Assignment 1 simulation

Joris Schut

Saturday, February 28, 2015

Introduction

This document was made in the context of the statistical inference MOOC by Johns Hopkins University as part of the Data Science specialization on Coursera. In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem.

1. Show the sample mean and compare it to the theoretical mean of the distribution.

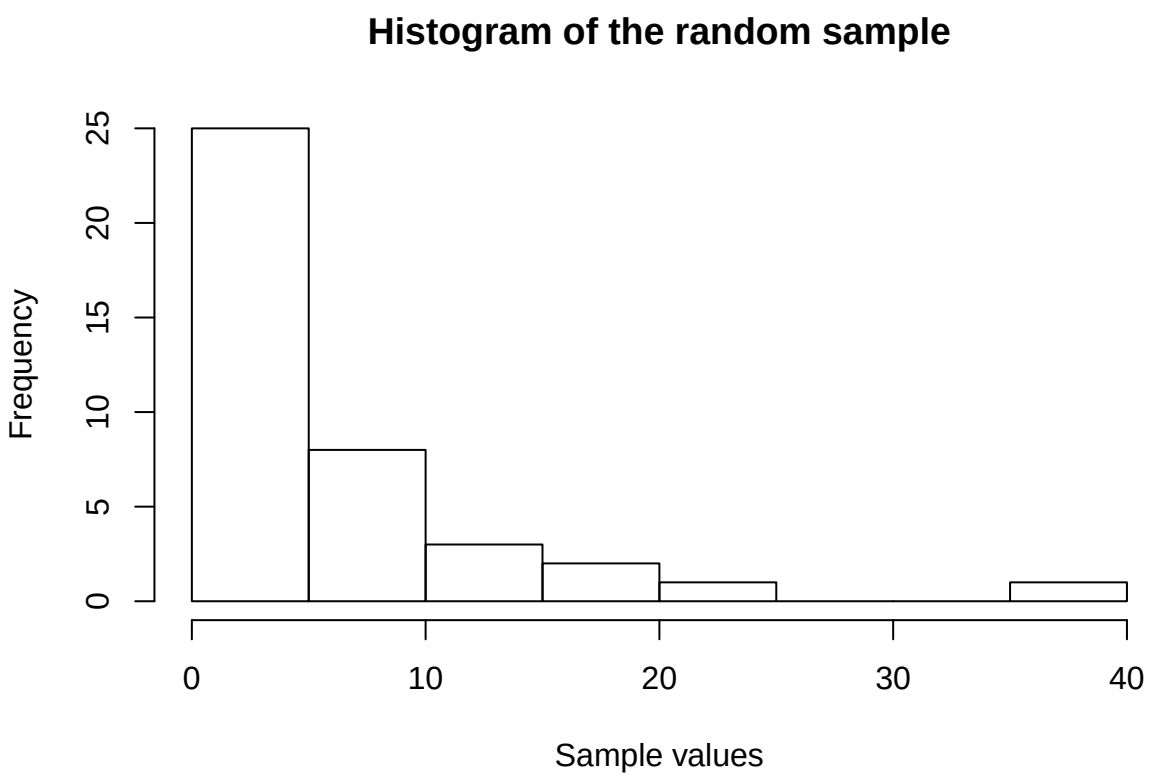
First, the parameters for the sample size (n), lambda and the number of experiments (runs) was set (values given in the assignment details). In order to create a reproducible analysis, the seed was set to 111.

```
#Set parameters
set.seed(111)
n = 40
lambda = 0.2
runs = 1000
```

Using the above variables the experiential mean of the experiment was determined. The values of the experiment were then plotted and the value of the mean was printed.

```
#Determine the experimental mean
expvalues <- rexp(n, lambda)
expmean1 <- mean(expvalues)

hist(expvalues, main="Histogram of the random sample", xlab="Sample values")
```




```
print(expmean1)
```

```
## [1] 5.720126
```

Based on the characteristics of the exponential distribution, the theoretical mean was calculated.

```
#Calcualte the theoretical mean[1]
theomean1 <- 1/lambda
print(theomean1)
```

```
## [1] 5
```

The difference between the experiential and theoretical value is given by subtracting both values and taking the absolute value of the result.

```
#Determine the absolute difference between the theoretical and experimental mean
diffmean1 <- abs(theomean1 - expmean1)
print(diffmean1)
```

```
## [1] 0.7201257
```

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution

Using the above variables the experiential variance of the experiment was determined.

```
#Determine the experiental variance
expvar1 <- var(expvalues)
print(expvar1)
```

```
## [1] 48.77622
```

Based on the characteristics of the exponential distribution, the theoretical mean was calculated for a distribution of 40 exponentials.

```
#Calculate the theoretical variance[1]
theovar1 <- 1/(lambda^2)
print(theovar1)
```

```
## [1] 25
```

The difference between the experiential and theoretical value is given by subtracting both values and taking the absolute value of the result.

```
#Determine the absolute difference between the theoretical and experimental mean
diffvar1 <- abs(theovar1 - expvar1)
print(diffvar1)
```

```
## [1] 23.77622
```

3 Show that the distribution is approximately normal

1000 experiments were conducted were the mean value of the experiment was stored in the mns variable. If the distribution of means is normal, the histogram of the means be should bell-shaped and can be approximated by a normal distribution density function with the theoretical average for the distribution of a large collection of averages of 40 exponentials and the corresponding standard deviation (Central limit theorem). To compensate for the number of runs the values of the density function should be multiplied by the binwidth of the histogram*number of experiments[2].

