A simulation exercise, Statistical Inference Course Project Week 4

Jose Manuel Coello

NOTE: My native language is not English, so I apologize for any grammatical errors.

Simulation Exercise

In this first part of project we start by showing how the Central Limit Theorem (CLT) works. For it we take "n" samples of size 40 from the exponential distribution and compute the mean of each sample, thus creating the sampling distribution, this sampling distribution will be nearly normal with parameters mean = population mean $\mu = \frac{1}{lambda}$ and a standard deviation equal to the population standard deviation divided by the square root of the sample size = $\frac{\frac{1}{lambda}}{\sqrt{40}}$.

Before we do that, we need to check the conditions for the CLT.

1. Independence: Sampled observations must be independent. We have random samples, so this condition is covered.
2. Sample size/skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large, we will take samples of size 40, so this condition is also covered.

Simulations

We proceed to create the sampling distribution, we take 1000 samples of size 40 from the exponential distribution and compute the mean of each sample. The mean of the exponential distribution is $\frac{1}{lambda}$ and the standard deviation is also $\frac{1}{lambda}$. We set lambda = 0.2 for all simulations.

```
my_sampling_dis <- c()
lambda <- 0.2
n <- 40
nsim <- 1000

# I set the seed so that the result is reproducible
set.seed(12345)
for(i in seq_len(nsim)){
  my_sampling_dis[i] <- mean(rexp(n = n, rate = lambda))
  }
```

