Statistical Inference Course Project 1: Exploring the Central Limit Theorem with the Exponential Distribution

J. Ramos

Saturday, July 25, 2015

Overview

This project aims to apply the Central Limit Theorem to the exponential distribution and prove that the means and variances of several samples obtained randomly from it approach a normal distribution, thus effectively adhering to the CLT. The contents are as follows:

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.

Simulations

The data for our proof will come from simulations. We will obtain 1000 samples of 40 observations extracted randomly from an exponential distribution with mean and standard deviation of `1/lambda`. Random observations will be extracted with the `rexp` function.

```r
# First and foremost, set seed to make this code reproducible
set.seed(230779)

# Then load relevant packages
library(dplyr)
library(ggplot2)

# Set up lambda, number of simulations, observations per simulation
lambda <- 0.2
n <- 40
sim <- 1000

# Set theoretical mean and variance according to the given lambda
theomean <- 1/lambda
theovar <- (1/lambda)^2/n

# Create dataframe with simulations
df <- as.data.frame(matrix(rexp(sim*n,lambda), sim, n))
# Create 'avg' column with mean of each row
df <- df %>% mutate(avg=rowMeans(df))

# Calculate realized mean and variance
realmean <- mean(df$avg)
realvar <- var(df$avg)
```

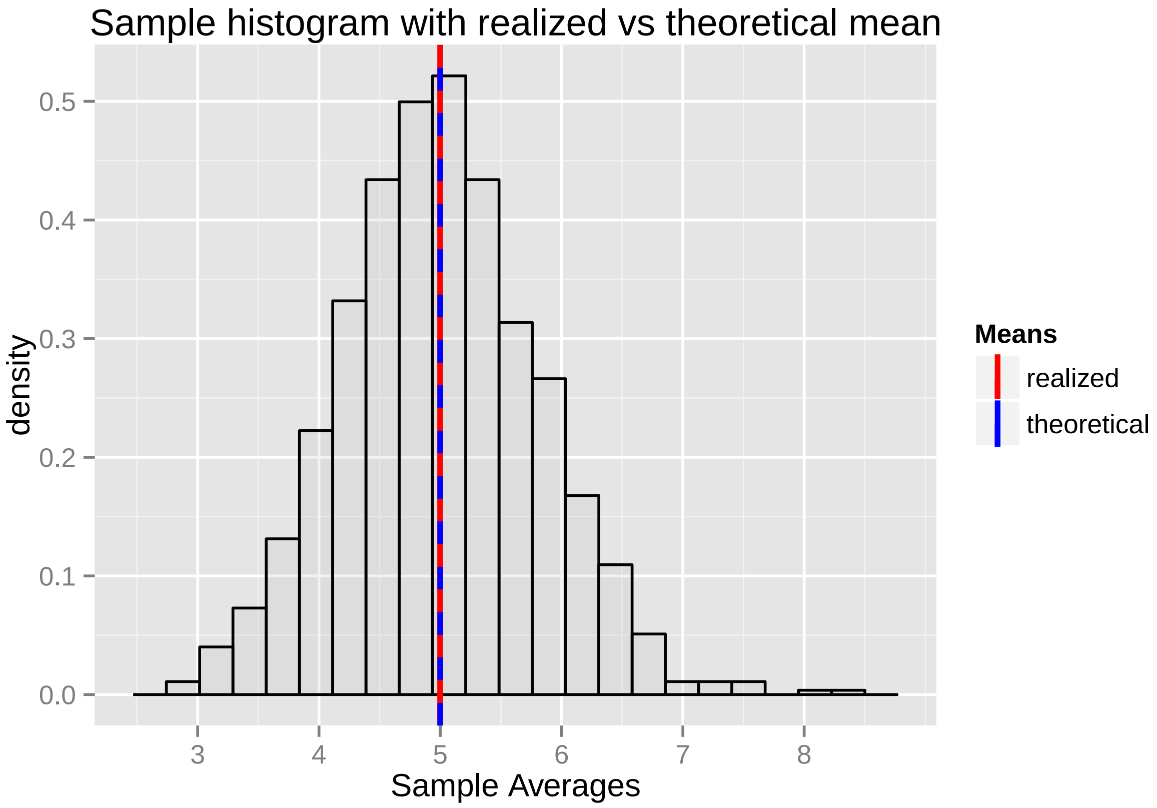
We have created a data frame from a *1000 x 40* matrix, and have added an extra column with the mean for each sample in the rows. We can now compare the theoretical mean with the realized mean from the 1000 samples.

