Statistical Inference: Course Project - Part 1

Nikita Kirnosov

Contents

Overview

In this project we 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 λ is the rate parameter. The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$.

We will perform a thousand simulations to investigate the distribution of averages of 40 exponentials with λ set to 0.2. The validity of the CLT will be evaluated and discussed.

Simulations

The seed was set to 111 to make the results reproducible.

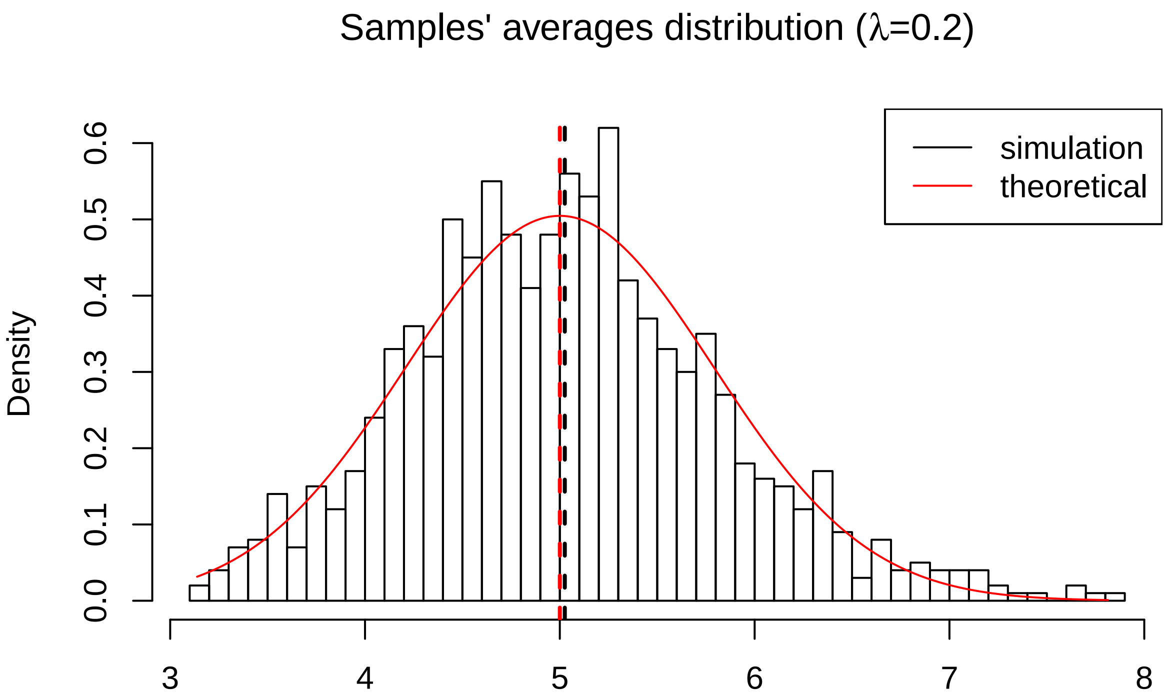
```r
set.seed(111)
lambda <- 0.2
num_sim <- 1000
sample_size <- 40
```

Since we only using means of each simulated normal distribution, we will simulate 40000 numbers and store them in a matrix with 1000 columns and 40 rows. Taking the mean of the rows will leave us with a vector of sample means.

```r
means <- rowMeans(matrix(rexp(num_sim*sample_size, rate=lambda), num_sim, sample_size))
```

