Part 1. A simulation exercise.

RE

Statistical Inference. Course Project.

Part 1. A simulation exercise.

Problems.

In this project you 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. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

You should

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Solution

1. Show the sample mean and compare it to the theoretical mean of the distribution.

Lets calculate the theoretical mean and standard deviation for the exponential distribution .

```
nosim <- 1000; n <- 40; lambda <- 0.2;
mean_theoretical <- 1/lambda
sd_theoretical  <- 1/lambda /sqrt(n)
```

Lets get the simulation data into a 1000*40 matrix, and a vector of 1000 means of the every 40 samples.

```
simulation_data <- matrix(rexp(nosim * n, lambda), nosim)
x_bars <- apply(simulation_data, 1, mean)
c('empirical and theoretical mean:')

## [1] "empirical and theoretical mean:"
```

```
c(mean(x_bars),mean_theoretical)
```

```
## [1] 4.979678 5.000000
```

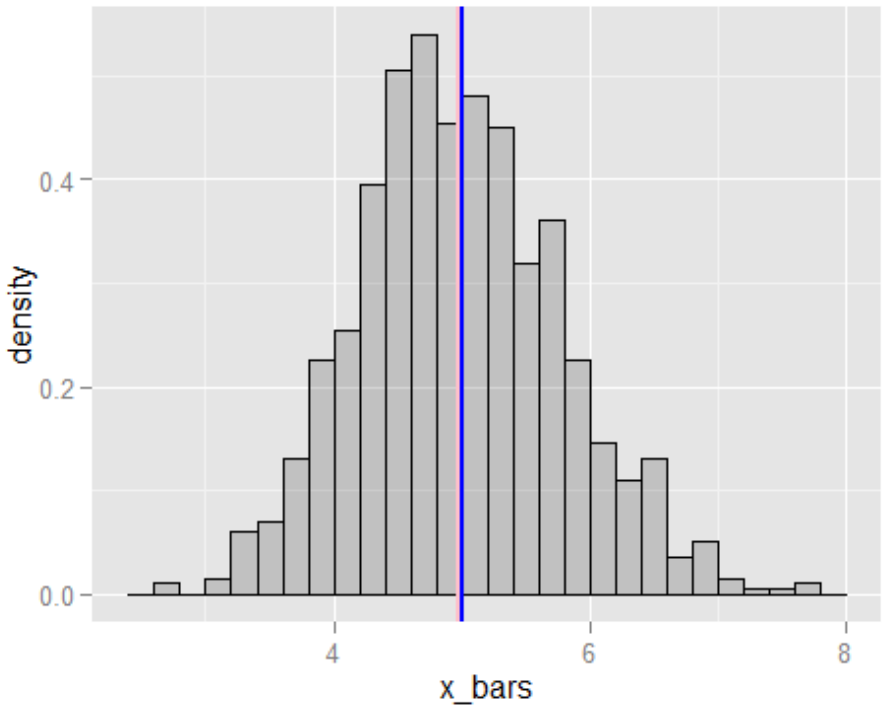
The mean of sample means is near to the theoretical population mean.

Because of the LLN says that the sample mean of iid sample is consistent for the population mean.

The sample variance and the sample standard deviation of iid random variables are consistent as well.

Let's have a look of the following plot of the means. The average of the means line (the pink one) is near to the theoretical mean (the blue line) 'mean_theoretical':

```
library(ggplot2)
g = qplot(x_bars, geom = 'blank') +
  geom_histogram(aes(y = ..density..), alpha = 0.2, binwidth =
.20,colour="black")+
  geom_vline(xintercept = mean(x_bars), colour = 'pink',size=1)+
   geom_vline(xintercept = mean_theoretical, colour = 'blue',size=1)
print(g)
```



2. Show how variable the sample mean is (via variance) and compare it to the theoretical variance of the distribution.

Lets calculate the variance of the 1000 means and compare to theoretical varance.

```
c('Empirical variance of the simples means: ', var(x_bars))
```

```
## [1] "Empirical variance of the simples means: "
## [2] "0.624212230534057"
```

```
c('theoretical variance of simples means: ', (1/lambda)^2/n)
```

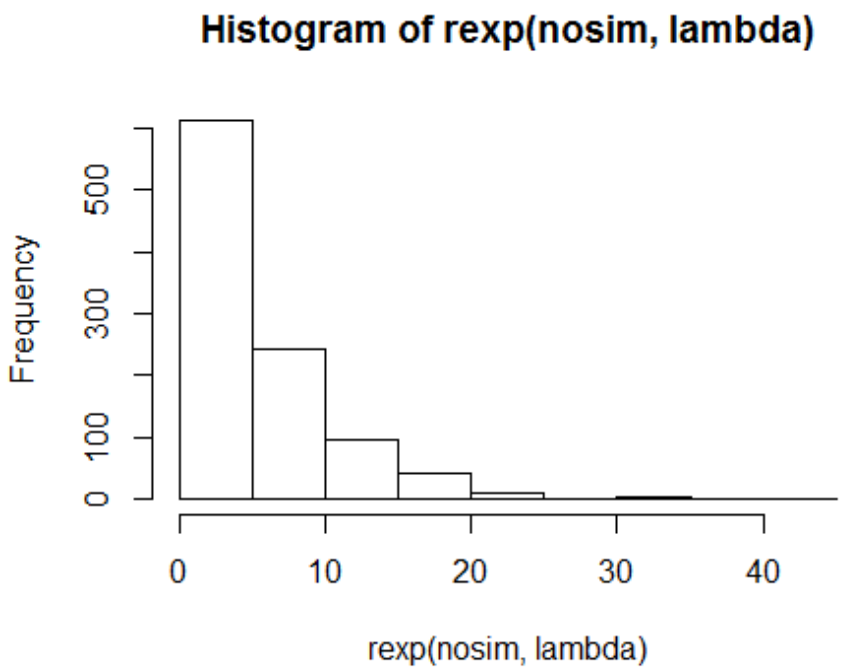
```
## [1] "theoretical variance of simples means: "
## [2] "0.625"
```

Thus the variance of the 1000 means is similar to their theoretical variance sd^2/n.

