Final Assignment for the Statistical Inference module: Part 1

In the first part of the assignment, we were tasked to investigate the distribution of a 1000 means of 40 variables from the exponential distribution. Then, we will compare it with the Central Limit theorem.

Firstly, we shall simulate the variables, while setting the "lambda" parameter equal to 0.2, as stated in the assignment. We shall store the 1000 exponentials in a variable "expn".

```
set.seed(234112)

expn<-rexp(1000, 0.2)

str(expn)
```

```
##  num [1:1000] 4.07 4.35 2.9 1.57 7.3 ...
```

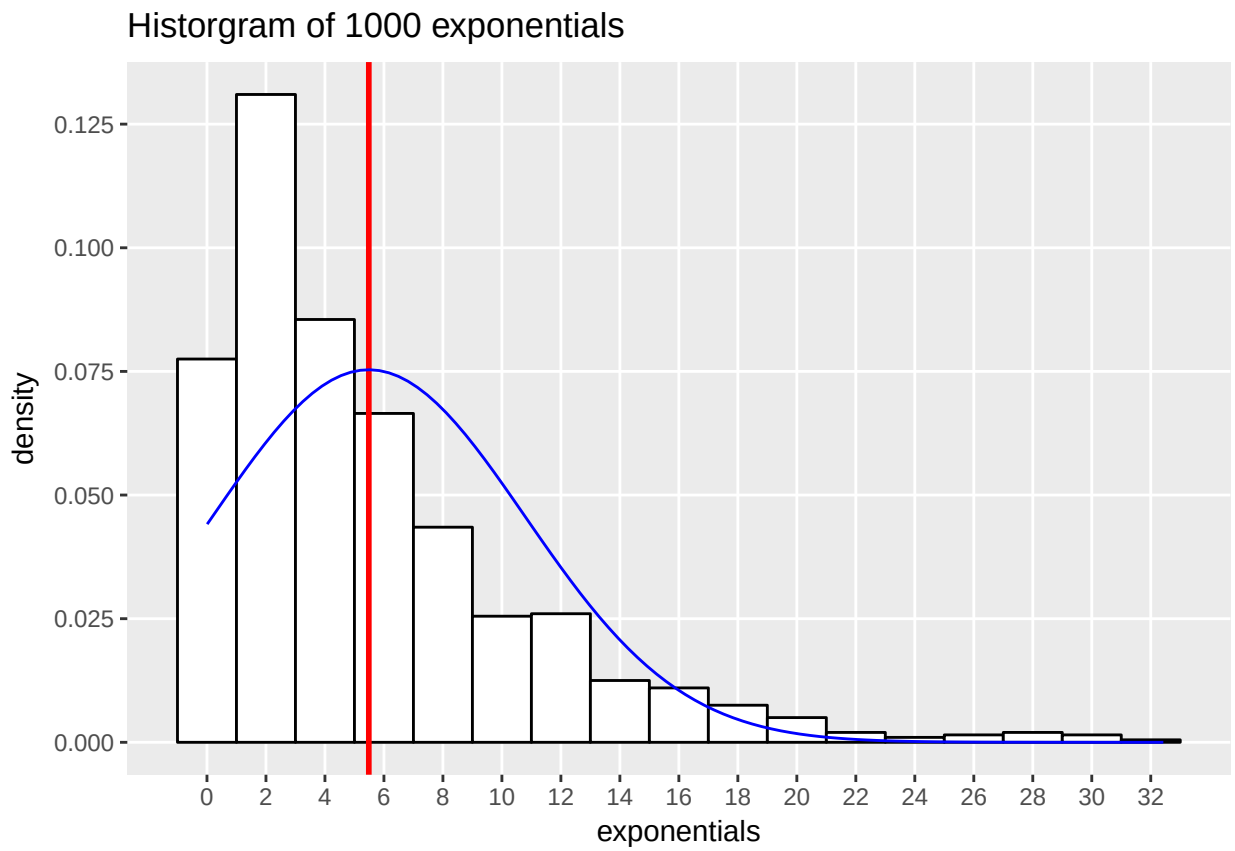
The theoretical mean of this sample is 1/lambda or 5, while the variance should be equal to (1/lambda)^2 or 25. Lets calculate the actual mean and variance of "expn".

```
u1<-mean(expn)

var1<-var(expn)

c(u1,var1)
```

```
## [1]  5.487744 28.041531
```

It seems that the values are quite close to to theoretical predictions. Lastly, lets take a look at the shape of our distribution by building a histogram.

```
plot1<-as.data.frame(expn)

library(ggplot2)

ggplot(data = plot1, aes(x=plot1))+

    geom_histogram(col= "black", fill="white",aes(y=..density..), binwidth = 2)+

    scale_x_continuous(name = "exponentials", breaks = seq(0, 32, 2), minor_breaks = NULL)+

    scale_y_continuous(breaks = seq(0, 0.15, 0.025), minor_breaks = FALSE)+

    geom_vline(xintercept = u1, col="red", lwd=1)+

    ggtitle("Historgram of 1000 exponentials")+

    stat_function(fun=dnorm, color="blue", args=list(mean=u1, sd=sqrt(var1)))
```



The red line on the histogram represents the mean value of "expn". The blue line is a normal distribution with the same mean and std as the sample. The distribution is clearly skewed to the right and is not very Gaussian in shape.

