Central Limit Theorem vs Exponential Distribution study

Daniel Rodrigues Ambrosio

June 20th, 2015

1. Overview

This is the part 1 of the Project for the Statistical Inference course in Data Science Specialization track from Coursera.

The goal of this assignment is to investigate the exponential distribution in R and compare it with the Central Limit Theorem and illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. The study shall:

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.

1.1 Basis for the study

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. We sill set lambda = 0.2 for all of the simulations and investigate the distribution of averages of 40 exponentials.

1.2 Environment

Being able to reproduce every step of a data analysis is a crucial aspect of the data science. That being said, all the libraries used as support for this analysis are listed below and so is the system information.

```
library(ggplot2)
```

2. Simulations

The snippet below will simulate the data with a thousand simulations of 40 exponentials.

```
    # variables to control the simulation
    numSim <- 1000;     # number of simulations
    n <- 40;            # number of exponentials
    lambda <- 0.2;      # lambda used for rexp

    ## create a data matrix with the simulations
    set.seed(1928737)  # set the seed for reproducibility
    data <- matrix(rexp(numSim * n, rate=lambda), numSim);

    ## for each simulation calculate the mean
    data.means <- apply(data, 1, mean);
```

3. Sample Mean vs Theoretical Mean

The expected mean u of a exponential distribution of rate lambda is u=1/lambda

```
u <- 1/lambda
u
```

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

Let X^{-} be the average sample mean of 1000 simulations of 40 randomly sampled exponential distributions.

```
meanOfMeans <- mean(data.means)
meanOfMeans
```

```
## [1] 4.992711
```

This shows that the expected mean of the exponential ditribution and the average sample mean of a randomly sample exponential distribution are very close.

4. Sample Variance versus Theoretical Variance

The expected standard deviation v of a exponential distribution of rate lambada is v=(1/lambda)/sqrt(n)

```
stdDev <- (1/lambda)/sqrt(n)
stdDev
```

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

The variance Var of standard deviation v is Var = v ˆ 2

```
variance <- stdDev ^ 2
variance
```

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

Let Varx be the variance of the average sample mean of 1000 simulations of 40 randomly sampled exponential distribution, and vx the corresponding standard deviation.

```
stdDev_x <- sd(data.means)
stdDev_x
```

```
## [1] 0.7947586
```

```
variance_x <- var(data.means)
variance_x
```

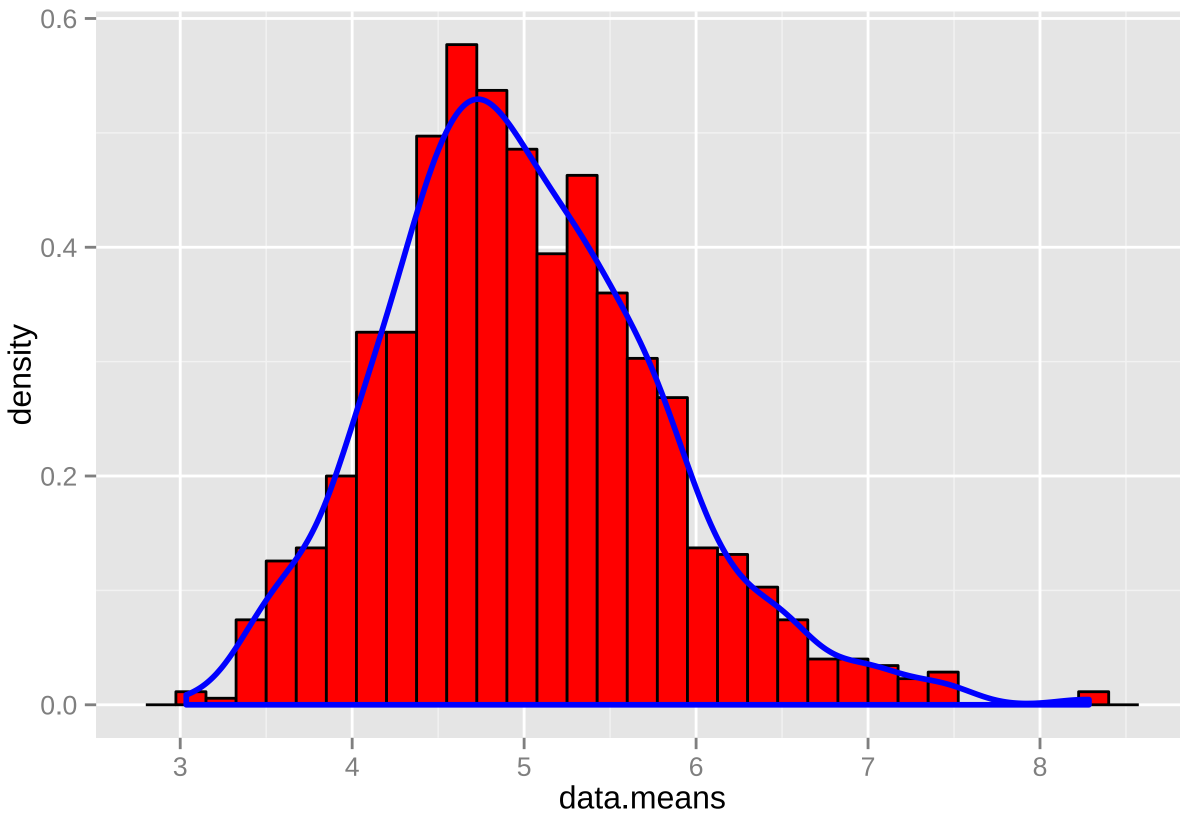
```
## [1] 0.6316413
```

This shows that the expected standard deviation and variance of the exponential ditribution and the sample standard deviation and variance of a randomly sample exponential distribution are very close. Since the variance is based on a squared value, the difference is a bit larger among the two, but still close enough.

5. Distribution

The distribution of the simulated data is very close to the normal distribution.

```
plotdata <- data.frame(data.means);
m <- ggplot(plotdata, aes(x =data.means))
m <- m + geom_histogram(aes(y=..density..), colour="black",fill = "red")
m + geom_density(colour="blue", size=1);
```



5. Appendix

1. You can find the original RPub file used to build this document on Daniel Ambrosio's repository: RPub original document