Peer-Graded Assignment Part 1: Simulation Exercise

Rogelio Caballero

May 26, 2017

Abstract

As part of the assignment for the Statistical Inference Course from John's Hopkins University at Coursera, we simulate numbers from an exponential distribution in *batches*, take the mean in each batch to compare the features of the distribution of the means to those of the original distribution.

Seed and Definitions

First we are going to set a seed in order to make our results reproducible:

```
set.seed(1)
```

The following define the size of the batches `n` over which we are going to take the mean, the number of batches `nosim` that will conform the distribution of the means and the parameter of the exponential distribution that will generate the numbers `lambda`.

```
n <- 40
nosim <- 1000
lambda <- 0.2
```

The inverse of the parameter `lambda` of the exponential distribution is equal to the mean and the standard deviation of the distribution (more info here). This is important because we are interested in comparing the mean and the standard deviation of the distribution of the means with the original distribution.

Now we are going to generate the random numbers and store them in a 1000x40 matrix whose rows represent each one of the batches over which we are going to take the mean:

```
m <- matrix(rexp(nosim*n, rate = lambda), nosim)
mn <- apply(m, 1, mean)
```

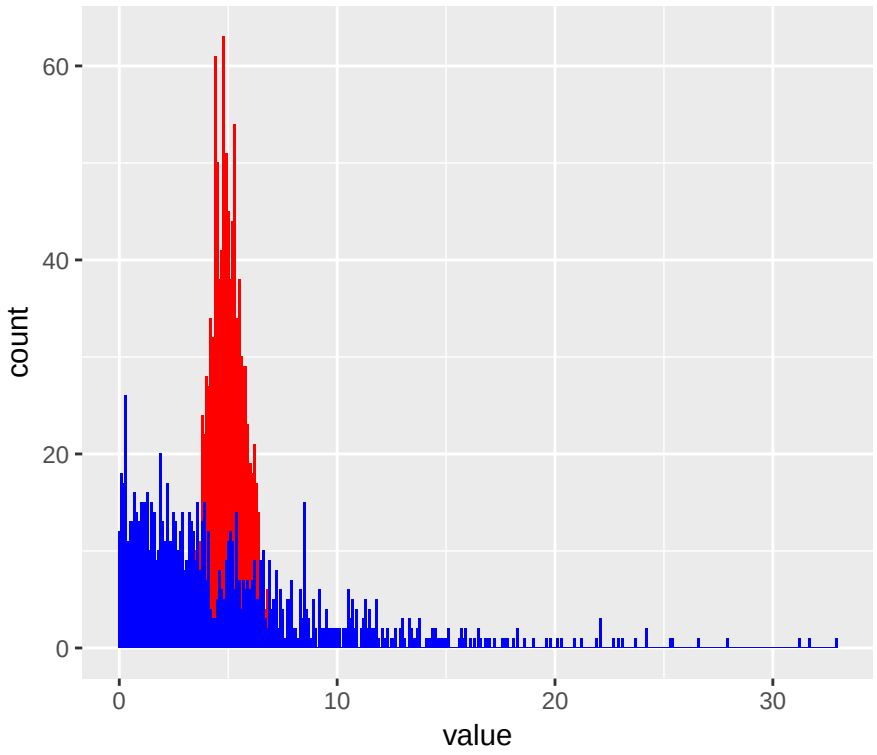
Distribution of the means

The *Central Limit Theorem* tells us that if we sample a distribution in groups of size `n` (what we've called batches) and take the mean in each group, as `n` approaches infinity, the distribution of the means approaches a normal distribution.

We illustrate this result with the exponential distribution, but first we need to arrange in a dataframe the means that we stored in `mn`.

