Statistical Inference Part 1

Kris Jacomet

8/14/2020

```
rm(list=ls(all=TRUE))
library(ggplot2)
library(knitr)
```

Overview

In this project I 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. Setting lambda = 0.2 for all of the simulations, I will investigate the distribution of averages of 40 exponentials. Note that I will need to do a thousand simulations.

The results will Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. I shall: 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.

In point 3, I focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Simulation

I will run a series of 1000 simulations to create a data set for comparison to theory. Each simulation will contain 40 observations and the expoential distribution function will be set to “rexp(40, 0.2)”.

I simulate 1000 samples for each size 40 with exponential distribution λ=0.2 by using rexp(n, lambda). The mean of exponential distribution is 1/λ. The standard deviation is also 1/λ. I generate the samples and calculate the average of each sample.

```
no_simulation <- 1000   # number of simulations
lambda <-  0.2
n <- 40             # sample size


simulated_data <- matrix(rexp(n= no_simulation*n,rate=lambda), no_simulation, n)
sample_mean <- rowMeans(simulated_data)
```

Sample Mean Vs. Theoretical Mean

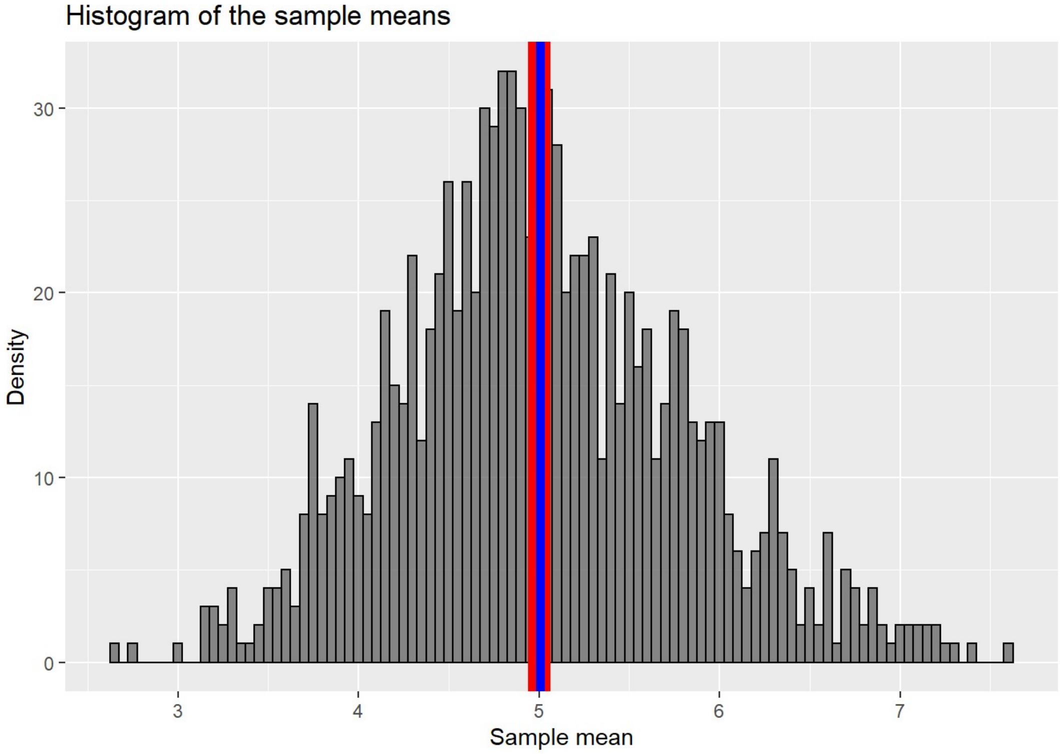
The theoretical mean of the average of samples will be : 1/λ .The following shows that the average from sample means and the theoretical mean are very close.

```
actual_mean <- mean(sample_mean)
theoretical_mean <- 1/ lambda

result1 <-data.frame("Mean"=c(actual_mean,theoretical_mean),
                     row.names = c("Mean from the samples ","Theoretical mean"))

result1
```

```
##                             Mean
## Mean from the samples  5.007836
## Theoretical mean       5.000000
```

The simulation mean of 4.983227 is close to the theoretical value of 5. Histogram plot of the exponential distribution n = 1000

```
sampleMean_data <- as.data.frame (sample_mean)

ggplot(sampleMean_data, aes(sample_mean))+
  geom_histogram(alpha=.7, position="identity", col="black", binwidth=.05)+
  geom_vline(xintercept = theoretical_mean, colour="red",size=5, show.legend=TRUE)+
  geom_vline(xintercept = actual_mean, colour="blue", size=2, show.legend=TRUE)+
  ggtitle ("Histogram of the sample means ")+
  xlab("Sample mean")+
  ylab("Density")
```



###

Sample Variance Vs. Theoretical Variance

The theoretical variance of the average of samples will be (1/λ)^2/n. The following shows that the variance of sample means and the theoretical variance are very close in value.