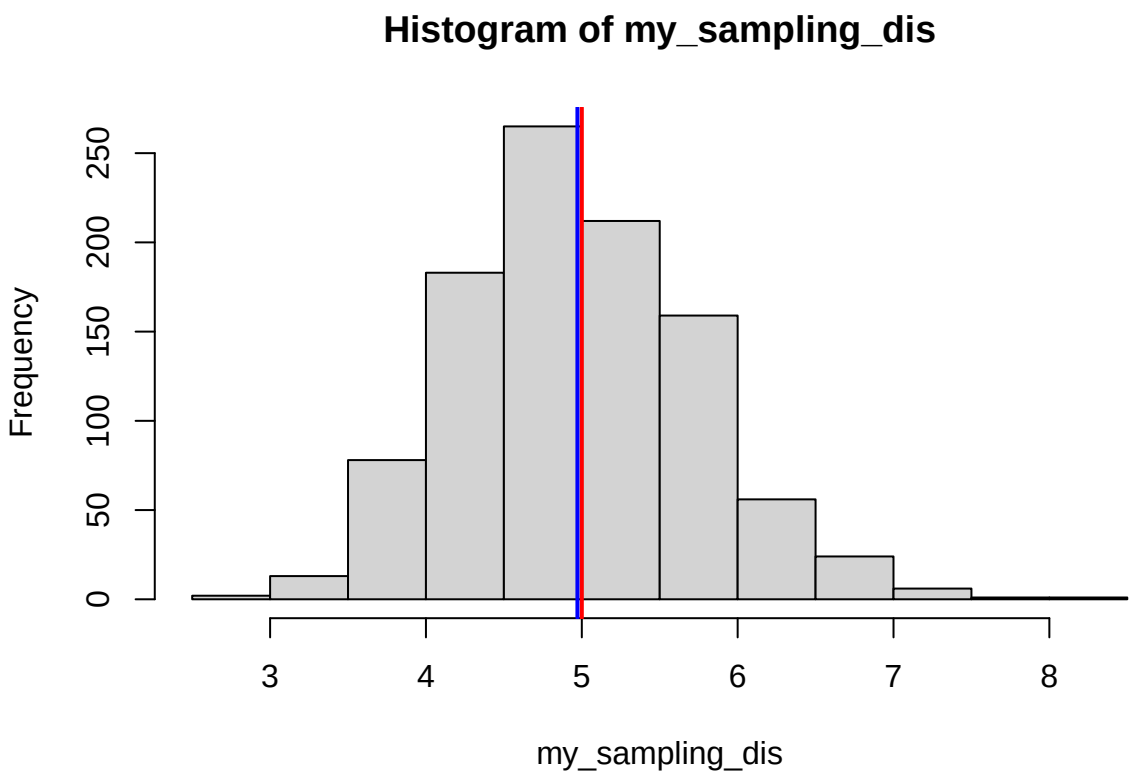
Sample Mean versus Theoretical Mean

Now we compare the mean of the sampling distribution with the theoretical mean of the exponential distribution. For it we make a plot of the sampling distribution and observe its mean and the population mean, these values by the CLT must be very similar.

```
avg_sampling_dis <- mean(my_sampling_dis)
avg_exponential_dis <- 1/lambda
c('mean_sampling_distribution' = avg_sampling_dis,
  'mean_exponential_distribution' = avg_exponential_dis)
```

```
##    mean_sampling_distribution mean_exponential_distribution
##                      4.971972                      5.000000
```

```r
hist(my_sampling_dis)
abline(v = avg_exponential_dis, col = 'red', lwd = 2)
abline(v = avg_sampling_dis, col = 'blue', lwd = 2)
```



We can see that the sampling distribution is nearly normal centered in the population mean, in the histogram, the red line is the population mean and the blue line is the mean of the sampling distribution, both values are very similar.

Sample Variance versus Theoretical Variance

We have by the CLT that the standard deviation of the sampling distribution is the standard deviation of the theoretical distribution divided by the square root of the sample size, so the standard deviation of the sampling distribution must be a value very close to $\frac{\frac{1}{0.2}}{\sqrt{40}} = \frac{5}{\sqrt{40}} = 0.79$. This value is called the standard error

```r
# standard deviation of the sampling dis
se_sampling_dis <- round(sd(my_sampling_dis), 2)
# standard error
standard_error <- round((1/lambda)/sqrt(n), 2)

se_sampling_dis
```

```
## [1] 0.77
```

```r
# variance of the theoretical distribution
var_exponential_dis <- round((1/lambda)**2, 2)
# variance of the sampling distribution
var_samplig_dis <- round(var(my_sampling_dis), 2)
```

We can see that the standard deviation of the sampling distribution is: 0.77 this value is very close to the

standard error 0.79.

With regard to the variance, the variance of the sampling distribution it's always less than the variance of the theoretical distribution, mathematically it's very easy to see, because the variance of the sampling distribution is $\sigma^2 = \left(\frac{\frac{1}{lambda}}{\sqrt{n}}\right)^2 = \left(\frac{5}{\sqrt{40}}\right)^2 = \frac{5^2}{40}$ while the variance of the theoretical distribution is $\sigma^2 = \left(\frac{1}{lambda}\right)^2 = \left(\frac{1}{0.2}\right)^2 = 5^2$.

We can see that $0.6 < 25$ this is because the variance of the sampling distribution is always divided by the sample size.