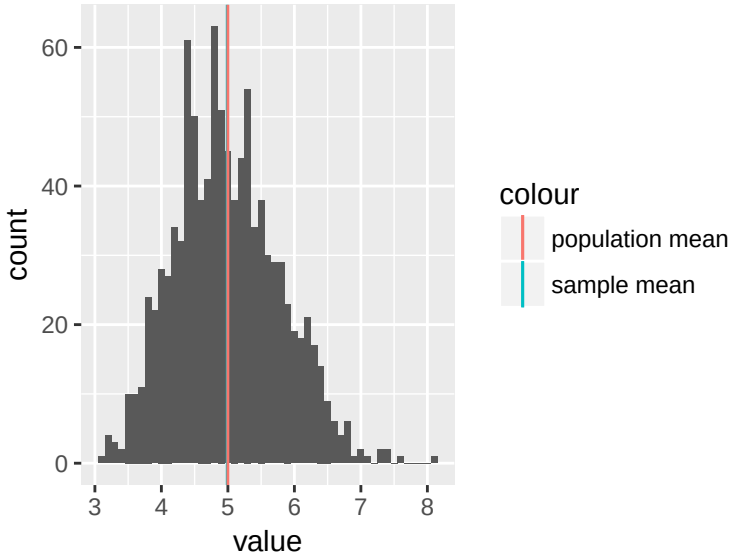
```
df <- data.frame(mean = mn, exponential = m[,1])
library(reshape2)
df <- melt(df)
```

The dataframe defined above contains numbers from the exponential distribution and the means of the batches. We see in the following figure that they follow different distributions, the red one being normal (sample means) and the blue one being exponential (population).



The other part of the *Central Limit Theorem* says that the sample mean, the mean of the distribution of means, is equal to the population mean, the inverse of `lambda` in this case. The following plot illustrates that:

```
ggplot(df, aes(x = value)) +
        geom_histogram(data = subset(df, df$variable == "mean"), binwidth = 0.1) +
        geom_vline(aes(xintercept = mean(mn), colour = "sample mean")) +
        geom_vline(aes(xintercept = 1/lambda, color = "population mean"))
```

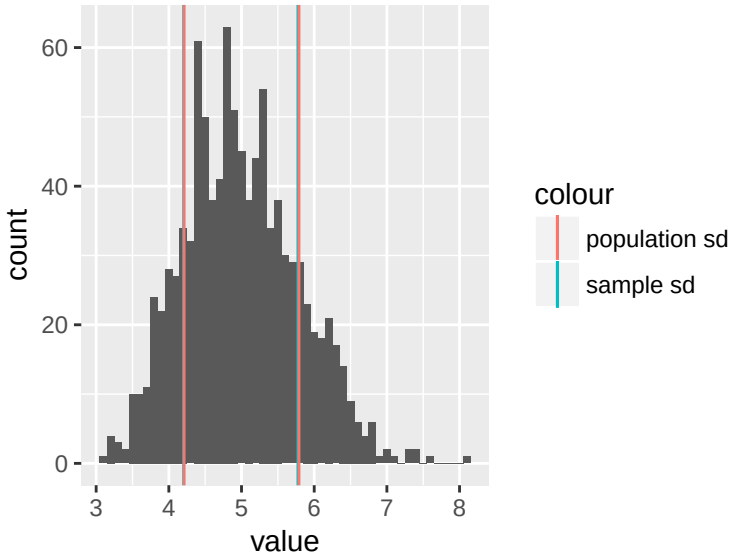


The Central Limit Theorem also says something about the relationship between the standard deviation of the sample Σ and the standard deviation of the population σ:

$$\Sigma = \frac{\sigma}{\sqrt{n}}$$

The following fuigure illustrates that fact. We've highlighted the interval between two standard deviations with center in the mean:

```
ggplot(df, aes(x = value)) +
        geom_histogram(data = subset(df, df$variable == "mean"), binwidth = 0.1) +
        geom_vline(aes(xintercept = mean(mn) - sd(mn), colour = "sample sd")) +
        geom_vline(aes(xintercept = mean(mn) + sd(mn), colour = "sample sd")) +
        geom_vline(aes(xintercept = 1/lambda - (1/lambda)/sqrt(n), color = "population sd")) +
        geom_vline(aes(xintercept = 1/lambda + (1/lambda)/sqrt(n), color = "population sd"))
```



As we can see the interval calculated via the population standard deviation is almost equal to the interval calculated directly from the sample.