Statistical Inference Course Project: The Central Limit Theorem (CLT) and Simulation Experiment

Michela Ieva

Thursday, May 21, 2015

Overview

We report about the results of a CLT simulation experiment. In this analaysis we will investigate the distribution of averages of 40 exponentials, performing 1000 simulations. According the CLT the resulting distribution looks like a bell curve with mean and standard deviation compatible with the theoretical values of *Normal* distribution.

```
library(knitr)
library(ggplot2)

echo = TRUE  # Always make code visible
```

Simulation

The exponential distribution can be simulated in R with the function rexp(n, lambda) where lambda is the rate parameter. The theoretical mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda.

First of all, for reproducibility we need to seet the seed than we set the simulation parameters.

```
set.seed(1234) # for reproducibility purpose

nosim <- 1000
n <- 40
lambda = 0.2
meanexp = 1/lambda
sigmaexp = 1/lambda
```

We use the funcion *replicate* to perform a thousand of simulations of a sample of 40 exponential distributions and take the mean of each sample.

```
dat <- replicate(nosim, mean(rexp(n, lambda)))
```

Our dataset looks like:

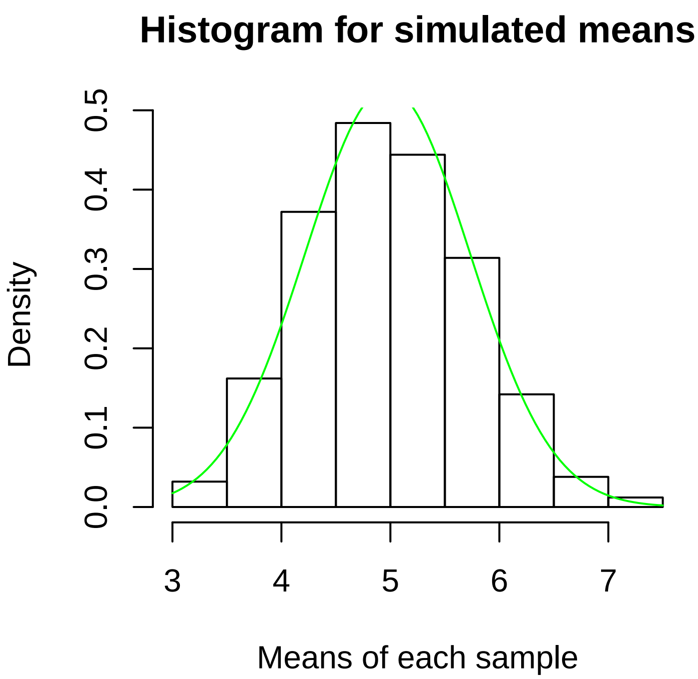
```
str(dat, 10)
```

```
##  num [1:1000] 4.97 5.75 3.32 6.42 4.67 ...
```

Data analysis

In order to study the behaviour of the averages distribution we plot the simulated dataset.

```
h = hist(dat, prob=TRUE, main = "Histogram for simulated means", xlab = "Means of each sample")
curve(dnorm(x, mean=mean(dat), sd=sd(dat)), add=TRUE, col = "green")
```



As you can see the distribution looks like a bell curve. We superimpose, futhermore a *normal* distribution (green line) with the mean and standard deviation coming from our simulated dataset.

Sample Mean versus Theoretical Mean

Using the simulated sample the mean is:

```
# sample mean
mean_dat <- mean(dat)
mean_dat
```

```
## [1] 4.974239
```

Instead, as known, the theoretical mean of the distribution is supposed to be `1/lambda`:

```
mean_theor<-1/lambda
mean_theor
```

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

The two values agree very well. Furthermore we can compare theoretical and simulated parameters using the 95% confidence interval for the averages.

```r
mean_dat + c(-1, 1)*qnorm(0.975)*sd_dat/sqrt(n)
```

```
## [1] 4.740137 5.208341
```

As you can see the theoretical value lies perfectly inside this interval.

Sample Variance versus Theoretical Variance

```r
# sample standard deviation and variance
sd_dat <- sd(dat)
sd_dat
```

```
## [1] 0.7554171
```

```r
var_dat <- var(dat)
var_dat
```

```
## [1] 0.5706551
```

```r
sd_theor <- (1/lambda)/sqrt(40)
sd_theor
```

```
## [1] 0.7905694
```

```r
var_theor <- sd_theor^2
var_theor
```

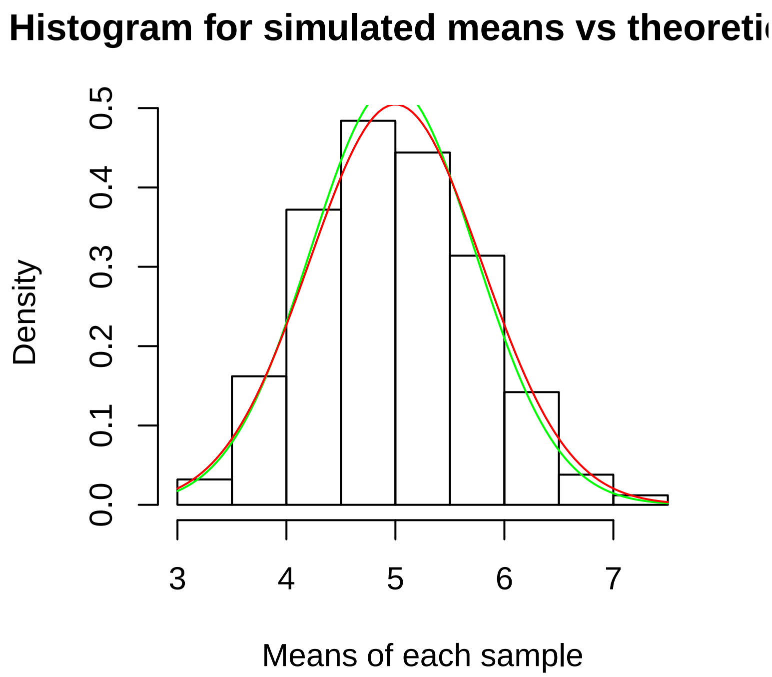
```
## [1] 0.625
```

Also in this case sample and theorethical values agree very well.

Comparison with Guassian distribution

In order to compare theoretical and simulated distribution we superimpose to the previous plot a *normal* distribution with theoretical parameters (red line).

```r
h = hist(dat, prob=TRUE, main = "Histogram for simulated means vs theoreticals", xlab = "Means of each s
curve(dnorm(x, mean=mean(dat), sd=sd(dat)), add=TRUE, col = "green")
curve(dnorm(x, mean=mean_theor, sd=sd_theor), add=TRUE, col = "red")
```



As you can see, the two distributions agree very well confirming the distribution is approximately normal.

Conclusions

We studied the distribution of the averages of 1000 samples of 40 exponentials, by simulation. we found this distribution looks like a bell curve with sample parameters, mean and variance, agree very well with theorethical values.