Statistical Inference - Course project Part 1

Statistical Inference - Course project Part 1

(from assignment)

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. I will investigate the distribution of averages of 40 exponentials. I will do a thousand simulations.

Exploration

What's a rexp? - let's do 1000 values and plot them

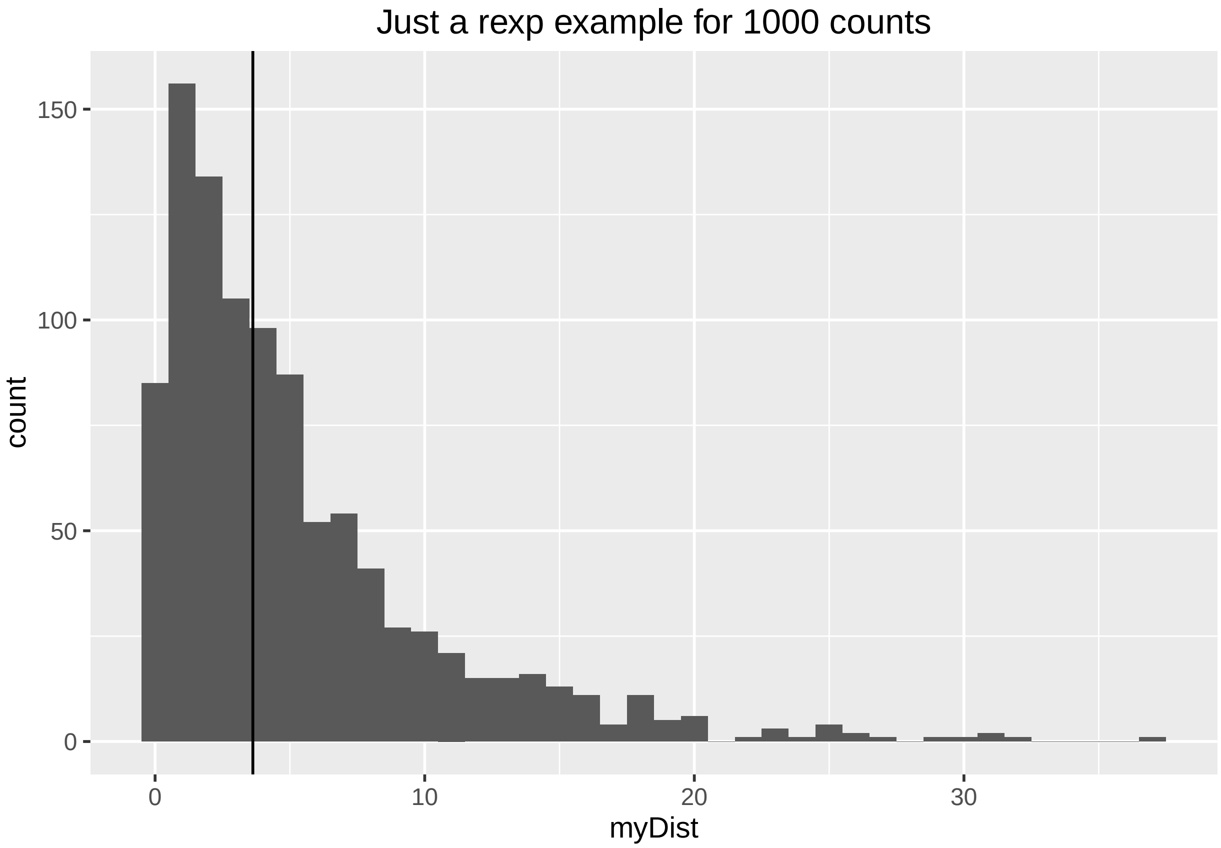
```
library(ggplot2)
set.seed(10101)

#Since this rexp thing is probably random, I will save it
myDist <- rexp(1000,.2)
head(myDist)
```

```
## [1] 9.5437361 0.7892039 1.7509186 3.7975125 0.4325101 1.1197484
```

```
summary(myDist)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00146  1.54900  3.62600  5.26300  7.15400 37.30000
```

```
#There were warnings asking to set binwidth, so I did it.
qplot(myDist, binwidth=1)+geom_vline(xintercept = median(myDist)) +
  ggtitle('Just a rexp example for 1000 counts')
```



Simulations

Now let's do 1000 of these with size of 40 per project requirement. I will save them to a matrix with 1000 rows and 40 columns

