Johns Hopkins Coursera - Statistical Inference - Project Part 1

shostiou

05/09/2020

Part 01 - Simulation Exercise

Project objectives :
The purpose of this project is to 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.
We will investigate the distribution of averages of 40 exponentials.

Simulation

Let's start by defining the distribution which will be used for simulation purposes

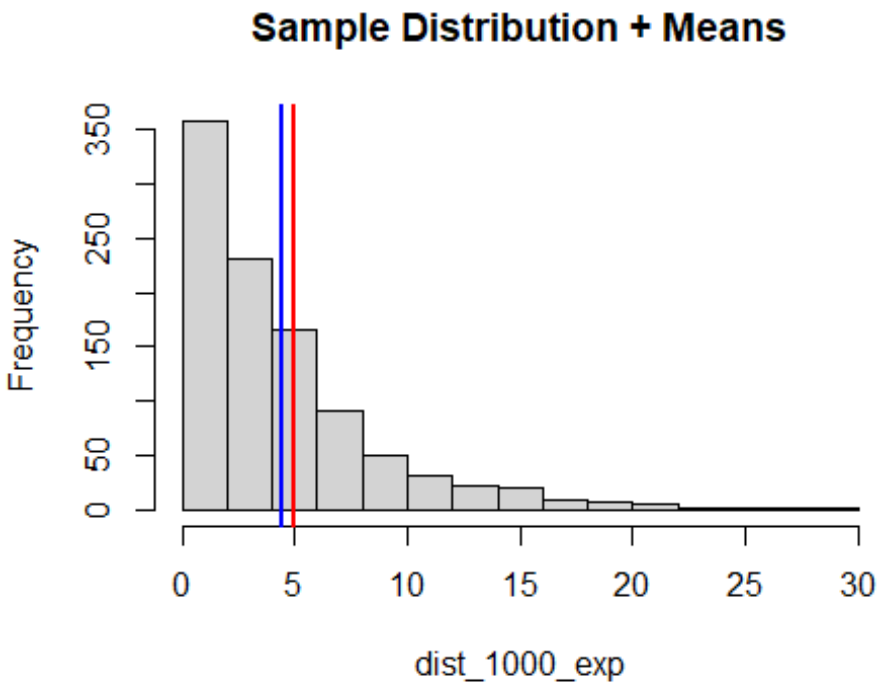
```
# Setting the value of lambda to 0.2 as requested
lambda <- 0.2
# Creating a distribution of 1000 exponentials
dist_1000_exp <- rexp(1000, lambda)
# Creating a distribution of 1000 averages of 40 random exponentials
dist_avg = NULL
for (i in 1 : 1000) dist_avg = c(dist_avg, mean(rexp(40, lambda)))
```

