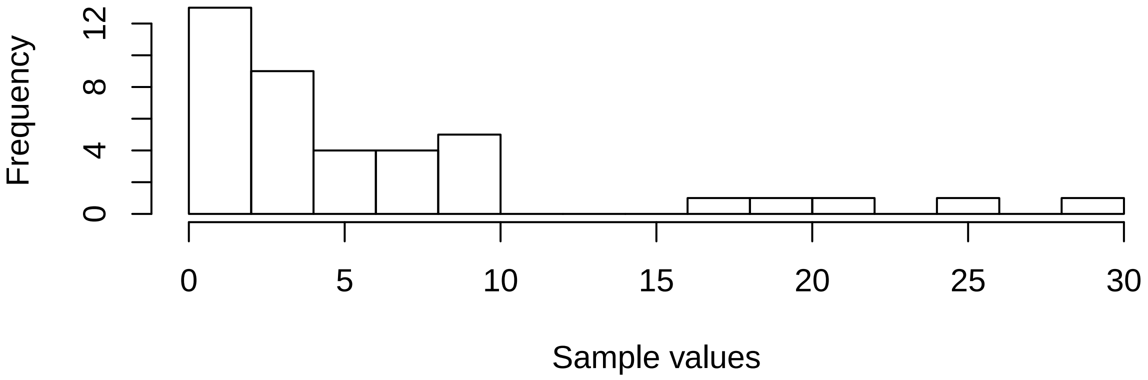

FIGURE 1: Example of the distribution for simulated 40 exponential

Following the representation of the 40 simulted exp(0.2)s, the mean of 40 of these distributions was estimated into the R variable mns

```r
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(40,0.2))) #simulate 40 exponentials 1000 times
myhist<-hist(mns,xlab="Sample means",main="") #keep the histogram information
```

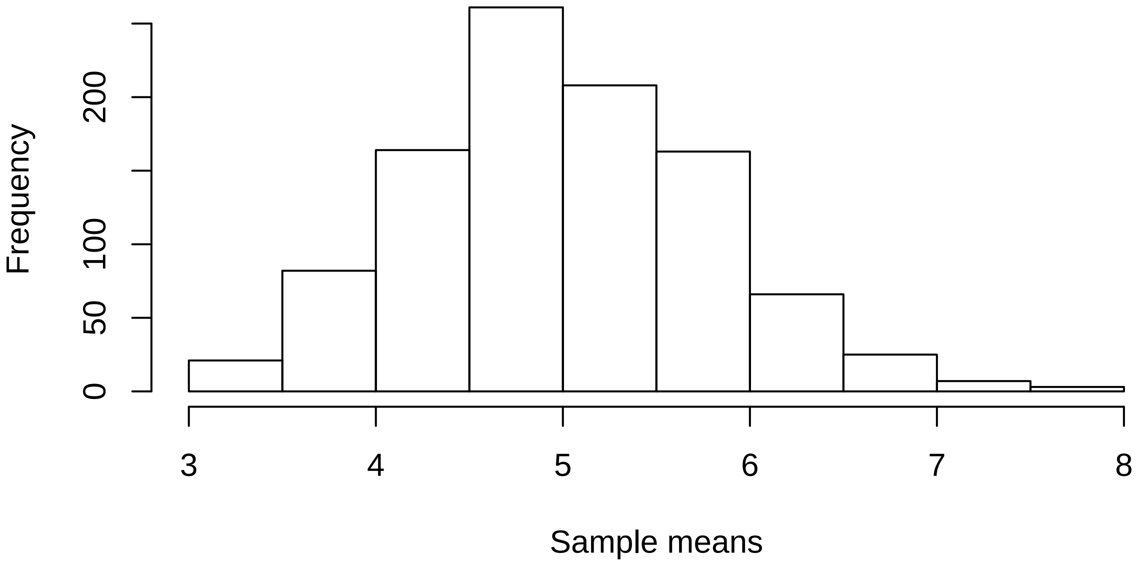

FIGURE 2: Distribution of the means 40 exponential observations simulated 1000 times

Aim:

To illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s.