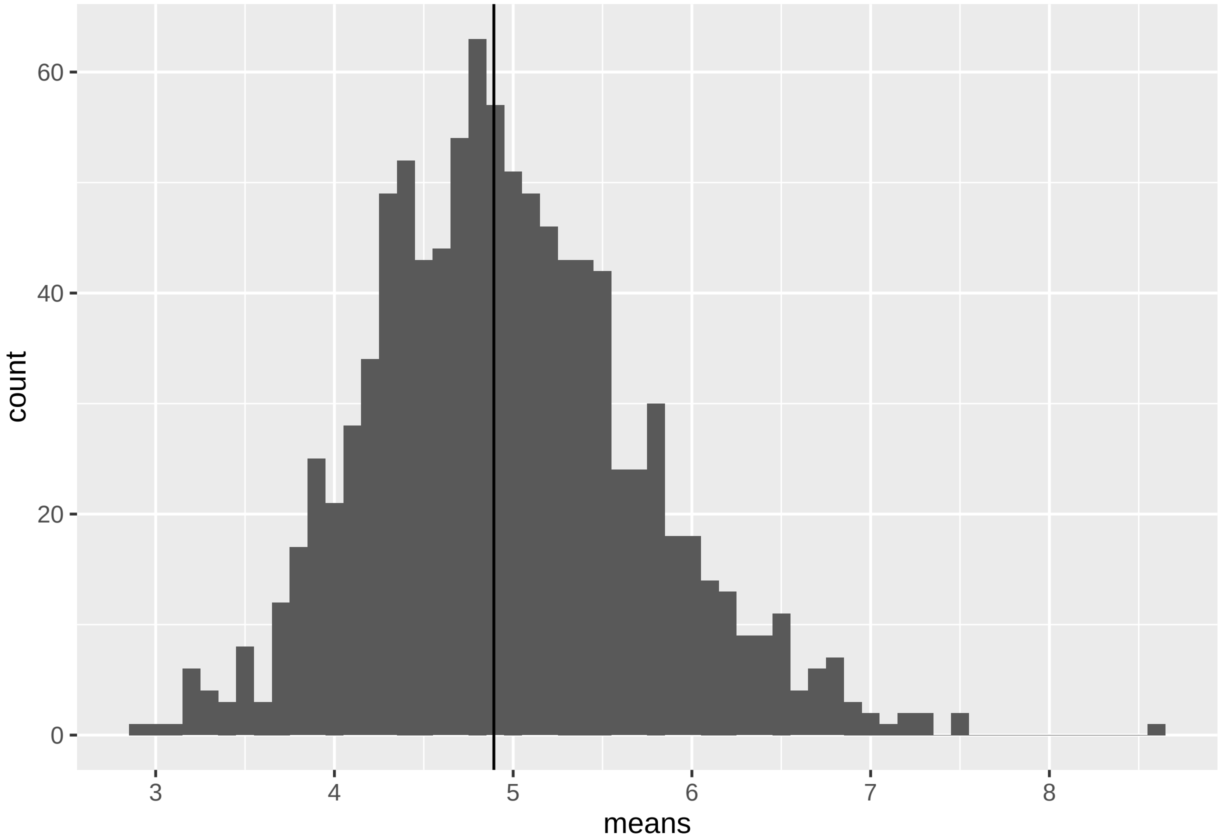
```
samples <-1000
n <- 40
lambda <- 0.2
dataSet <-matrix(data=rexp(n * samples, lambda), nrow=samples)
```

Let's take means of every row (observation)

```
means <- apply(dataSet, 1, mean)
summary(means)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.871   4.435   4.892   4.966   5.441   8.611
```

```
#Let's plot them too!
g<-qplot(means, binwidth=.1)+geom_vline(xintercept = median(means))
g
```



Mean of means turns out to be different every time I run knitr - I should set seed up top. With seed 10101 mean of means is constant - 4.966

Sample Mean versus Theoretical Mean

Condition of the project is that mean is 1/lambda

```
1/lambda
```

```
## [1] 5
```

Theoretical mean 5 and sampled mean 4.966 are pretty close!

