Statistical Inference, Course Project - Part1

Ivan Maksimov

08.01.2015

1 Introdaction

1.1 The exponential distribution

In this project we should investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution (a.k.a. negative exponential distribution) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. The probability density function (pdf) of an exponential distribution is

$$\begin{cases} f(x;\lambda) = \lambda e^{-\lambda x}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

Here $\lambda > 0$ is the parameter of the distribution, often called the rate parameter. The distribution is supported on the interval $[0, \infty)$. The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$.

1.2 Objectives

In this project we should

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

2 Data Processing

2.1 Exponential sampling parameters

The exponential distribution can be randomly generated in R with the function `rexp(n, lambda)` where n is the number of observations and λ is the rate parameter. Set $\lambda = 0.2$ for all of the simulations. We investigate the distribution of averages of 40 exponentials and need to do a thousand simulations.

```
lambda <- 0.2 #Set rate to 0.2
sampleSize <- 40 #Set sample size to 40
nSamples <- 1000 #Draw 1000 samples
```

Let's do a thousand simulated averages of 40 exponentials.

```
expoDist <- replicate(n = nSamples, expr = rexp(n = sampleSize, lambda))
```

2.2 Calculate theoretical and sample summary statistics

2.2.1 Means

```
theoMean <- round((1/lambda), 3) #Theoretical
sampleMean <- round(mean(colMeans(expoDist)), 3) #Sample
```

2.2.2 Standard Deviations

```
theoSd <- round((1/lambda * (1/sqrt(sampleSize))), 3) #Theoretical
sampleSd <- round(sd(colMeans(expoDist)), 3) #Sample
```

2.2.3 Variance

```
theoVar <- round((theoSd^2), 3)
sampleVar <- round((sampleSd^2), 3)
```

2.3 Summary statistics and inferences

The theoretical and empirical summary statistics of the simulated exponential distribution for this exercise can be verified in Table 1 bellow.

Table 1: Summary statistics

	Mean	Standart Diviation	Variance
Theoretical	5	0.791	0.626
Sample	4.981	0.788	0.621

2.3.1 Answer to objective 1

The center of the simulated sampling distribution (4.981) is very close to the theoretical center (5) of the distribution.

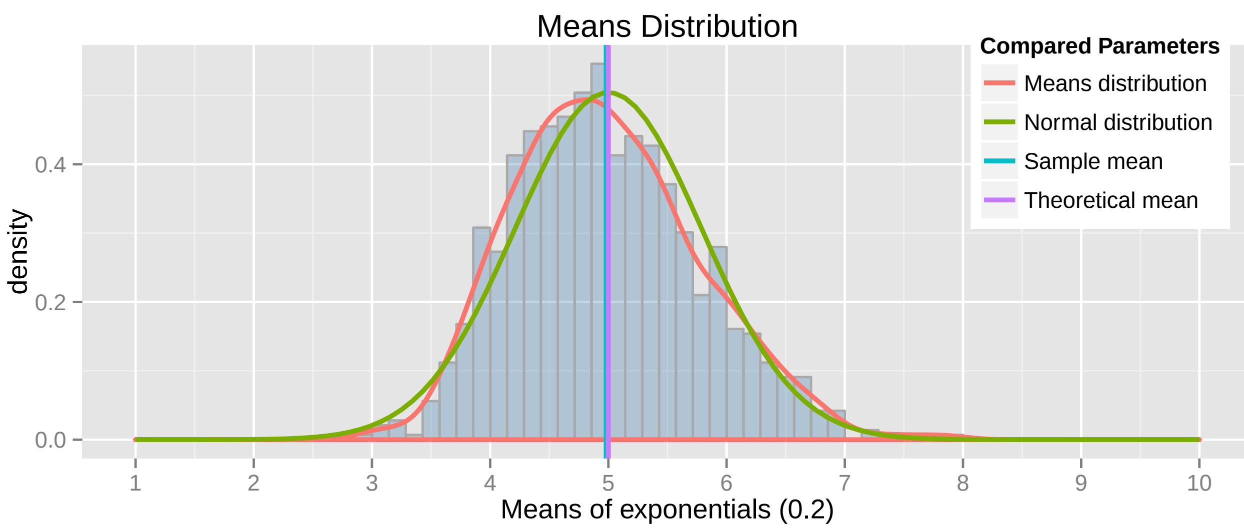
2.3.2 Answer to objective 2

The standard deviation of sampling 0.788 is also close to the theoretical standard deviation of 0.791.

2.3.3 Answer to objective 3

Due to the central limit theorem, the averages of samples follow normal distribution. The figure bellow also shows the density computed using the histogram and the normal density plotted with theoretical mean and variance values.

```r
library(ggplot2)
hist.df <- data.frame(Means = colMeans(expoDist))
ggplot(data = hist.df, aes(x = Means)) +
        geom_histogram(aes(y=..density..), fill = "steelblue",
                       binwidth = 1/7, color = "darkgrey", alpha = 1/3) +
        geom_density(aes(color = "Means distribution"), size = 1, show_guide=FALSE)+
        stat_function(fun=dnorm, arg=list(mean=theoMean, sd=theoSd),
                      aes(color = "Normal distribution"), size = 1)+
        geom_vline(aes(xintercept=sampleMean, colour="Sample mean"), size = 1)+
        geom_vline(aes(xintercept=theoMean, colour = "Theoretical mean"), size = 1)+
        theme(legend.justification=c(1,0), legend.position=c(1,0.5))+
        labs(title = "Means Distribution", x = "Means of exponentials (0.2)")+
        scale_x_continuous(limits = c(1, 10), breaks=1:10)+
        scale_color_discrete(name ="Compared Parameters")
```



The aproximate normality of this distribution can by also confirmed by the normality test depicted in the Q-Q plot.

```r
qqnorm(colMeans(expoDist),col = "steelblue")
qqline(colMeans(expoDist), col = "red",lwd=5)
```