Computational Verification of the Central Limit Theorem by Simulation of Means from an Exponential Distribution

Christopher Skyi

August 23, 2015

Part I: Computational Verification of the Central Limit Theorem

In part I of the final project for the Johns Hopkins' *Statistical Inference* class, I'll validate the Central Limit Theorem by using R to generate 1000 means from an exponential distribution (created with a random sample of 40 random deviates from rexp()).

The Simulation:

- The exponential distribution will be simulated in R with random deviates from rexp(n = 40, lambda) where lambda is the rate parameter.
- The mean & standard deviation of the exponential distribution is 1/lambda.
- I'll set lambda = 0.2 for all of the simulations.

Script Output:

Produces a distribution of 1000 sample means of 40 exponential random variables

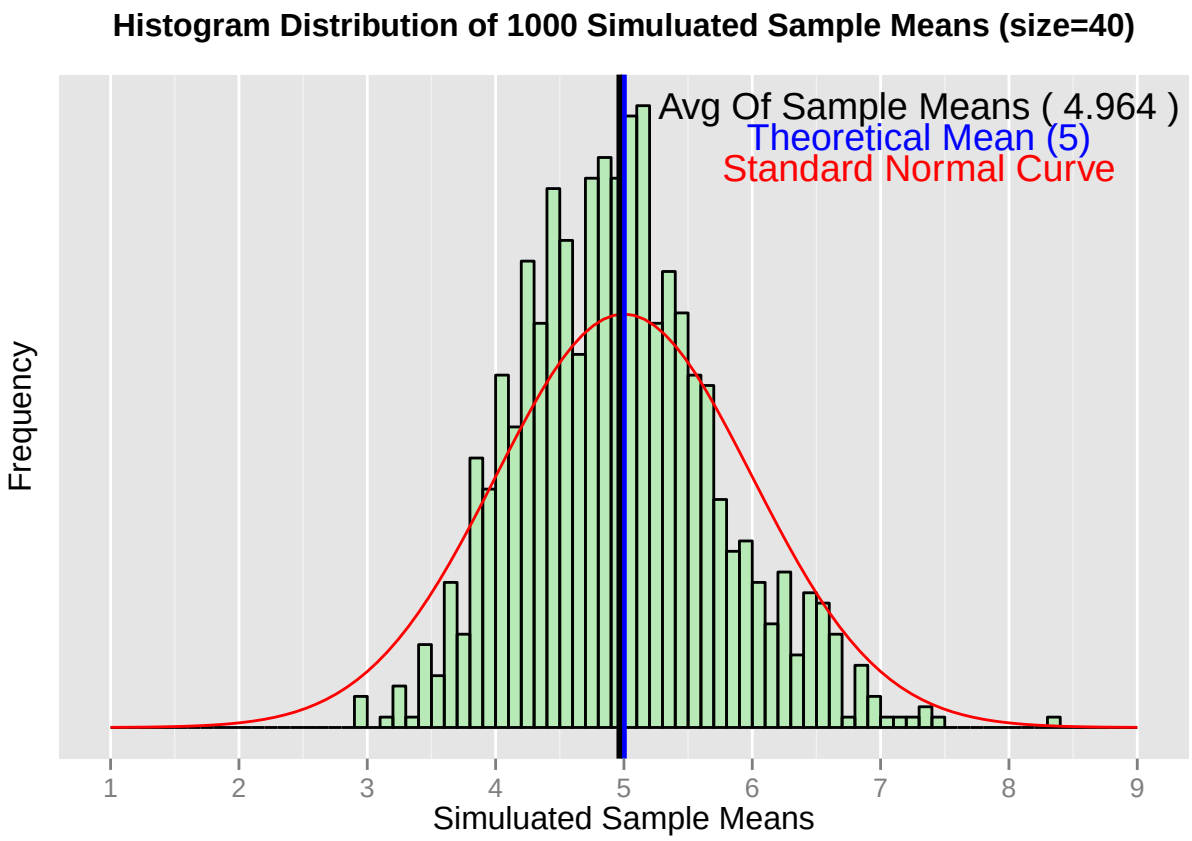
```
nosim <- 1000                  # number of samples (of size 40)
lambda = 0.2                   # rate parameter
exp.mean = exp.sd = 1/lambda   # theorectial exponential mean and standard deviation = 5
sample.size = 40               # sample size

# Simulate 1000 means from 1000 sample sizes of size 40
set.seed(20) # ensures reproducibility of the sequence of random numbers
sample.means = NULL
for (i in 1 : nosim) sample.means = c(sample.means, mean(rexp(sample.size, lambda)))

# get the average of those 1000 means
mean.of.sample.means = sum(sample.means)/nosim
```

Graphical Output

- Shows a distribution of 1000 simulated sample means of 40 exponential random variables
- Shows where the distribution is centered (4.964), close to the theoretical mean, verifying the Law of Large Numbers
- Shows a comparison of theoretical center of the distribution (5) to it's actual center (4.964)
- Using an overlay of a standard normal (see red plot), we see that our simulated distribution is approximately normal, verifying the **Central Limit Theorem**



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Sample Variance versus Theoretical Variance

The theoretical variance of this distribution (derived from the std. error of the mean, given on p. 37 of *Statistical Inference For Data Science* by Brian Caffo, PhD) is the variance of the exponential distribution, 25, divided by the sample size of a sample mean, 40, i.e.,

$25/40 = 0.625$

The sample varianace of the above simulated distribution of means can be calculated a couple of different ways:

```
# maually calculate the sample variance:
diff.sqr <- (sample.means - mean.of.sample.means)^2
sample.var = sum(diff.sqr)/(length(diff.sqr)-1) # returns 0.5919712 (using a simulation seed of 20)

# use R's var() to caluated it
var(sample.means) # returns 0.5919712 (using a simulation seed of 20)
```

```
## [1] 0.5919712
```

If we round our sample variance to one decimal place, it completely agrees with the theoretical Variance:

```r
round(sample.var,1) # returns 0.6
```

```
## [1] 0.6
```

```r
round(25/40,1)      # returns 0.6
```

```
## [1] 0.6
```

Summary

So, in this simulation, we simulated 1000 means of 40 exponentals with lambda = 0.2. Our theory says the variance of averages of 40 standard normals must be the theoretical variance (25) divided by the sample size, 40 (see p. 37 in *Statistical Inference For Data Science* by Brian Caffo, PhD)). Taking the variance of the 1000 means yields nearly exactly that, 0.5919712. Note that it's only close, 0.5919712 versus 0.625. To get it to be exact, we'd have to simulate infinitely many means.