First, the experiment is run for n=1000 (runs variable) times. The results are appended to the mns variable

```
#Determine the experimental mean
mns <- NULL
for (i in 1 : runs){
  mns = append(mns, mean(rexp(n, lambda)))
}
```

Second, the values of the normal approximation are calculated for the distribution of a large collection of averages of 40 exponentials. Further, a normal density function is run to obtain the values used for the approximation.

```r
#Calculate theoretical values for the average and the variance
theomean2 <- theomean1
theovar2 <- (1/(lambda^2))/n
print(theomean2)
```

```
## [1] 5
```

```r
print(theovar2)
```

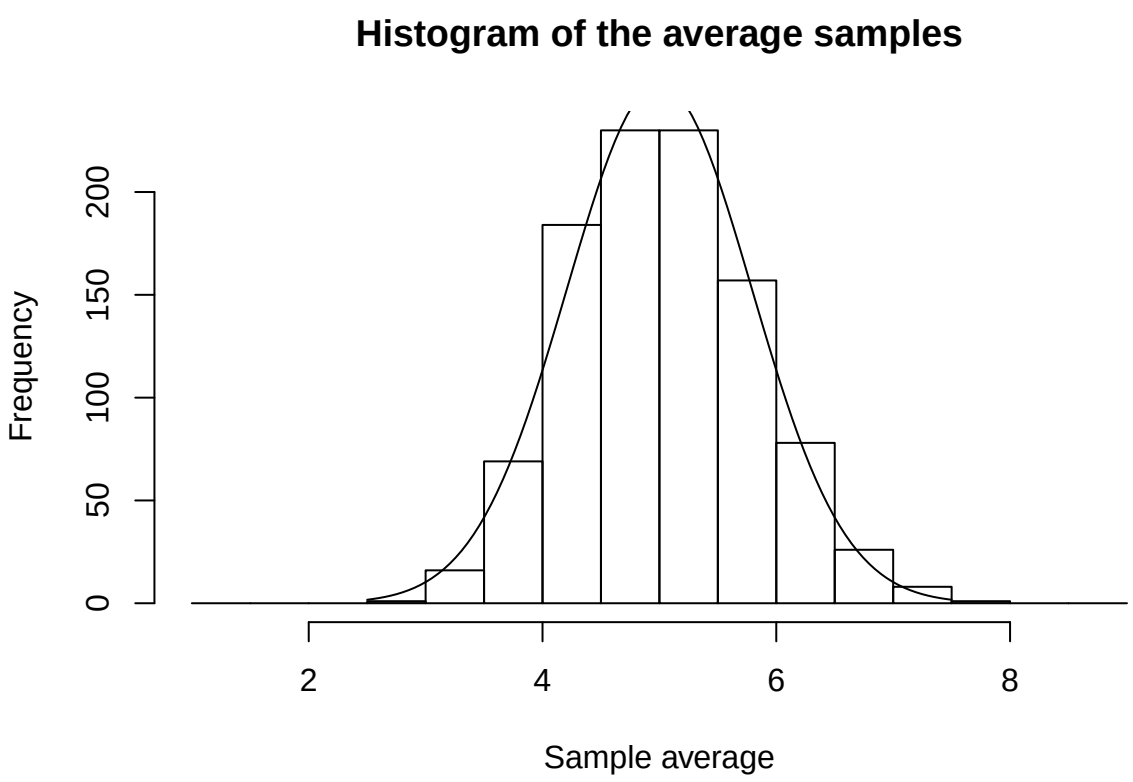
```
## [1] 0.625
```

Third, a normal approximation can be simulated by using the the theoretical for the distribution of a large collection of averages of 40 exponentials and the corresponding standard deviation (Central limit theorem). To compensate for the number of runs the values of the density function should be multiplied by the binwidth of the histogram*number of experiments[2].

```r
#Create a an normal approximation
binwidth <- 0.5
x <- seq(-4, 4, length=100)*theovar2 + theomean2
hx <- binwidth*runs*dnorm(x, mean=theomean2, sd=sqrt(theovar2))
```

Fourth, the results of both the mns variable and the normal approximation are plotted in a single plot.

```r
#Plot the histogram of the experiments and the normal approximation
hist(mns, breaks=seq(1,9,by=binwidth),main="Histogram of the average samples",
     xlab="Sample average")
lines(x, hx)
```



As the normal approximation fits pretty good with the histogram, normality of the distribution of the mean can be assumed. This is as stated in the central limit theorem.

External sources used for reference

[1] [Wikipedia](http://en.wikipedia.org/wiki/Exponential_distribution#Mean.2C_variance.2C_moments_and_median)

[2] [Fitting a Gausian curve to a histogram](http://www.theinformationlab.co.uk/2013/11/04/fitting-a-gaussian-normal-distribution-curve-to-a-histogram-in-tableau/)