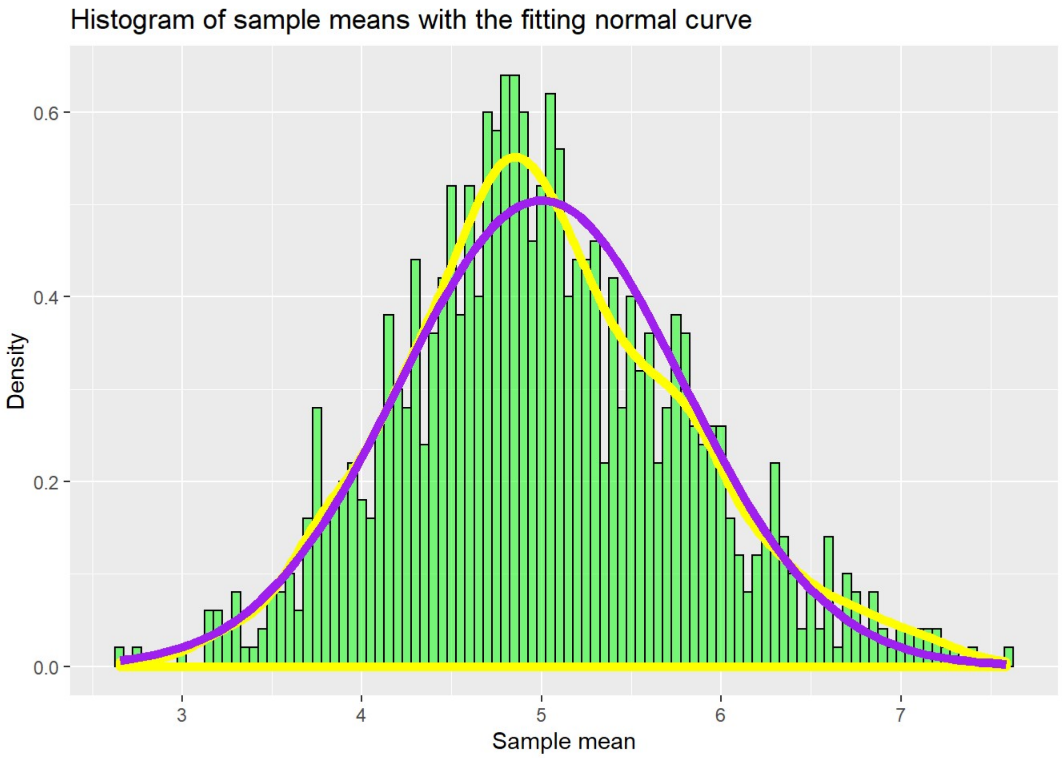
```
actual_variance <- var(sample_mean)

theoretical_variance <- (1/ lambda)^2 /n

result2 <-data.frame("Variance"=c(actual_variance, theoretical_variance),
                     row.names = c("Variance from the sample ","Theoretical variance"))
```

Distribution

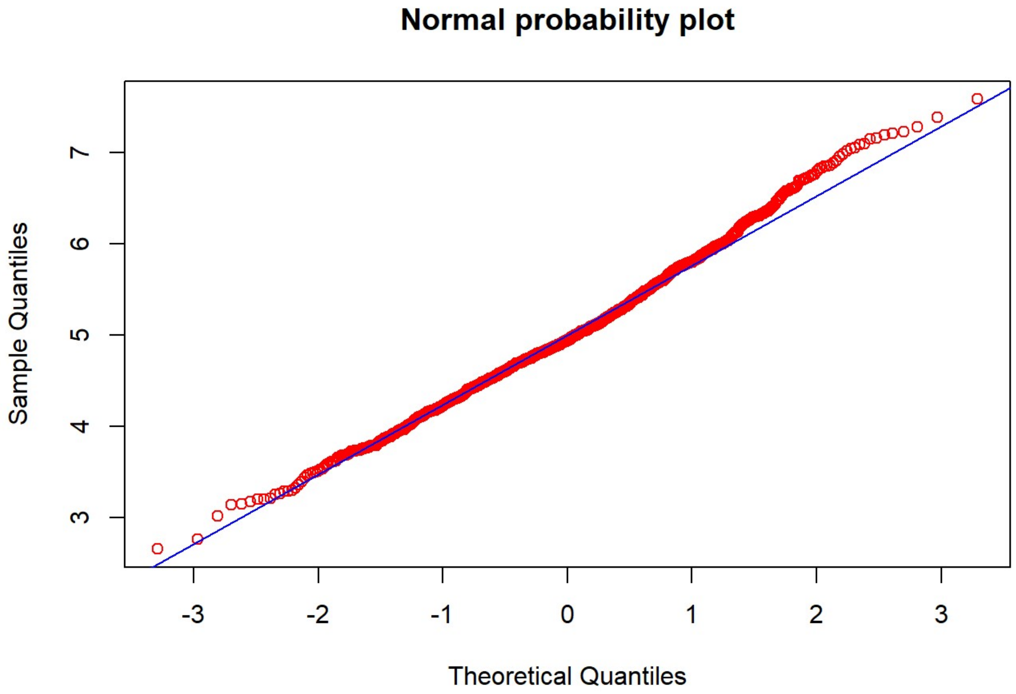
According to the central limit theorem (CLT), the averages of samples follow normal distribution.

The following plot shows that the distribution of the sample means almost matches the normal distribution. Also I created a Normal Probability Plot of Residuals below to confirm the fact that the distribution of sample means matches the theoretical normal distribution.

```
ggplot(sampleMean_data, aes(sample_mean))+
  geom_histogram(aes(y=..density..),alpha=.5,position="identity", fill="green",col="black",binwidth=0.05)+
  geom_density(colour="yellow", size=2)+
  stat_function(fun = dnorm, colour = "purple", size=2, args = list(mean = theoretical_mean, sd = sqrt(theoret
ical_variance)))+
  ggtitle ("Histogram of sample means with the fitting normal curve ")+
  xlab("Sample mean")+
  ylab("Density")
```



```
qqnorm(sample_mean, main ="Normal probability plot",col="2")
qqline(sample_mean,col = "4")
```



Both histogram and the normal probability plot show that distribution of averages is approximately normal.