Sample Mean vs Theoretical Mean

To assess the difference (or, hopefully, similarity) between the theoretical mean of *5* and the realized mean of *4.9988764*, we will plot a histogram and overlay a normal distribution curve, and lines at the theoretical and realized means.

```
# Histogram
p <- ggplot(aes(x=df$avg), data=df) +
    geom_histogram(aes(y= ..density..), binwidth=(max(df$avg)-min(df$avg))/20,
                   fill=I("gray"), col=I("black"), alpha=I(.2)) +
    ggtitle('Sample histogram with realized vs theoretical mean') +
    scale_x_continuous(breaks=c(0:8)) +
    xlab('Sample Averages') +
    geom_vline(aes(xintercept=realmean, color='realized'), linetype=1,
               size=1, show_guide = T) +
    geom_vline(aes(xintercept=theomean, color='theoretical'), linetype=2,
               size=1, show_guide = T) +
    scale_colour_manual(name="Means",
                        values=c('theoretical'='blue','realized'='red'))

print(p)
```



We can surmise from the plot that both the realized means from the samples and the theoretical mean are very close (indeed, the realized mean is only *0.001* less than the theoretical mean), so the principle of the CLT of

> [. . .] the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, **regardless of the underlying distribution.**

holds even for the exponential distribution.

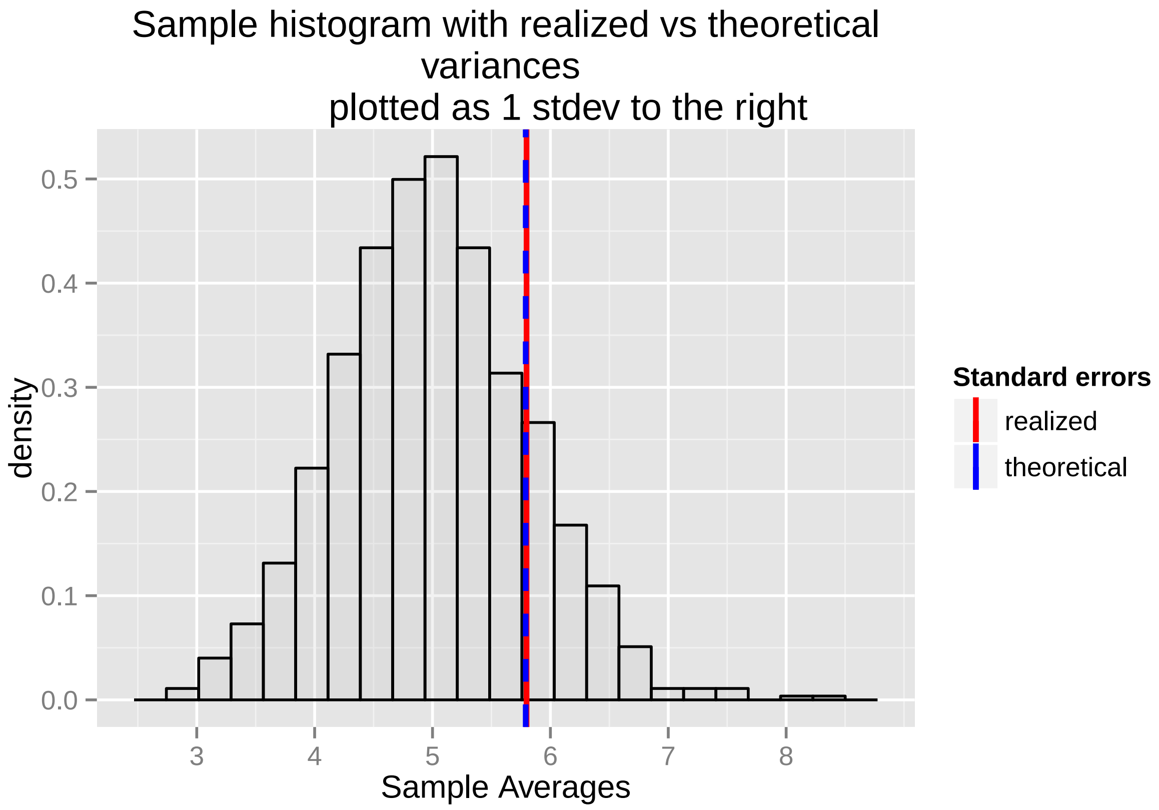
Sample Variance vs Theoretical Variance

To further proof the CLT not only in terms of means but of variances as well, we will plot *1 standard error of the mean* to the right along with *1 theoretical standard deviation* and examine if they fall in the same point of the distribution.

For this exercise, we will need to get the square root of both the theoretical variance of *0.625* and the realized var of *0.6381557*.

```r
# Histogram
r <- ggplot(aes(x=df$avg), data=df) +
    geom_histogram(aes(y= ..density..),
                   binwidth=(max(df$avg)-min(df$avg))/20,
                   fill=I("gray"), col=I("black"), alpha=I(.2)) +
    ggtitle('Sample histogram with realized vs theoretical\nvariances
            plotted as 1 stdev to the right') +
    scale_x_continuous(breaks=c(0:8)) + xlab('Sample Averages') +
    geom_vline(aes(xintercept=realmean+sqrt(realvar),
                   color='realized'), linetype=1, size=1, show_guide = T) +
    geom_vline(aes(xintercept=theomean+sqrt(theovar), color='theoretical'),
               linetype=2, size=1, show_guide = T) +
    scale_colour_manual(name="Standard errors",
                        values=c('theoretical'='blue','realized'='red'))

print(r)
```



Indeed, the difference between both variances is just *-0.013*, and it is virtually impossible to distinguish when plotted as standard deviation in the histogram, serving as further proof of the CLT.