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

According to the Central Limit Theorem, the distribution of the means of the simulated exponentials should approximate to a normal distribution, given a large enough number of samples were simulated. This can be confirmed by looking at the distribution of the means in figure 2. Consequently, the distribution of the means is expected to be centered in the population mean.

To find where the distribution is centered, the mean of the R variable mns was taken

```
mean<-mean(mns)
diff.mean<-abs(real.mean-mean) #compare with the real mean
```

The distribution of means obtained for the simulated exp(0.2) distributions is 5.0009 while the theorical center 5. This corresponds to a difference of 9.1831 × 10-4.

The comparison is more immediate when represented graphically

```
hist(mns,xlab="Sample means",main="")
abline(v=real.mean,col="green",lwd=5)
abline(v=mean,col="red",lwd=5,lty=3)
```

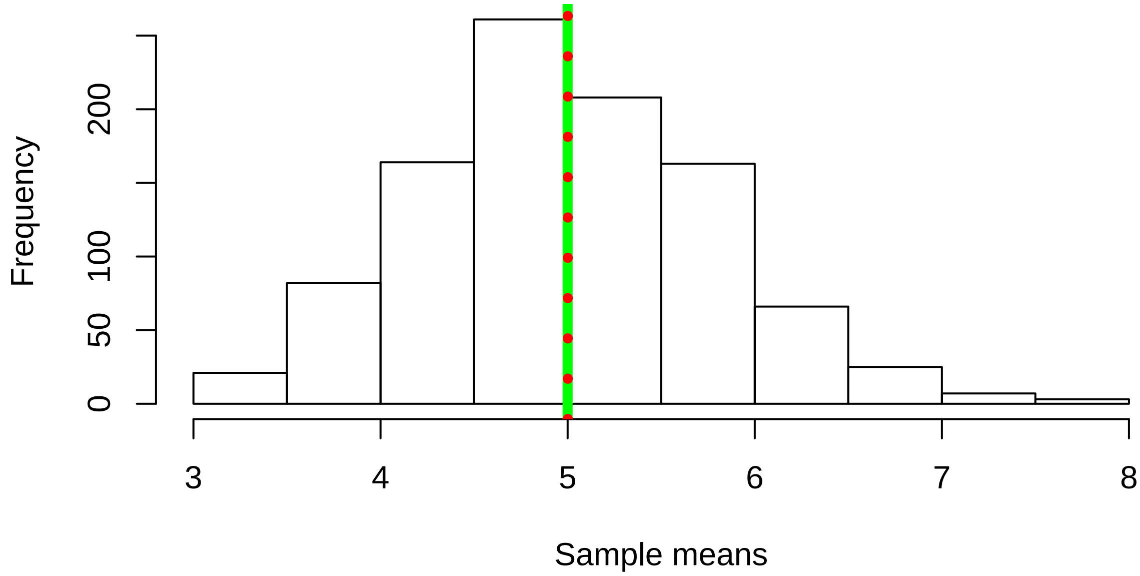

FIGURE 3: Graphical comparison between the estimated center of the simulated exp(0.2) *(dotted red)* and the theorical center of the exp(0.2) *(green line)*

As conclusion, since the number of sample simulations is large enough for the Central Limit Theorem to apply, the center of the means for the simulated exponential samples is very close to that of the population of exp(0.2)

2. Show how variable it is and compare it to the theoretical variance of the distribution.

The variance of the mean of a population is thus expected to be sigma squared divided by the sample number (40 exponentials), where sigma is the standard deviation of the exp(0.2) distributions. Since the standard deviation of an exponential distribution is the same as the mean, that is 1 divided by lambda, the expected standard error of the mean was calculated as

```
real.dev<-sqrt(real.mean^2/40)
```

While the observed standar error of the mean for the samples simulated would be calculated by

```
mean.dev<-sd(mns)
diff.dev<-abs(real.dev-mean.dev) # compare with real standard deviation
```

The standard error of the means obtained for the simulated exp(0.2) is 0.7891 while the theorical center for such distribution is 1/lambda, that is 0.7906. This corresponds to a difference of 0.0014.

