Statistical Inference Course Project - Part 1

Zohar Peleg

Tuesday, August 12, 2014

This report covers the simulation excercize from the first part of the statistical inference course project. The report does not show the full R code. The complete reproduceable code can be found at my github account at https://github.com/zoharpeleg/Statistical-Inference-Course-Project

In this excercise I have investigated the properties of the distribution of the mean of 40 random exponential variables, using R simulation of the exponential distribution. The simulation is using the *rexp(n,lambda)*, where *lambda* is the rate parameter, and *n* is the sample size. I used *lambda=0.2* and *n=40* for all simulations,as required by the project instructions. In order to get an idea about the distribution of the sample's properties, I have repeated the simulation for a large number of times *(nosim=1000)*.

Setting the simulation parameters:

```
lambda<-0.2 ; n<-40 ; nosim<-1000
```

The simulated samples are stored in a matrix, where each row contains one sample of 40 random exponential variables. The mean and standard-deviation were calculated for each sample, and stored in two vectors - *meanx*, and *sdx* respectively:

```
x<-matrix(rexp(n*nosim,lambda),nosim) #nosim samples of n observations
meanx<-apply(x, 1, mean)              #the means of all samples
sdx<-apply(x, 1, sd)                  #the std-deviations of all samples
```

1. Showing where the distribution is centered at, and comparing it to the theoretical center of the distribution:

The average mean of all the samples is:

```
mean(meanx)
```

```
## [1] 4.993
```

compared with the theoretical mean of the exponential distribution, which is $\mu = \frac{1}{\lambda} = 5$

2. Showing how variable the sample mean is, and comparing it to the theoretical variance of the distribution:

The calculated variance of the sample means is:

```
var(meanx)
```

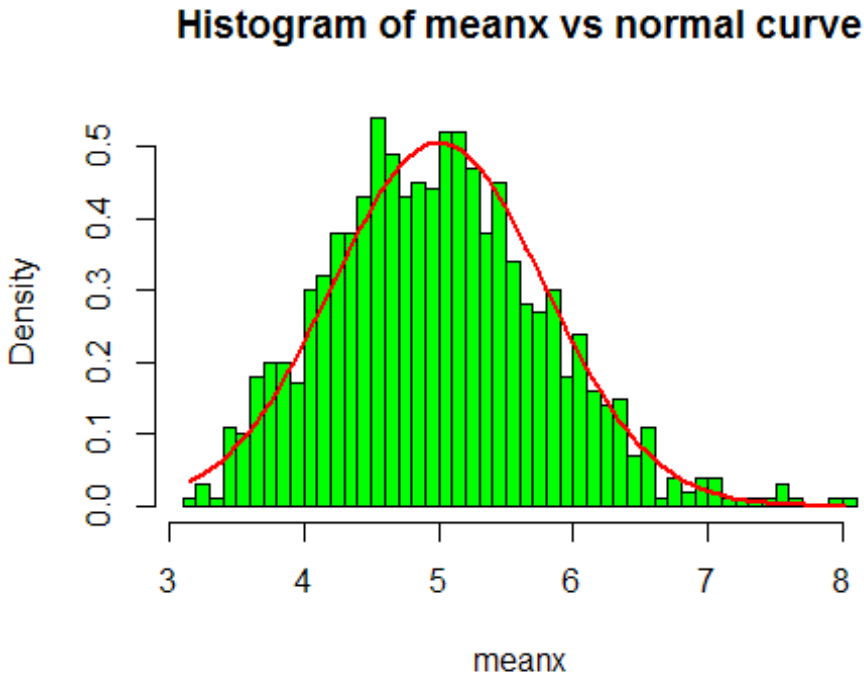
```
## [1] 0.6444
```

compared with the theoretical variance of the sample mean, for sample-size of *n*, which is