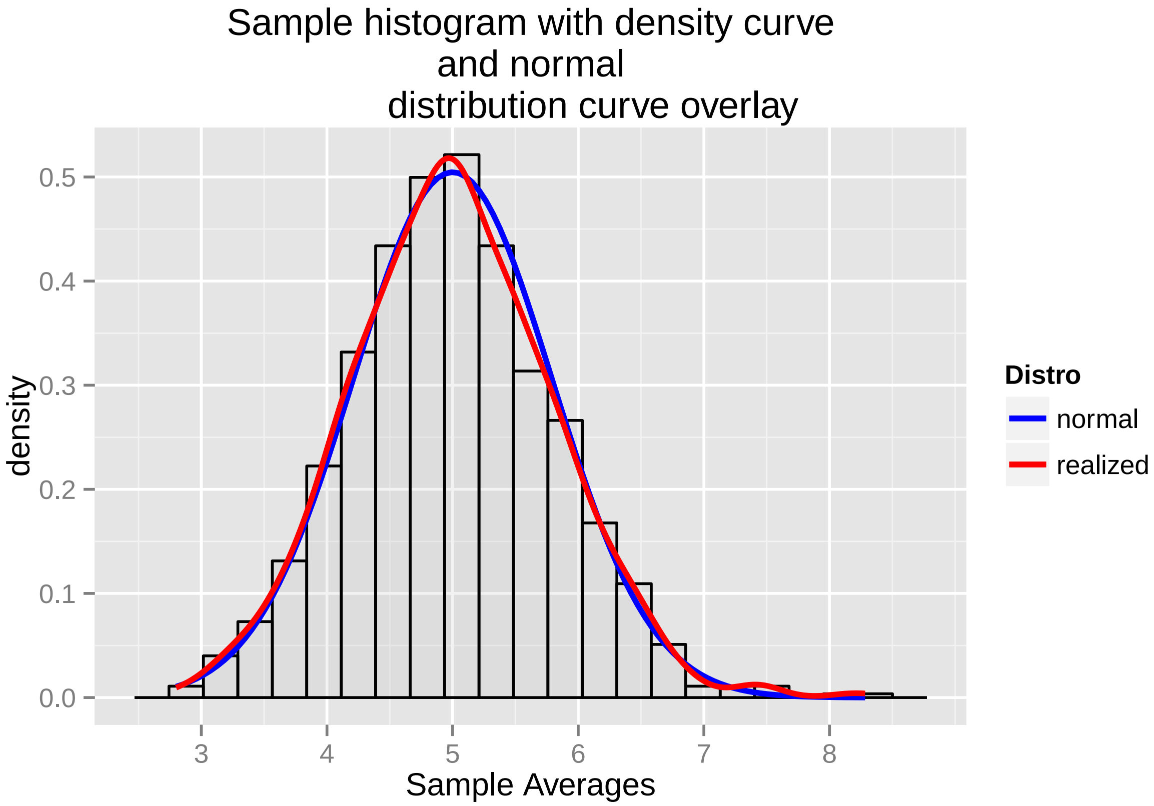
Distribution

Finally, when plotting the histogram of the realized samples, and overlaying a plot of a normal distribution following ~**N(5, 0.791)**, together with a plot of the density curve we clearly observe that, save for some outliers and the maximum density, the curves are virtually the same.

```r
# Histogram
q <- ggplot(aes(x=df$avg), data=df) +
    geom_histogram(aes(y= ..density..), binwidth=(max(df$avg)-min(df$avg))/20,
                   fill=I("gray"), col=I("black"), alpha=I(.2)) +
    ggtitle('Sample histogram with density curve\n and normal
            distribution curve overlay') +
    scale_x_continuous(breaks=c(0:8)) + xlab('Sample Averages') +
    stat_function(aes(color='normal'), geom='line', fun=dnorm, size=1,
                  arg=list(mean=theomean, sd=sqrt(theovar)), show_guide=T) +
    stat_density(aes(color='realized'), geom='line', size=1,
                  show_guide=T) +
    scale_colour_manual(name="Distro",
                        values=c('normal'='blue','realized'='red'))

print(q)
```

```
## ymax not defined: adjusting position using y instead
```



Conclusion

The CLT is one of the most powerful inferential tools in statistics, for it allows us to look past strenous details and features of the data and get directly to the values that matter.