Statistical Inference Course Project, Data Science Specialization on Coursera. John Hopkins University, 2015-5-15

Kun Zhu

The purpose of this project assignment is to demonstrate of asymptotical phenomena of a exponential distribution with a large size. A distribution, referred to as sample distribution, is created with one thousand averages from samples from an exponential distribution. Specifically, the theoretical mean and standard deviation from studied exponential distribution are compared with mean and standard deviation of the sample distribution in this work.

The theoretical mean and standard deviation of the exponential population are both 1/lambda. By plugging in the lambda of the studied exponential population, 0.2, the theoretical mean and distribution of the population are 5. As there are one thousand observations in the sample distribution, the standard deviation of the exponential distribution of interest is calculated as 5/ sqrt(1000)= 0.7906.

The sample distribution includes a thousand averages from 40 samples from the studied exponential distribution. The histogram and box plot of the sample distribution are presented in Figure 1. below.

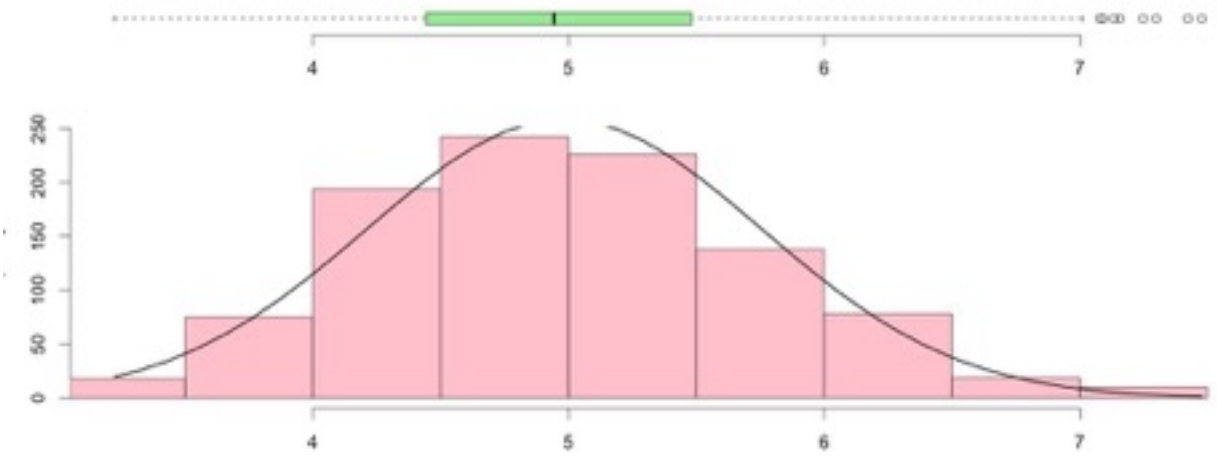

Figure 1 Histogram (in pink) and box plot of the sample distribution and histogram of the normal distribution with mean=1/lambda and sd=1/lambda/sqrt(1000) (in black)

It is apparent that the histogram of the sample distribution follows the pattern of the plotted normal distribution. For comparison purpose, the box plot of the sample distribution is also provided. The mean and standard deviations are presented in the Table 1.

	Mean	Standard Deviation
Exponential Distribution	5	0.7906
Sample Distribution	5.0183	0.7896
Difference	0.0183	0,0010

Table 1 Mean and standard deviation of the studied exponential distribution and sample distribution.

It is clear that the mean and standard deviation of the studied exponential distribution and sample distribution are very close. Therefore, the study results agree with Central Limit Theorem that the distribution of averages of iid variables becomes that of a normal distribution when the sample size is large.

Appendix, R code

```
library(ggplot2)

# n1 is the number of averages of the samples from the studied exponential distribution.
# n is number of exponentially distributed samples used to calculate averages.
# lambda is the expotential distribution rate
n  <- 1000
n1  <- 40
lambda  <- 0.2

# Calulcate the theoritical mean and standard deviation of the exponential distribution.
theo_mean  <- 1/lambda
theo_sd <- 1/lambda

# Create a new distribution including n number of exponential distribution mean.
mns = NULL
for (i in 1 : n) mns = c(mns, mean(rexp(n1, lambda)))

# Calulcate the mean and standard deviation of the averages of the samples from the exponential
distribution.
# These values are referred to as sampled mean and standard deviation.
smp_mean  <- mean(mns)
smp_sd <- sd(mns)

# Calulcate the difference between the theoritical and sampled mean and standard deviation.
delta_mean <- sqrt((smp_mean -theo_mean)^2)
delta_sd  <- sqrt((smp_sd -theo_sd/sqrt(n1))^2)

# Plot the histogram and boxplot of the sampled averages.
# together with a normal distribution with mean= 1/lambda and standard deviation = 1/lambda.
nf <- layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(1,3))
par(mar=c(3.1, 3.1, 1.1, 2.1))
boxplot(mns, horizontal=TRUE,  outline=TRUE,ylim=c(min(mns),max(mns)), frame=F, col =
"lightgreen")
g  <- hist(mns, xlim=c(min(mns),max(mns)), xlab="Frequency", main="", col="pink")
xfit<-seq(min(mns),max(mns),length=40)
yfit<-dnorm(xfit,mean=theo_mean,sd=theo_sd/sqrt(n1))
yfit <- yfit*diff(g$mids[1:2])*length(mns)
lines(xfit, yfit, col="black", lwd=2)
```