Understanding the Central Limit Theorem (CLT) via Simulating the Exponential Distribution

Zhongyi Lin

Aug 16, 2015

Introduction

In this project of the Statistics Inference course on Coursera, we investigate and try to understand the Central Limit Theorem (CLT) via simulating the exponential distribution. Particularly we show that when the sample size is large, the average and standard deviation of the sample tend to be closed to the theoretical value of the distribution, and also the distribution is centered at the theoretical mean with a theoretical standard deviation.

Parameter Preparation

The ggplot2 plotting system is used in this project. Firstly it is loaded to the the environment.

```
library(ggplot2)
```

The lambda parameter in the exponential distribution is set to 0.2 and the number of simulation is set to 1000 in the whole project.

```
lambda=0.2
nosim <- 1000
meanExp <- 1/lambda
sdExp <- 1/lambda
```

To make a reproducible investigation we set a seed for the random sampling activity. Next we randomly sample according to exponential distribution using the rexp function. Notice that we sample four groups of data with sample sizes 10, 20, 30, 40 times the simulation number. These groups will be resized into matrix of 1000 rows and 10, 20, 30, 40 columns respectively in the last step.

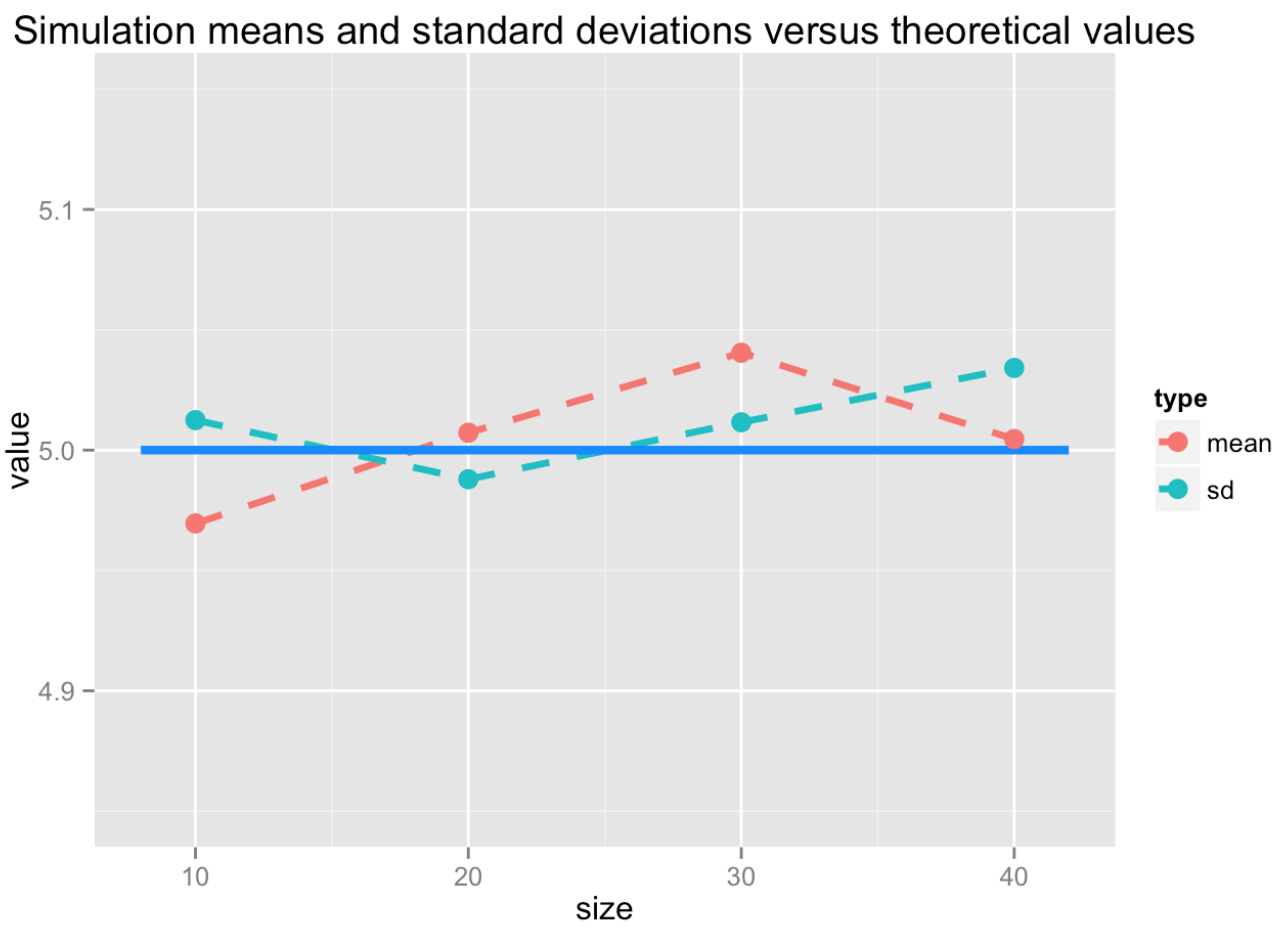
```
set.seed(3833)
rexp10 <- rexp(nosim*10, lambda)
rexp20 <- rexp(nosim*20, lambda)
rexp30 <- rexp(nosim*30, lambda)
rexp40 <- rexp(nosim*40, lambda)
```