Analysing the means

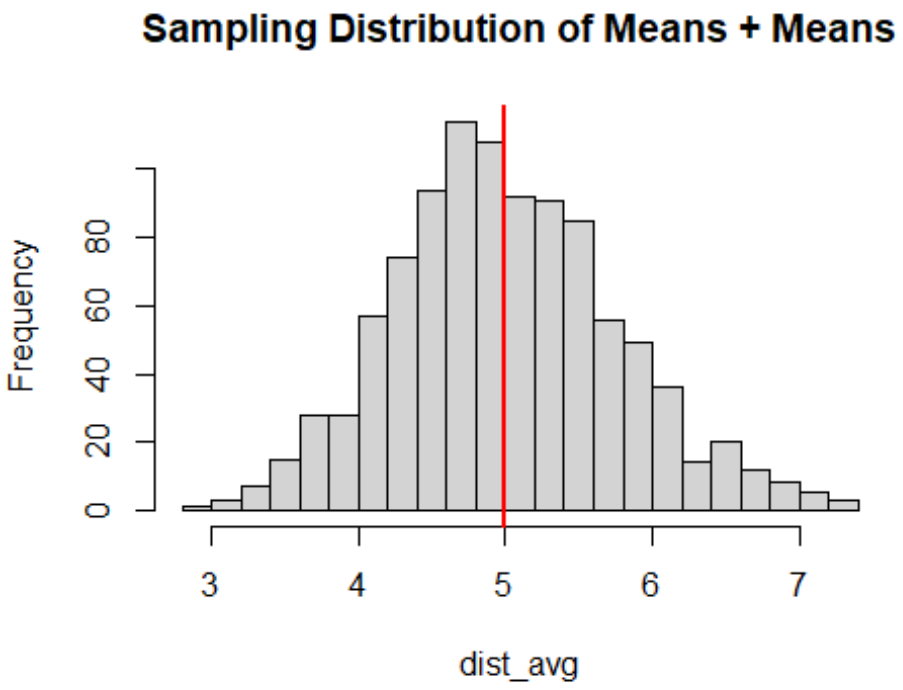
In this section we will build investigations around the means.
First recall that for an exponential distribution, we have mean = 1/lambda and std_dev = 1/lambda.
Those "true" values will be compared to the the sample mean and sampling distribution of the means.

```
# Let's calculate the true value of the exponential distribution mean
theo_mean <- 1/ lambda
paste0('True dist mean = ', theo_mean)

## [1] "True dist mean = 5"

# Comparing true mean to sample mean
sample_mean <- mean(dist_1000_exp)
paste('True mean = ', theo_mean, " ; sample mean = ", sample_mean, "
Diff :",theo_mean-sample_mean)
```

```
## [1] "True mean =  5  ; sample mean =  4.44410415112898  Diff :
0.555895848871018"

# Applying Central Limit theorem
# Calculating the sampling distribution mean
sampling_mean = mean(dist_avg)
paste0('Sampling Distribution of the means mean = ', sampling_mean)

## [1] "Sampling Distribution of the means mean = 4.99915452845423"

# Let's visualy explore those calculations
hist(dist_1000_exp, breaks=20, main = "Sample Distribution + Means")
abline(v = sample_mean, col = "blue", lwd = 2)
abline(v = theo_mean, col = "red", lwd = 2)
```




```
hist(dist_avg, breaks=20, main = "Sampling Distribution of Means + Means")
abline(v = sampling_mean, col = "blue", lwd = 2)
abline(v = theo_mean, col = "red", lwd = 2)
```



As stated in the course, we can see that the sample mean, the true mean (theoritical) and the sampling distribution of the means mean are similar (equal to 5).
The sample distribution is right skewed (this shape corresponds to an exponential distribution).
As expected by the Central Limit Theorem, **the sampling distribution of the means is centered on the sample mean (which corresponds also to the true population mean)**
Referring to the illustrations : Blue line is the sample / sampling distribution mean while red is the true distribution mean.

Analysing Variance

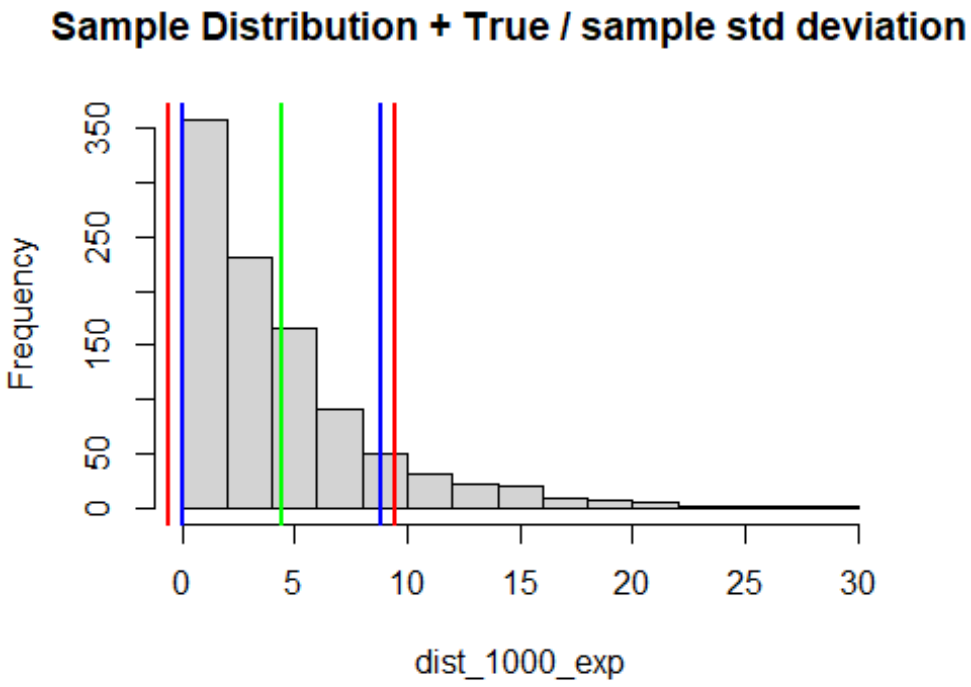
In this section, we will compare the variances of the distributions (true / sample / sampling).

```
# Let's calculate the true value of the exponential distribution variance
theo_var <- (1/ lambda)^2
# Comparing true variance to sample distribution variance
sample_var <- var(dist_1000_exp)
paste('True variance = ', theo_var, " ; sample var = ", sample_var, "
Diff :",theo_var-sample_var)
```

```
## [1] "True variance =  25  ; sample var =  19.470231682782  Diff :
5.52976831721802"
```

```
# Let's visualy explore those calculations
hist(dist_1000_exp, breaks=20, main = "Sample Distribution + True / sample
std deviations")
```

```
abline(v = sample_mean+sqrt(sample_var), col = "blue", lwd = 2)
abline(v = sample_mean-sqrt(sample_var), col = "blue", lwd = 2)
abline(v = sample_mean+sqrt(theo_var), col = "red", lwd = 2)
abline(v = sample_mean-sqrt(theo_var), col = "red", lwd = 2)
abline(v = sample_mean, col = "green", lwd = 2)
```




```
# Let's now observe the variance of the sampling distribution of means
sampling_var <- var(dist_avg)
paste0('Sampling Distribution of the means variance = ', sampling_var)

## [1] "Sampling Distribution of the means variance = 0.575230061544135"
```

