Investigating the Exponential Distribution in R

By: Oleg Rozenshtain

Overview

This report is written as a solution to the first part of the "statistical inference" course project, as part of the "data science" specialization by Johns Hopkins University on Coursera.

The purpose of the project is to investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). I am going to illustrate via simulation the properties of the distribution of the mean of exponentially distributed iids.

Method

Using R I randomly simulated 1000 samples of 40 exponentially distributed iids with rate $\lambda = 0.2$. I used the *replicate()* function which repeatedly evaluates an expression to generate a matrix with the samples as columns.

```
> simulationNumber <- 1000
> lambda <- 0.2
> sampleSize <- 40
> set.seed(12)
> sampleVector <- replicate(simulationNumber, rexp(n = sampleSize, rate = lambda))
```

Next, I calculated for each sample the mean in order to create a sample mean distribution. I used the *colMeans()* function, that returns a vector of the mean of each column in the matrix. Apply *mean()* and *var()* on the sample mean vector to get the estimated mean and variance of the distribution.

```
> sampleMeanVector <- colMeans(sampleVector)
> mean(sampleMeanVector)
[1] 5.010015
> var(sampleMeanVector)
[1] 0.599168
```

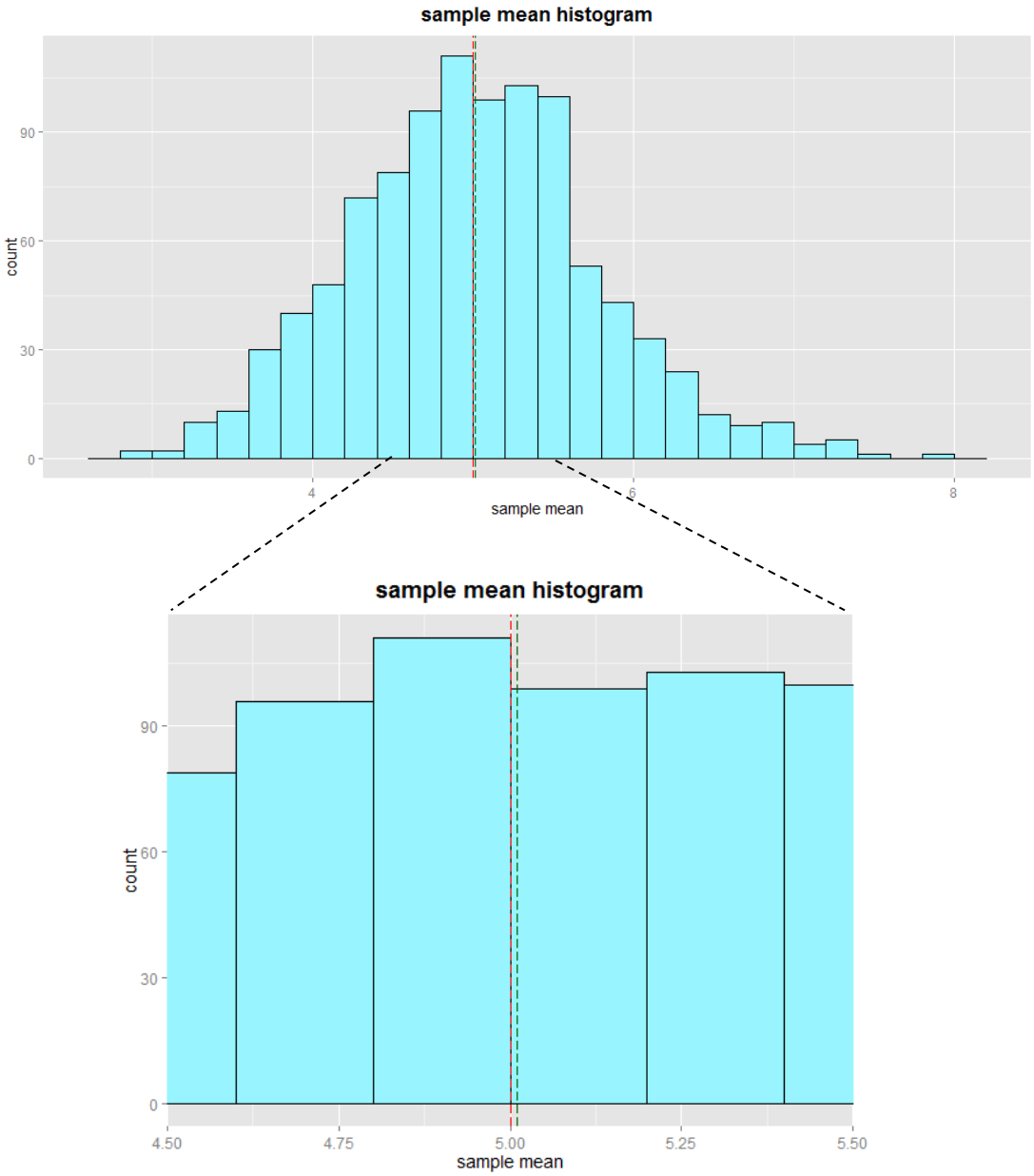
To demonstrate the CLT I am going to show that the distribution is approximately normal and compare its density function to the relevant normal distribution density function.

