Testing the Validity of the Central Limit Theorem applied to Exponential Distribution

Vasil Yordanov aka b1ck0

15 June, 2017

Problem Definition

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$. Set $\lambda = 0.2$ for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should:

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

Introduction

According the to Central Limit Theorem (CLT) the sum of random variables (from unknown distribution with mean μ and variance σ) is also a random variable which follows the normal distribution with mean $\bar{X} = \mu$ and variance $s = \frac{\sigma}{n}$.

This is what we are going to test in this simulation

Investigation

We base our numerical experiment on the exponential distribution which has probability density function (pdf):

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

In our case $\lambda = 0.2$ which results in mean $\mu = \frac{1}{\lambda} = \frac{1}{0.2} = 5$ and standard deviation $\sigma = \frac{1}{\lambda} = \frac{1}{0.2} = 5$

Before starting the simulation we need to define some general variables:

```
library(ggplot2)    # loading the plotting library

rate = 0.2          # rate of the exponential distribution
mean = 1/rate       # theoretical mean of the exponential distribution
sd = 1/rate         # theoretical standard deviation of the exponential distribution
n = 40              # number of sample variables from each distribution
num.sim = 1000      # number of distributions
binwidth = 0.2      # parameter used to scale the width of the histogram

sim = NULL          # initializing the vector of sample varaibles
mns = NULL          # initializing the vector of sample means
sds = NULL          # initializing the vector of sample standard deviations
```

Now we will continue by simulating 1000 exponential distributions with $\lambda = 0.2$ from which we will each time take only 40 random variables and compute their mean and standard deviation:

```
for (i in 1 : num.sim) {
    set.seed(i)
    sim = rexp(n = n, rate = rate)  # drawing n samples from the exponential distribution
    mns = c(mns, mean(sim))         # appending the sample mean to the vector of sample means
    sds = c(sds, sd(sim))           # appending the sample standrad deviation to the vector of standrd
}

df = data.frame(mns,sds)            # arranging the sample means and standard deviations to a data.fram
```

One we have the data in our hands we can calculate the mean of the sample means and sample standard deviations:

```
sim.mean = mean(df$mns)  # calculating the mean of sample means
sim.sd = mean(df$sds)    # calculating the mean of sample variances
```

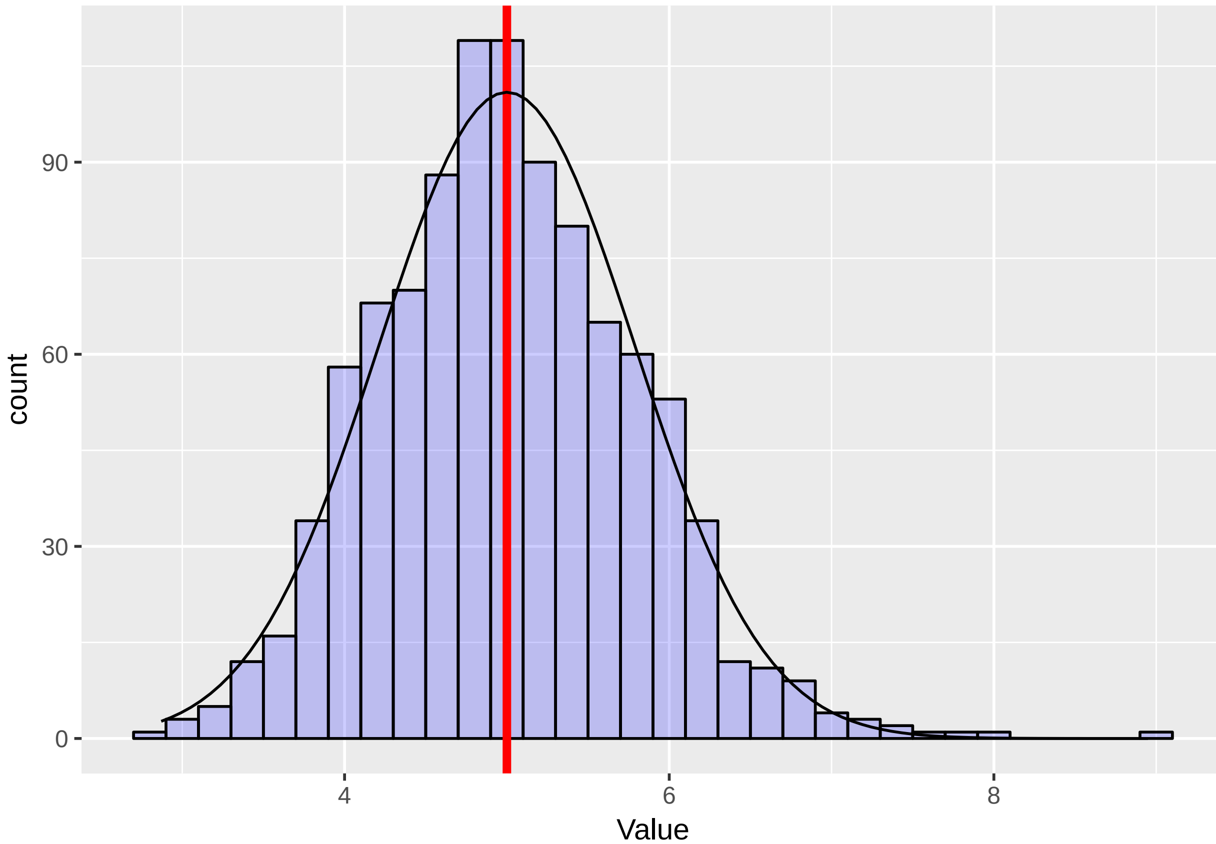
from our simulation we got sim.mean $= 5.002$ (compared with $\mu = 5$) and sim.sd $= 4.848$ (compared with $\sigma = 5$) which is pretty close.

However lets stop with all this writting and look at some graphs. Here is everything in only a single graph:

```
# plotting the variable with the simulated means
p = qplot(df$mns,
      geom="histogram",
      xlab = "Value",
      fill=I("blue"),
      col=I("black"),
      alpha=I(.2),
      binwidth=binwidth)

# plotting the theoretical mean as a vertical red line
# fitting a normal distribution N(mean, sd/sqrt(n)) to the histogram

p + geom_vline(xintercept = mean, size = 1.5, col="red") +
    stat_function(
        fun = function(x, mean, sd, n, bw){
            dnorm(x = x, mean = mean, sd = sd) * n * bw
        },
        args = c(mean = mean, sd = sd/sqrt(n), n = num.sim, bw = binwidth))
```



Some explanations. The bars represent the frequency plot of our sample means (histogram), the red line represends the theoretical mean of the exponential distribution and the black curve is representing a normal distribution with mean μ and standard deviation $\frac{\sigma}{n}$.

Conclusions

From the graph shown above we can conduct that in fact the CLT works and the distribution of sample means follows a normal distribution. As a little bonus here is a graph of the sample variances:

```
# plotting the variable with the simulated means
p = qplot(df$sds,
      geom="histogram",
      xlab = "Value",
      fill=I("green"),
      col=I("black"),
      alpha=I(.2),
      binwidth=binwidth)

# plotting the theoretical mean as a vertical red line
# fitting a normal distribution N(mean, sd/sqrt(n)) to the histogram

p + geom_vline(xintercept = mean, size = 1.5, col="red") +
    stat_function(
        fun = function(x, mean, sd, n, bw){
            dnorm(x = x, mean = mean, sd = sd) * n * bw
        },
```

```
       args = c(mean = mean, sd = sd/sqrt(n), n = num.sim, bw = binwidth))
```