The comparison is more immediate when represented graphically

```
hist(mns,xlab="Sample means",main="")
segments(real.mean,100,real.mean+real.dev,100,col="green",lwd=3)
segments(mean,105,mean+mean.dev,105,col="red",lwd=3)
```

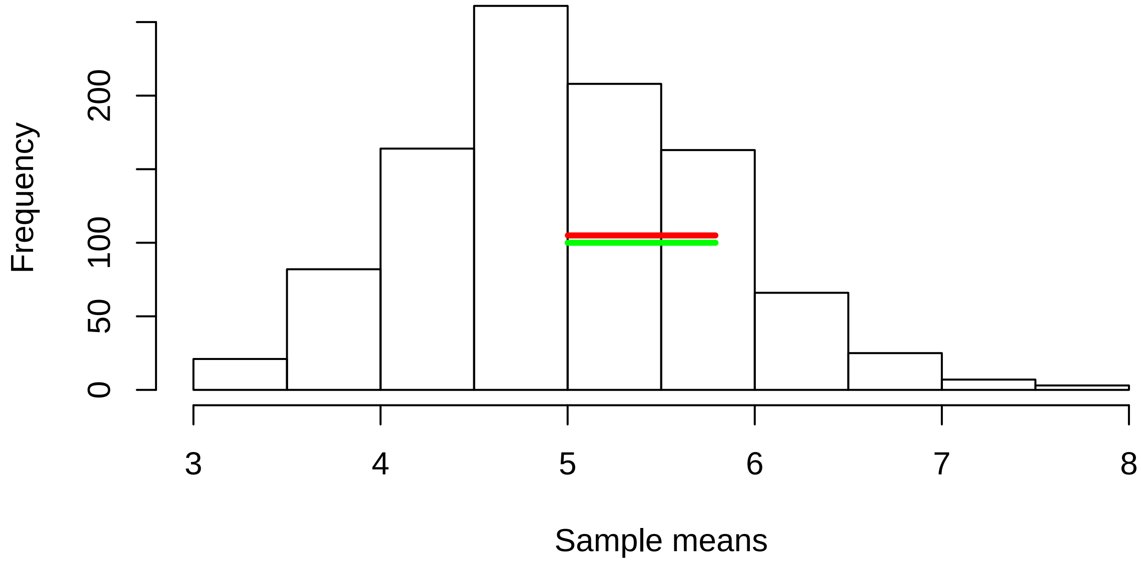

FIGURE 4: Graphical representation of 1 sd from the mean for the estimated standard deviations for the mean distribution of the simulated exp(0.2) *(red line)* and the theorical cneter of the exp(0.2) *(green line)*.

The theorical value for the standard error of the mean corresponds to the limit of the estimated standard error for the samples when the number of simulations tends to infinity. As such, with a large enough number of simulations, the standard error of the population mean can be estimated.

3. Show that the distribution is approximately normal.

With both the mean and the standard error of the mean estimated, I can draw a normal distribution that has these same parameters and compare the simulated distribution of the means with the obtained normal approximation.

```
library(ggplot2)
mns2<-as.data.frame(mns)
g <- ggplot(mns2, aes(x = mns))+ xlab("Sample means") +
    geom_histogram(alpha = .20, binwidth=.3, colour = "black",aes(y = ..density..))
g <- g + stat_function(fun = dnorm,color="red",arg = list(mean=mean(mns),sd=sd(mns)),lwd=1)
g
```