3. Show that the distribution is approximately normal.

Lets see what the exponential distribution looks like:

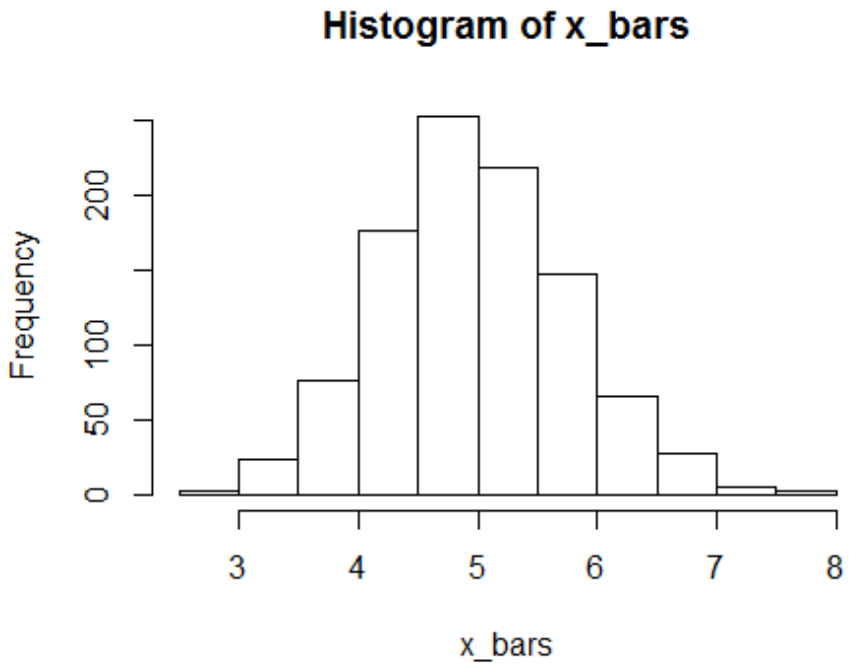
```
hist(rexp(nosim, lambda))
```



Now lets see at the distribution of the 1000 means of the samples sized 40.

The CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases

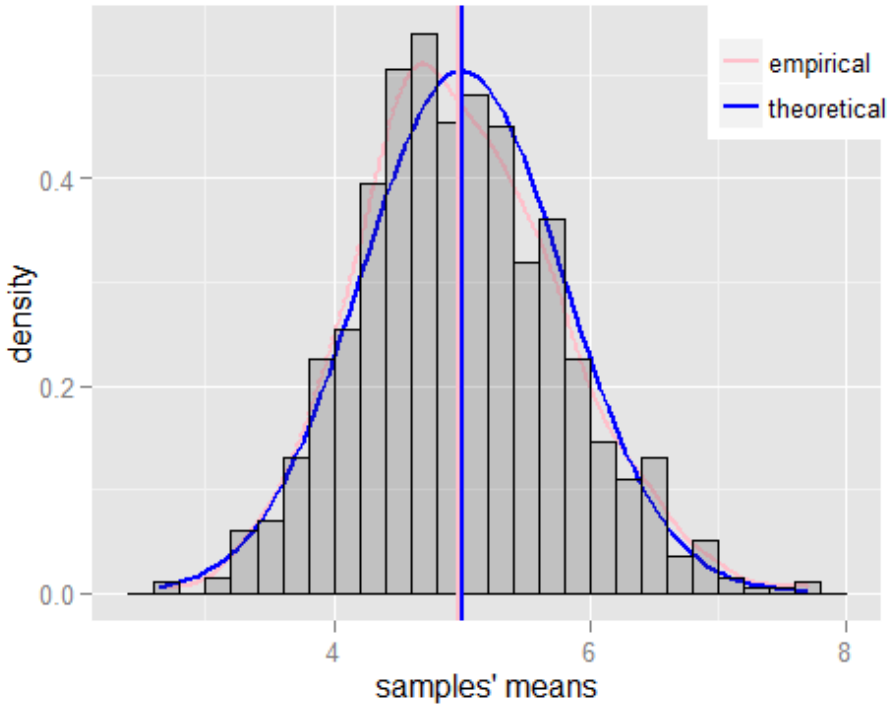
```
hist(x_bars)
```



The graph comforms to a normal distribution.

Lets print

```
g = qplot(x_bars, geom = 'blank') +
  geom_line(aes(y = ..density.., colour = 'empirical'), stat = 'density',size=1)
+
  stat_function(fun=dnorm, args=list(mean=mean_theoretical, sd=sd_theoretical),
                aes(colour = 'theoretical'), size=1) +
  geom_histogram(aes(y = ..density..), alpha = 0.2, binwidth =
.20,colour="black")+
  geom_vline(xintercept = mean(x_bars), colour = 'pink',size=1) +
  geom_vline(xintercept = mean_theoretical, colour = 'blue',size=1)+
  scale_colour_manual(name='', values = c('pink', 'blue')) +
  theme(legend.position = c(0.9, 0.9)) +  xlab("samples' means")
print(g)
```



We can see from plot above the means' density line is just around the normal distribution line of N(1/lambda, (1/lambda)/sqrt(n))