Results

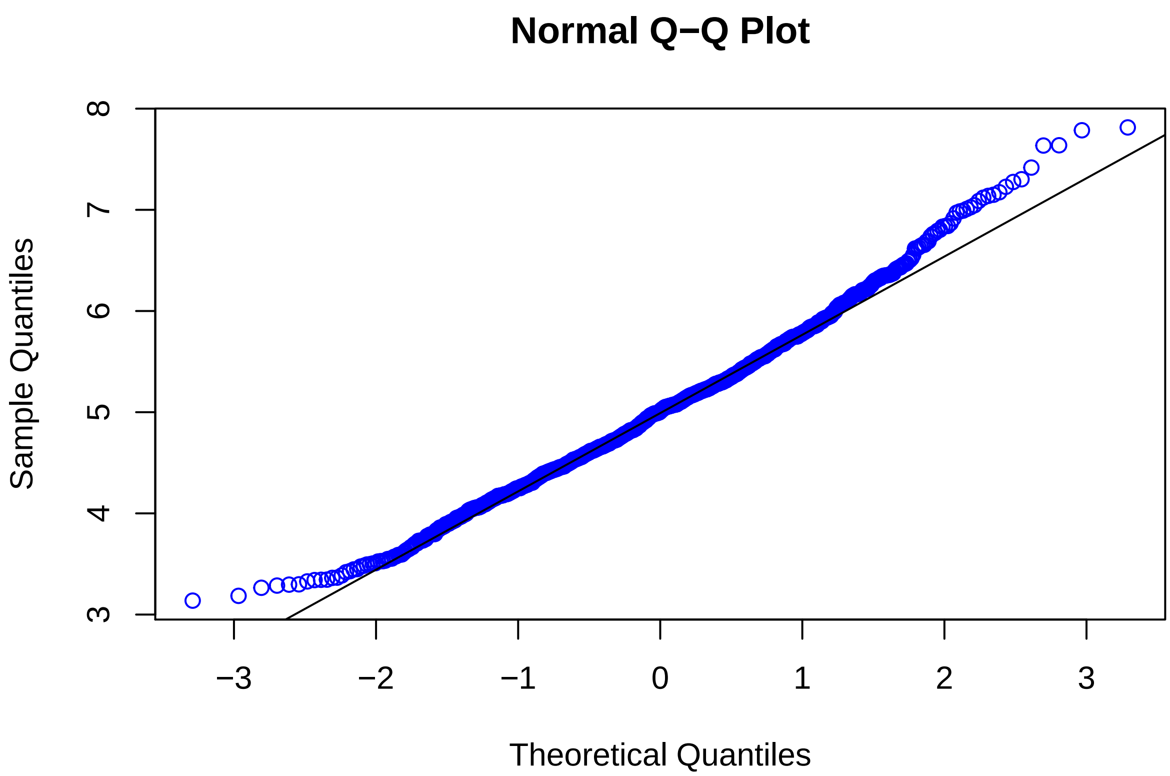
The histogram below overlays the distribution of samples' means and its mean (black dashed line) with the theoretically predicted shape and mean (in red).



Aside from the visual examination, the comparison of these distrubutions' mean and variance values reveals that they ar very close, thus supporting the hypothesis of distribution normality:

	Simulated	Theoretical
mean	5.03	5.00
variance	0.631	0.625

Another aid in discussion of sample normality is the normal Q-Q plot. The linearity of the points suggests that the data are normally distributed.



Conclusion

The obtained data suggests that the Central Limit Theorem is applicable in this case. This assumption can be also confirmed by Kolmogorov-Smirnov normality test, where high p-value would suggest the normal distribution:

```
ks.test(means,"pnorm",mean(means),sqrt(var(means)))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  means
## D = 0.04, p-value = 0.1
## alternative hypothesis: two-sided
```

Appendix

Plain R code used in this project:

```
library(knitr)
library(xtable)
opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
knit_hooks$set(inline = function(x) {
   if (is.numeric(x)) round(x, 2)})
options(digits=2)

set.seed(111)
lambda <- 0.2
num_sim <- 1000
sample_size <- 40

means <- rowMeans(matrix(rexp(num_sim*sample_size, rate=lambda), num_sim, sample_size))

hist(means, breaks=sample_size, prob=TRUE,
     main=expression(paste("Samples' averages distribution (",lambda,"=0.2)")),
     xlab="")
abline(v=mean(means),lty=2, lwd=2, col="black")
abline(v=1/lambda,lty=2, lwd=2, col="red")
xfit <- seq(min(means), max(means), length=num_sim)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(sample_size)))
lines(xfit, yfit, pch=22, col="red")
legend('topright', c("simulation", "theoretical"), lty=c(1,1), col=c("black", "red"))

rbind(c("","Simulated","Theoretical"),
        c("mean",signif(mean(means),3),"5.00"),
        c("variance",signif(var(means),3),signif(1/(lambda^2 * sample_size),3)))

qqnorm(means,col="blue")
qqline(means)

ks.test(means,"pnorm",mean(means),sqrt(var(means)))
```