Next, we shall simulate 1000 sets of 40 exponential variables (with lambda=0.2) and take the mean of each set. We shall store the 1000 means in "expn.means".

```
expn.means<- c(1:1000)

for (i in 1:1000) {x<-mean(rexp(40, 0.2))

    expn.means[i]<-x}

str(expn.means)
```

```
##  num [1:1000] 5.91 6.13 4.29 5.36 4.44 ...
```

Likewise, we will now compare the actual mean and variance of "expn.means" with their theoretical values. Since sample mean in an unbiased estimator of the population mean, the distribution of the sample mean will be Gaussian, even if the original population distribution it not.

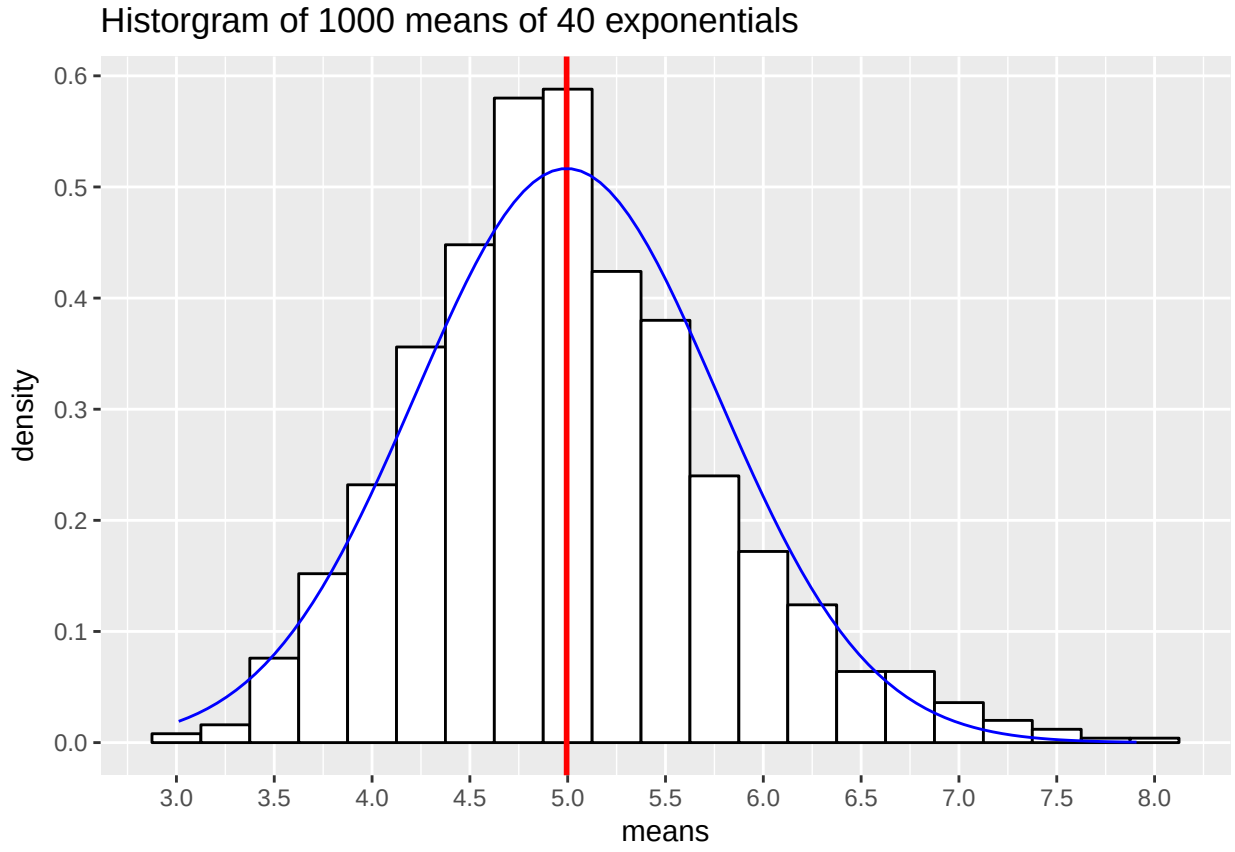
Hence, we expect the mean of the sample mean distribution to be 1/lambda=1/0.2=5 and the variance to equal to lambda^2/n=25/40=0.625.

```
u2<-mean(expn.means)

var2<-var(expn.means)

c(u2, var2)
```

```
## [1] 4.9945111 0.5963646
```

As we can see, the actual values are very close to their theoretical values. As a last step, lets build a histogram of this distribution and see if it is Gaussian in shape.

```
plot2<-as.data.frame(expn.means)

ggplot(data = plot2, aes(x=plot2))+

    geom_histogram(col= "black", fill="white", aes(y=..density..), binwidth = 0.25)+

    scale_x_continuous(name = "means" , breaks = seq(1, 10, 0.5)) +

    scale_y_continuous(breaks = seq(0, 0.6, 0.1), minor_breaks = FALSE)+

    geom_vline(xintercept = u2, col="red", lwd=1)+

    ggtitle("Historgram of 1000 means of 40 exponentials")+

    stat_function(fun=dnorm, color="blue", args=list(mean=u2, sd=sqrt(var2)))
```



Once more, the red line is the mean and blue line is a standard normal distribution with same mean and variance as "expn.means". This time, in accordance with the Central Limit Theorem, the distribution is a lot more Gaussian in shape, centered around the population mean.