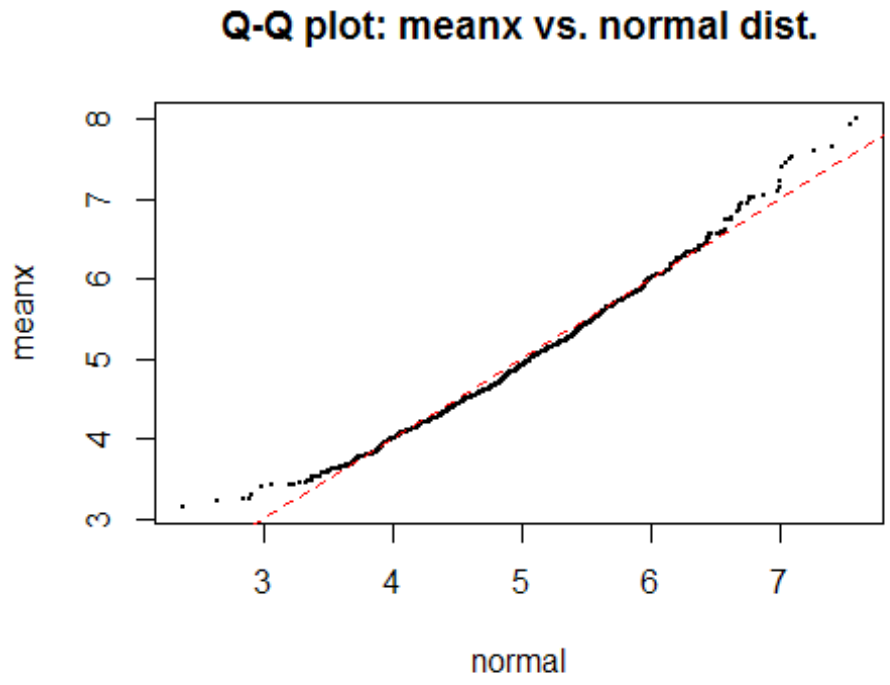
$$s^2 = \frac{\sigma^2}{n} = \frac{\left(\frac{1}{\lambda}\right)^2}{n} = \frac{5^2}{40} = 0.625$$

3. Showing that the distribution is approximately normal:

a. By plotting a normalized histogram of the sample means, compared to the density function curve of normal distribution of $\mu = \frac{1}{\lambda}$ and $\sigma = \frac{\frac{1}{\lambda}}{\sqrt{n}}$:



b. By plotting a Q-Q plot of the sample means compared to random normal sample with the same expected parameters as above:



4. Evaluation of the coverage of the confidence interval for $\frac{1}{\lambda}$ using $\overline{X} \pm 1.96 * \frac{s}{\sqrt{n}}$:

With the sample's mean and standard deviation, I have calculated the interval's lower limit and upper limit for each sample:

```
ul<-meanx+sdx*1.96/sqrt(n) #upper limit vector
ll<-meanx-sdx*1.96/sqrt(n) #lower limit vector
```

Now it is possible to evaluate the percentage of good intervals, which contain the *mean=5*, as expected:

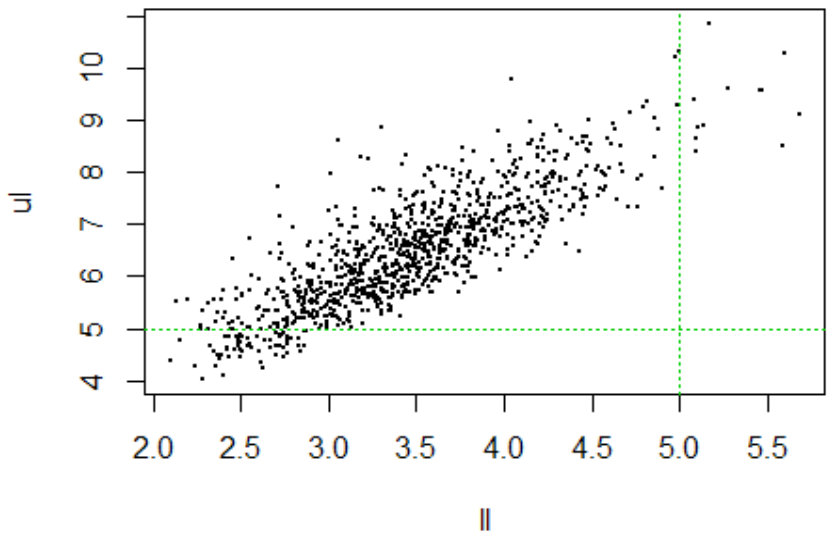
```
paste(mean(1/lambda<ul&1/lambda>ll)*100,"% of the intervals contained the
population mean",sep="")

## [1] "92.2% of the intervals contained the population mean"
```

Repeating the set many times, I have found that the coverage was consistently about 92%. I tried various sample sizes and figured out that when the sample size is increased from 40 to 400, the coverage goes up to 95%. My conclusion is that a sample size of 40 exponentials(0.2), is not sufficiently large for 95% confidence by CLT.

I also tried replacing the normal CLT interval with t-interval, by replacing the 1.96 with qt(0.975,n-1) = 2.0227. The coverage increased to about 93%, which is about 0.5%-1% improvement.

I am adding an interesting view of the distribution of the calculated intervals compared to the population's mean, by plotting each sample's interval as *(x=ll; y=ul)*:



The point (5,5) represents the ultimate interval (ul=ll=5). All the points in the upper-left rectangle represent good intervals which contain the population mean $\frac{1}{\lambda}$ (ll<5 and ul>5). The points in the upper-right area are the upper tail, in which the interval's lower limit is greater than 5, and the lower-left area is the lower tail, in which the interval's upper limit is smaller than 5.