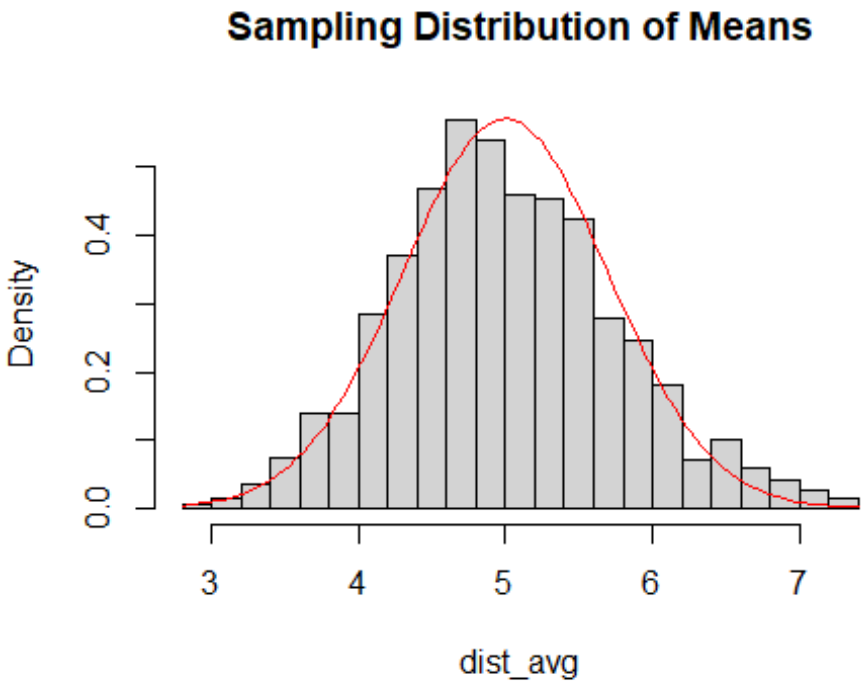
We can see that the sample and theoritical variances are quite similar. This was verified numerically and with a histogram were values of mean +/- 1 x std_deviation are displayed (blue : sample standard deviation / red : true std deviation).
As a reminded : std_deviation = sqrt(variance)
We can also see that the variance of the sampling distribution of the means as nothing to do with the sample / population variance. This is perfectly normal because this variance measures the variability of the approximation of the population (true) mean.

Sampling Distribution of Means - Normal approximation

The application of the Central Limit Theorem shows that the sampling distribution of the means can be approximated by a normal distribution :
- Mean = sample mean
- Standard Error = sample_std_dev / sqrt(means sample size)

Note : The normality conditions can be satisfied in this specific case because each observation has been randomly and independently generated.

```
#Let's compute the standard Error
sampling_SE <- sd(dist_1000_exp)/sqrt(40)
# Compating distribution + normal approximation
hist(dist_avg, breaks=20, main = "Sampling Distribution of Means", prob=TRUE)
curve(dnorm(x, mean=sampling_mean, sd=sampling_SE), add=TRUE, col="red")
```



We can see graphicaly that the sampling distribution of means can be approximated by a normal distribution (according to Central Limit Theorem)