Sample Variance versus Theoretical Variance

Let's find theoretical and practical standard deviation and variance

```
theo_sd <- (1/lambda)/sqrt(n)
theo_sd
```

```
## [1] 0.7905694
```

```
theo_var <- theo_sd^2
theo_var
```

```
## [1] 0.625
```

```
prac_sd <- sd(means)
prac_sd
```

```
## [1] 0.7681136
```

```
prac_var <- var(means)
prac_var
```

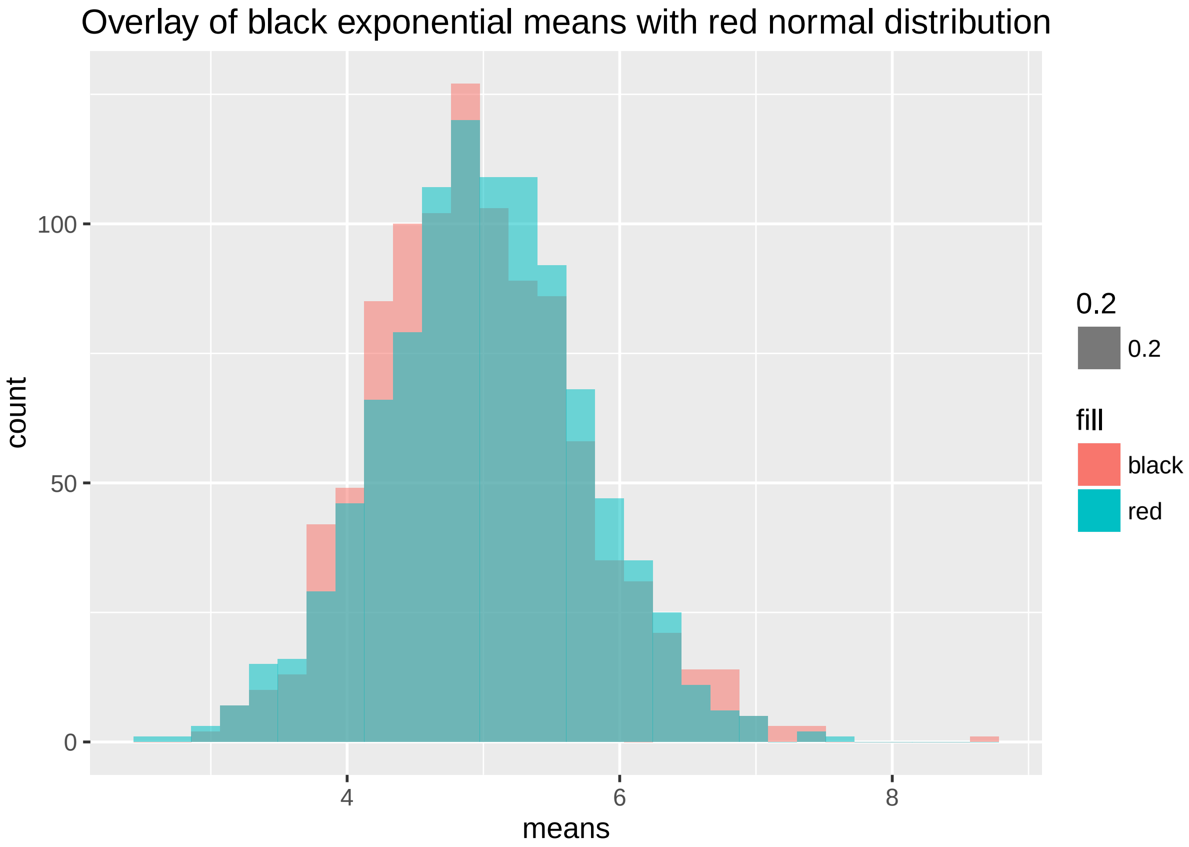
```
## [1] 0.5899986
```

Here we have sd off by couple of decimals(0.7681136 vs 0.7905694), and variance of course in the same area. I wish I had a few years of statistical background to conclude it it close enough, but I also have a gut feeling that in scope of this project the practice will fit the theory, so they are close.

Distribution

Here's a overlay with a random distribution with mean = 4.996 and sd = 0.7905694. It fits really well, so we conclude the distribution of means of rexp is approximately normal.

```
line <- rnorm(1000,5,0.7905694)
df <- data.frame(means=means,line=line)
ggplot(df, aes(means)) +
  geom_histogram(aes(means, bindwidth=.1, fill = "black",alpha = 0.2)) +
  geom_histogram(aes(line, bindwidth=.1, fill = "red", alpha = 0.2)) +
  ggtitle('Overlay of black exponential means with red normal distribution')
```


Overlay of black exponential means with red normal distribution
count
100
50
0
4
6
8
means
0.2
0.2
fill
black
red