Results

All plots shown in this section were made with *ggplot2* package. The full code can be found in the appendix.

First we want to compare the samples mean average value to the theoretical mean of the exponential distribution. The theoretical mean is: $E[X] = {}^{1}/_{\lambda} = {}^{1}/_{0.2} = 5.$

As expected the sample mean 5.01 calculated earlier is very close to the theoretical mean. The sample mean distribution is centered around the population mean as can be seen in the figures below (red - theoretical mean; green - sample mean):



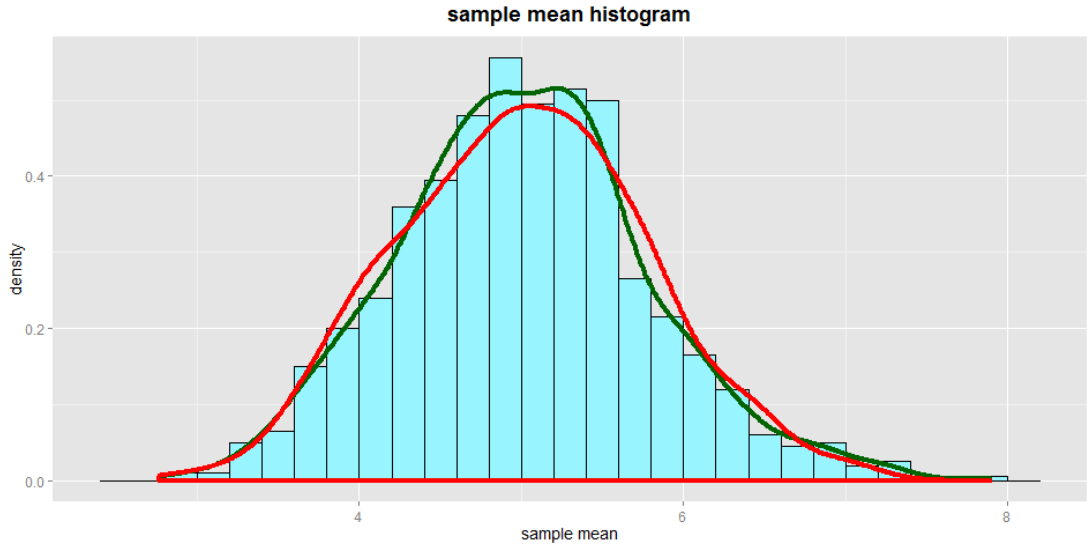
Now let's look at the theoretical variance of the exponential distribution: $Var(X) = \sigma^2 = \left({}^{1}/_{\lambda}\right)^2 = \left({}^{1}/_{0.2}\right)^2 = 25$, which is very different from variance I obtained earlier: 0.6. That's simply because those are different measures. The theoretical variance measures how variable the data within the sample, the sample mean variance measures how variable the mean of a random sample out of a population. The variance of the sample mean is the standard error squared, which is: $[SE(\bar{X})]^2 = {}^{\sigma^2}/_{n} = {}^{25}/_{40} = 0.625$, and now the similarity is noticed. Also, we can calculate the variance within each sample, and take the average of all these

variances to make sure it is close to the theoretical variance of the exponential distribution (which is 25 from above calculations) :

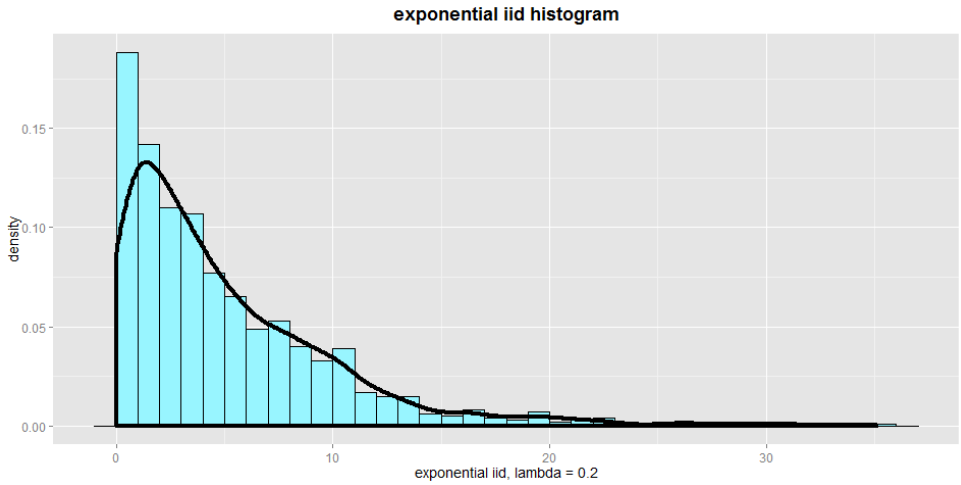
```
> mean(apply(sampleVector, 2, var))
[1] 24.72851
```