Distribution

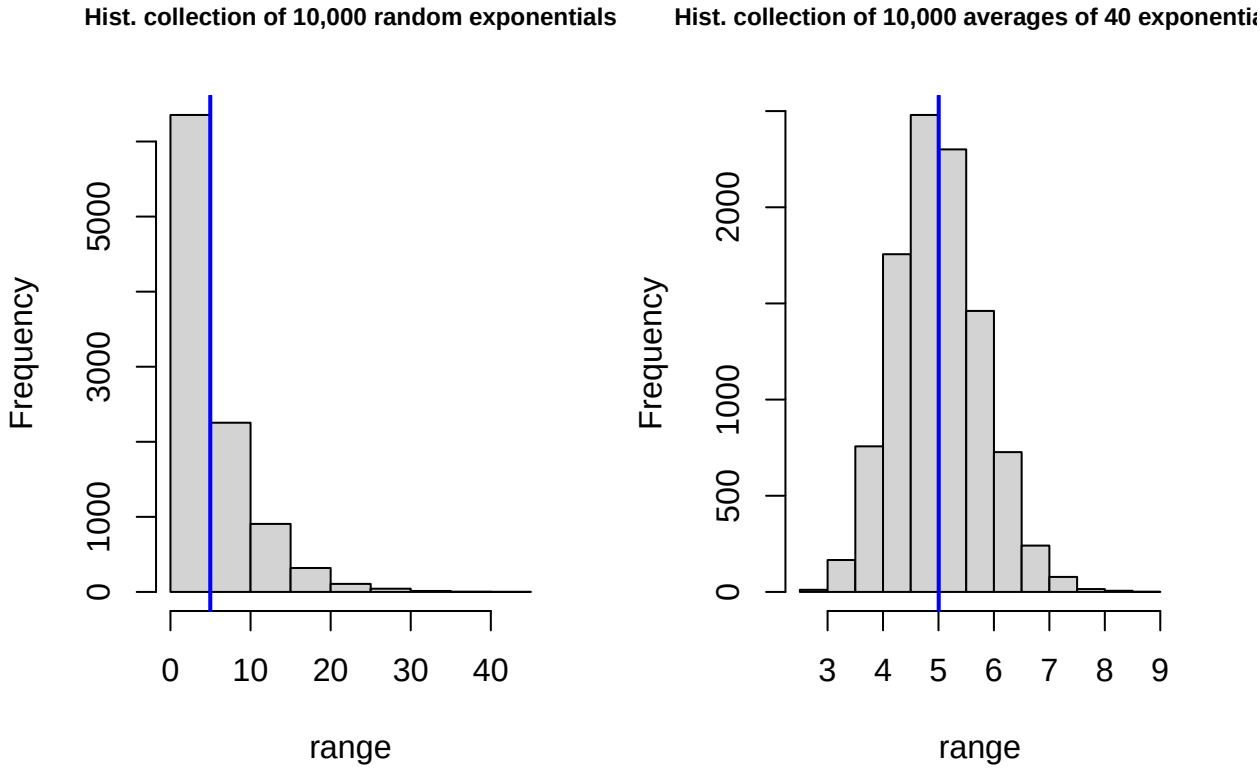
We take a large collection of random exponentials (10,000 records) and then proceed to create a new sampling distribution, but now we take a big number of samples than before, we take 10,000 samples of size 40 from the exponential distribution and compute the mean of each sample. Then we compare both of them. We keep lambda in 0.2 for all simulations.

```
# part three: show that the distribution is approximately normal.
nsim2 <- 10000
expo_sample <- rexp(n = nsim2, rate = lambda)

my_sampling_dis2 <- c()
sapply(X = seq_len(nsim2),
       function(x){
         my_sampling_dis2[x] <<- mean(rexp(n = n, rate = lambda))
         })
```

```
my_samples <- list(expo_sample, my_sampling_dis2)
my_titles <- c('Hist. collection of 10,000 random exponentials',
               'Hist. collection of 10,000 averages of 40 exponentials')

par(mfrow = c(1,2))
mapply(function(x, y){
  hist(x, main = y, cex.main = .75, xlab = 'range')
  abline(v = 1/lambda, col = 'red', lwd = 1)
  abline(v = mean(x), col = 'blue', lwd = 2)},
  my_samples, my_titles)
```




```
## [[1]]
## NULL
##
## [[2]]
## NULL
```

In the left side we have a plot showing a sample of 10,000 random exponentials, two lines were drawn, a red line with the population mean and a blue line with the sample mean, both lines overlap each other, showing that when the sample size increases, the sample mean converges to the population mean.

On the right hand side we have a plot showing a sampling distribution of 10,000 sample averages of size 40 from the exponential distribution, we can see that the sampling distribution is nearly normal, in both plots we only can see the blue line, it's centered in the population mean $\frac{1}{lambda} = \frac{1}{0.2} = 5$. This is a visual demonstration of the CLT it says that the mean of "n" samples of size "m" from a distribution is centered on the population mean μ with standard deviation $S = \frac{\sigma}{\sqrt{n}}$.