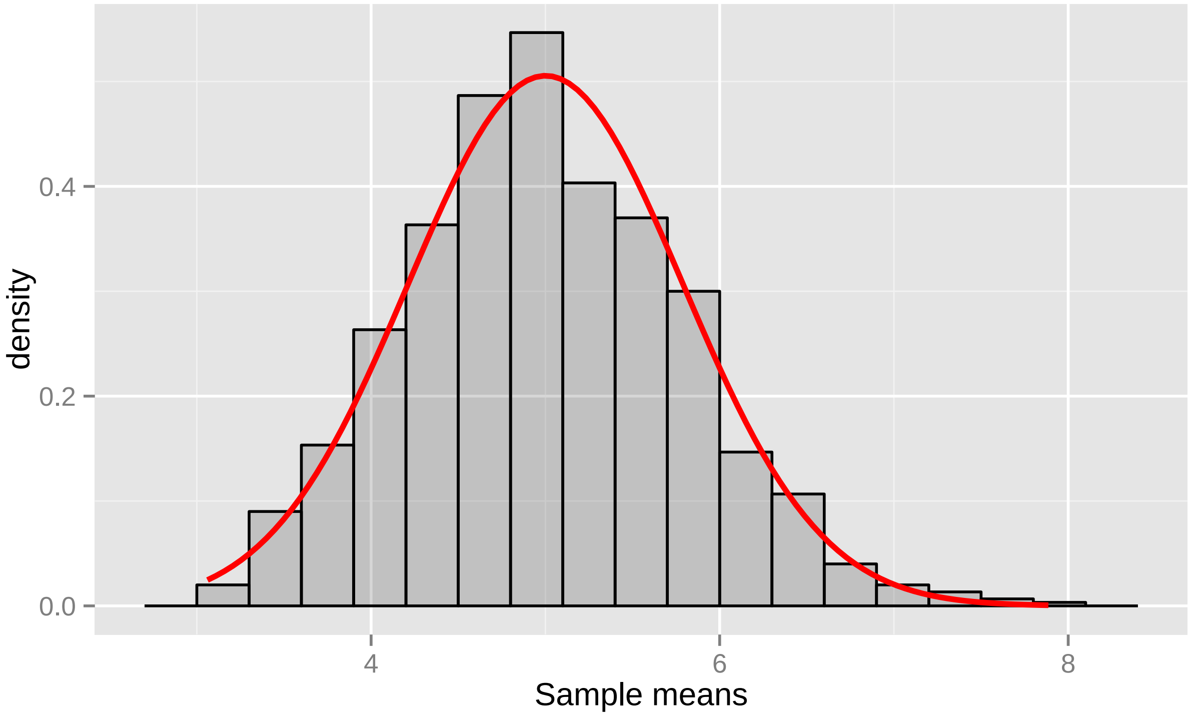

FIGURE 5: Graphical representation of the distribution of the means for the simulated exp(0.2) overlayed with the approximated normal approximation *(red line)*

The distribution of the means does seem to be well approximated by a normal distribution, confirming the validity of the Central Limit Theorem for this simulations. Additional confirmation was obtained by calculating the probabilities for various sample quantiles in both exponential and normal approximation

```
#Exponential probability values
#calculate frequencies
freq<-myhist$counts/sum(myhist$counts)
#Assign frequencies to quantiles
```

```r
hist.p<-cbind("breaks"=myhist$breaks[1:length(myhist$breaks)-1],"probability"=freq)


probs<-data.frame()
for (i in c(3,4,5,6,7)) {
    exp<-round(sum(hist.p[hist.p[,1]<i,2]),3)
    norm<-round(unname(pnorm(i,mean = mean(mns), sd = sd(mns))),3)
    probs<-rbind(probs,c(i,exp,norm,norm-exp))
}
names(probs)<-c("quantile","p.exponential","p.normal","difference")
probs
```

```
##   quantile p.exponential p.normal difference
## 1        3         0.000    0.006      0.006
## 2        4         0.103    0.102     -0.001
## 3        5         0.528    0.500     -0.028
## 4        6         0.899    0.897     -0.002
## 5        7         0.990    0.994      0.004
```

It can be seen that the probability for the quantiles is very similar between both distributions, with a maximum difference of 0.028, thus confirming numerically that the distributed means for the exponentials can be well approximated by a normal distribution.