Means and standard deviations compared with theoretical values

We calculate the means and standard deviations of the four sample groups and compare them with the theoretical mean and standard deviations, which are both 1/lambda. The result is shown in the figure plotted by the ggplot function.

```
means <- data.frame(size = c(10,20,30,40), value = c(mean(rexp10), mean(rexp2
0), mean(rexp30), mean(rexp40)), type="mean")
sds <- data.frame(size = c(10,20,30,40), value = c(sd(rexp10), sd(rexp20), sd(r
exp30), sd(rexp40)), type="sd")
meanSd <- rbind(means, sds)
meanSd$type <- as.factor(meanSd$type)

imgMeanSd <- ggplot(meanSd, aes(x=size, y=value)) +
             geom_line(aes(color=type), size=1.2, linetype="dashed") +
             geom_line(aes(x=c(8,42), y=1/lambda), size=1.5, color="#0088FF", l
inetype="solid") +
             geom_point(aes(x=size, y=value, color=type), size=3.5) +
             ylim(c(1/lambda-0.15, 1/lambda+0.15)) +
             ggtitle("Simulation means and standard deviations versus theoretic
al values")
imgMeanSd
```

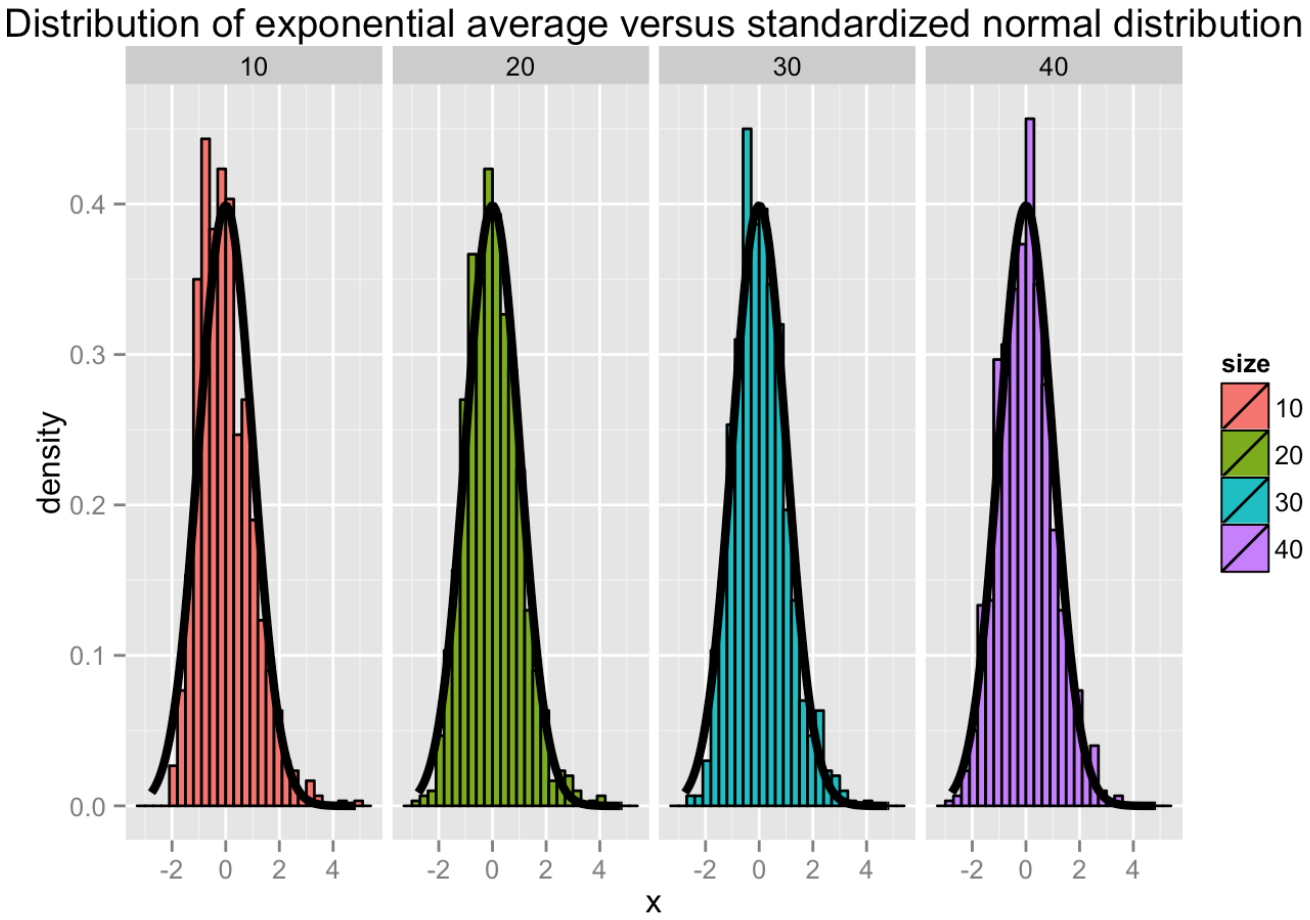


The blue solid line in the figure represents both the theoretical values of mean and standard deviation. It can be observed that with a large sample size both the mean and the standard deviation tend to be closed to the theoretical values, which is stated by CLT.

Is the distribution normal?