This code uses the *apply()* function to execute the *var()* function on each column (MARGIN=2) to calculate the variance within each sample. Finally the mean of all the variances is calculated.

Now let's check if the sample mean distribution is normal. The CLT states that the sample mean should be normally distributed with $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. In my example I expect $\bar{X} \sim N\left(5, \sqrt{0.625}\right)$. Let's look at a smooth density estimate curve to the sample mean and a smooth density estimate curve for the theoretical $\bar{X}$ (red - theoretical normal density plot; green - sample mean density plot).



It's important to notice the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials. As shown above the collection of averages of 40 exponentials is normally distributed, however a large collection of random exponentials is obviously exponentially distributed by definition. You can see it the following figure of a thousand randomly simulated exponential iids with $\lambda = 0.2$.



Appendix

```
investigateExponential<-function()
{
  library(ggplot2)

  # set exponential sample simulations parameters.
  # simulate 1000 different samples of size 40, out of exponential
  # distribution with rate 0.2
  simulationNumber<-1000
  lambda<-0.2
  sampleSize<-40

  set.seed(12)
  # simulate the samples. the result is a matrix with each sample as a column
  sampleVector<-replicate(simulationNumber, rexp(n = sampleSize, rate = lambda))
  # calculate the mean for each sample
  sampleMeanVector<-colMeans(sampleVector)

  # set title and axis themes for all future plots
  plotTheme<-theme(plot.title = element_text(size = 18, face = "bold", vjust = 2),
                   axis.title = element_text(size = 14),
                   axis.text = element_text(size = 12))

  # convert to a data frame, because ggplot works only with data frames
  dfSampleMeanVector<-data.frame(sampleMeanVector)
  # create a histogram out of samples means
  pMean<-ggplot(dfSampleMeanVector, aes(x = sampleMeanVector)) +
         labs(title = "sample mean histogram", x = "sample mean") +
         geom_histogram(binwidth = 0.2, colour="black", fill="cadetblue1") +
         plotTheme

  # add vertical lines to the histogram to show the theoretical mean of the
  # exponential distribution (=1/lambda) and the sample mean
  pMeanVerticalLines<-pMean +
       geom_vline(xintercept = c(1/lambda, mean(sampleMeanVector)) ,
                  colour = c("red", "darkgreen"), linetype = "longdash")

  # add a smooth density estimate curve to the histogram to show gaussian form.
  # add a smooth normal density estimate curve to show similarity
  pMeanDensity<-pMean + aes(y = ..density..) +
    geom_density(colour = "darkgreen", size = 1.5) +
    geom_density(data = data.frame(x = rnorm(n = simulationNumber,
                                             mean = 1/lambda,
                                             sd = (1/lambda)/sqrt(sampleSize))),
                 colour = "red", size = 1.5)
```

```
  # create a data frame with a 1000 iids exponntially distributed with lambda=0.2
  dfExponentialVector<-data.frame(exponentialVector = rexp(n = simulationNumber,
                                                           rate = lambda))
  # create a histogram of the exponntially distributed iids and add a smooth
  # density estimate curve to the histogram to show exponential form instead
  # of gaussian
  pExponentialDensity<-
    ggplot(dfExponentialVector, aes(x = exponentialVector, y = ..density..)) +
    labs(title = "exponential iid histogram", x = "exponential iid, lambda = 0.2") +
    geom_histogram(binwidth = 1, colour="black", fill="cadetblue1") +
    geom_density(size = 1.5) + plotTheme


  # save all plots to png files
  png("sample_mean_histogram.png", width = 960)
  print(pMeanVerticalLines)
  dev.off()

  png("sample_mean_histogram_zoomin.png", width = 600)
  print(pMeanVerticalLines + coord_cartesian(xlim = 1/lambda + c(1,-1)*0.5))
  dev.off()

  png("sample_mean_density.png", width = 960)
  print(pMeanDensity + ylab("density"))
  dev.off()

  png("exponential_density.png", width = 960)
  print(pExponentialDensity + ylab("density"))
  dev.off()
}
```