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"

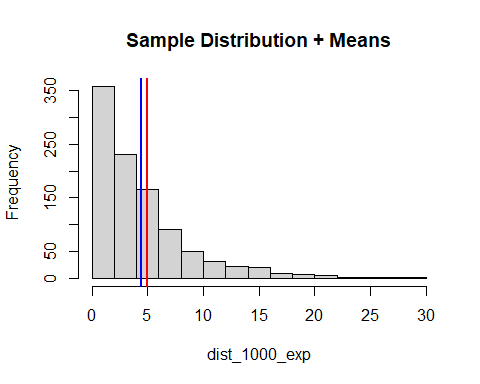
# Comparing true mean to sample mean   
sample\_mean <- mean(dist\_1000\_exp)  
paste('True mean = ', theo\_mean, " ; sample mean = ", sample\_mean, " Diff :",theo\_mean-sample\_mean)

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

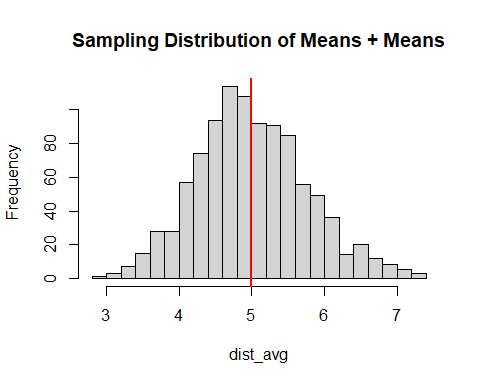
# Applying Central Limit theorem   
# Calculating the sampling distribution mean  
sampling\_mean = mean(dist\_avg)   
paste0('Sampling Distribution of the means mean = ', sampling\_mean)

## [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.

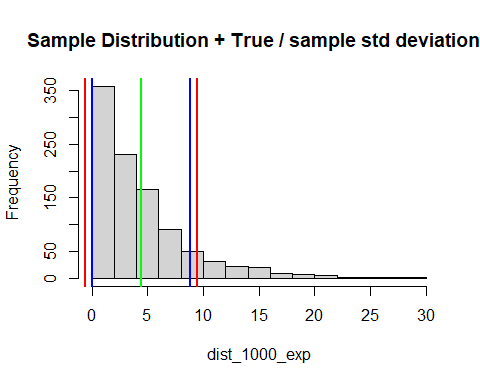
### Analysing Variance

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)

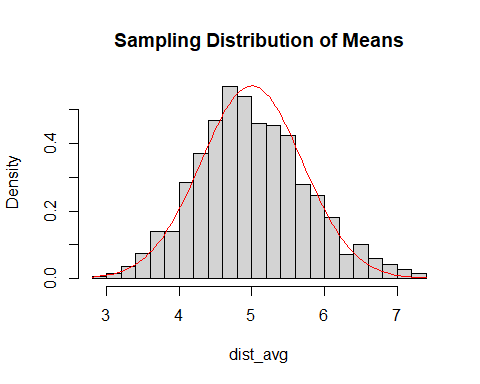
## [1] "Sampling Distribution of the means variance = 0.575230061544135"

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)