Next we will prove that with a large sample size the distribution will tend to be a normal distribution rather than the original exponential one. As introduced before the four sample groups are resized. Then we use the apply function on these four matrices and fit every row of them to a standardized normal distribution. Thus each calculation has a sample size of 10, 20, 30, and 40 respectively. The result is plotted and compared using facet_grid function of the ggplot2 package.

```
cfunc <- function(x, n) sqrt(n) * (mean(x) - meanExp) / sdExp
dat <- data.frame(
  x = c(apply(matrix(rexp10, nosim), 1, cfunc, 10),
        apply(matrix(rexp20, nosim), 1, cfunc, 20),
        apply(matrix(rexp30, nosim), 1, cfunc, 30),
        apply(matrix(rexp40, nosim), 1, cfunc, 40)
        ),
  size = factor(rep(c(10, 20, 30, 40), rep(nosim, 4))))
g <- ggplot(dat, aes(x = x, fill = size)) +
     geom_histogram(binwidth=.3, colour = "black", aes(y = ..density..)) +
     stat_function(fun = dnorm, size = 1.5) +
     facet_grid(. ~ size) +
     ggtitle("Distribution of exponential average versus standardized normal di
stribution")
g
```



In each subplot the standardized normal distribution functions are shown in bold black line. It can be observed that histograms of the distribution fit the normal ones very well, which means that with a large sample size (1000 in this example) the distribution tend to become a normal one rather than the original exponential one.