Sample-based analysis of exponential distribution

Michael Krämer

25.10.2015

Abstract

The report shows the result of an analysis of the distribution of the mean of 40 exponentials (values drawn from the exponential distribution) using R. The analysis is based on random generated sample data and compares properties of the sample set to the theoretical values of the distribution. 1000 samples (of 40 exponentials each) will be generated.

Data generation

The following code was used to generate 1000 samples of the mean of 40 random exponentials. Lambda was set to 0.2

The R code creates a list of 1000 number elements (as in intermediate step) and uses then the sapply function to replace every element of that list with the mean of 40 randomly generated number from the exponential distribution. The seed is used to ensure reproducible results, but its value is not particularly relevant.

```
set.seed(174) # just ensure reproducability
lambda = 0.2
size = 40
m_data <- data.frame(xb = sapply(rep(0, 1000), function(x){mean(rexp(size, lambda))}))
```

Expected properties

Let X_i be the ith of our generated sample. The sample data is expected to follow a normal distribution like the distribution of averages of any underlying distribution does.

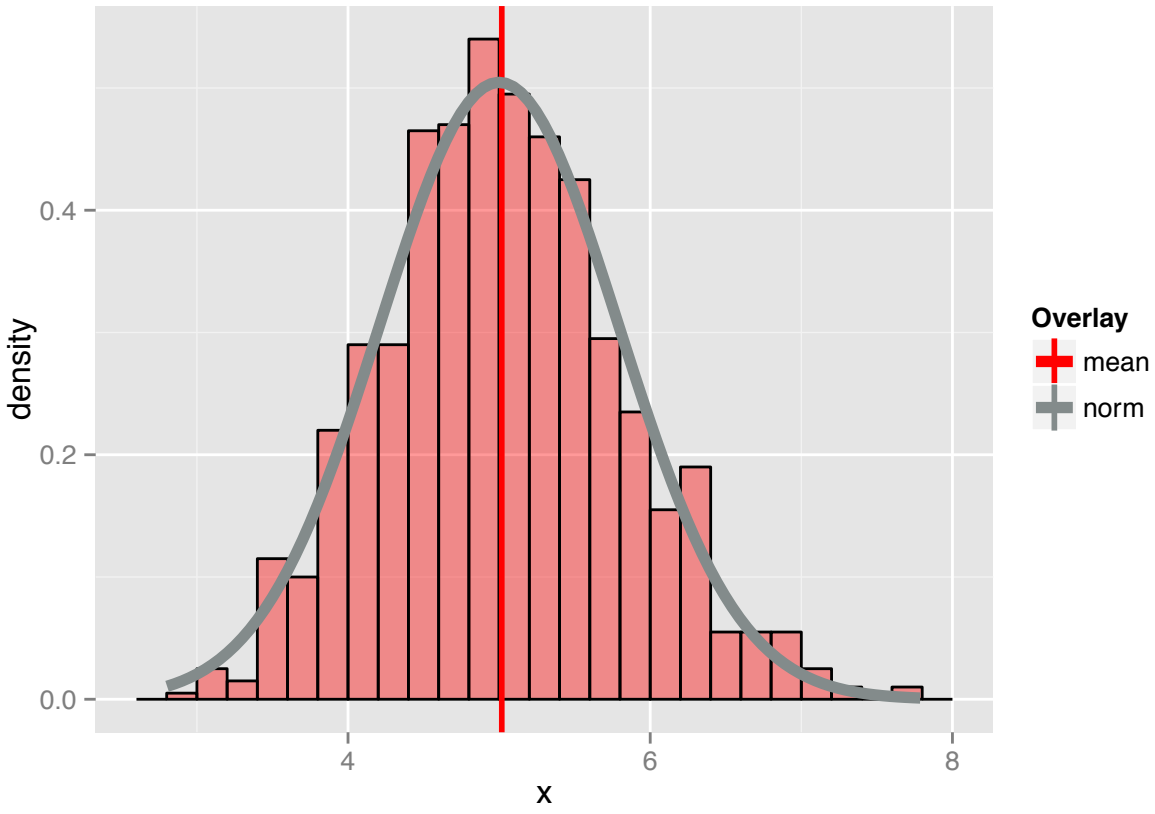
Since the mean of the chosen exponential distribution is $\mu = \frac{1}{\lambda}$ and $\lambda = 0.2$ we expect $E[X_i] = 5$ as the mean of our samples. This is due to the fact that averages converge to the value they are estimating, which is in our case the mean of the underlying exponential distribution.

The standard deviation of the underlying exponential distribution is defined as $\sigma = \frac{1}{\lambda} = 5$ and the variance $Var[Exp] = \frac{1}{\lambda^2} = 25$ with respect to our chosen value of $\lambda = 0.2$.

The distribution of averages of 40 exponentials in our case is expected to have a variance of $Var[X_i] = \frac{\sigma^2}{n} = \frac{25}{40} = 0.625$ and a standard error $SE = \frac{\sigma}{\sqrt{n}} = 0.79$. This follows from the relations given between the population mean (in our case the theoretical value from the distribution) and the sample mean.

Sample data histogram

A histogram is provided to give an overview of the generated sample data. As an overlay, the histogram shows the normal distribution with mean of 5 as a continuous curve.



xb
Min. :2.803
1st Qu.:4.476
Median :4.982
Mean :5.017
3rd Qu.:5.534
Max. :7.787

Achieved results

The plot and the data summary show that the sample mean of `mean(m_data$xb)` is close to the theoretical mean of 5.

```
var(m_data$xb)
```

```
## [1] 0.6201163
```

```
sd(m_data$xb)
```

```
## [1] 0.7874747
```

The empirical variance of the sample is 0.6201163 and the standard deviation of the sample is 0.7874747, both of which are quite close to the expected values from the theoretical consideration above. Again, it must be noted that the distribution of averages follows a normal distribution while the underlying single values are

drawn from an exponential distribution. The relations between these two distributions were already discussed in the section about expected properties of the data.

The overlay of the normal distribution in the plot shows as well, that the distribution of the samples is approximately normal

Sources

The sources for the report can be found on Github at https://github.com/mkraemerx/datasciencecoursera/tree/master/06StatisticalInference