An Exercise in the Central Limit Theorem

Tim Wise

May 2015

Overview

This is an exercise in the Central Limit Theorem (CLT) where we investigate the distribution of the average of exponential numbers and compare:

- The sample mean to the theoretical mean
- The sample variance to the theoretical variance, and
- The sample distribution to the normal distribution

This exercise was a project for the Coursea class *Introduction to Statistical Inference* which is part the *Data Science Specialization* from Johns Hopkins University.

Note to graders:

Since this was an exercise and a demonstration, rather than a formal paper or executive report, I chose to write this report in a narrative style interweaving discussion, code, and results. I think it is a good style for this exercise and let me practice the skills learned in the Reproducible Research class. The report is less than 6 pages, as per the rubric, but the code and details are not separated into an appendix. Grade it how you wish. Thanks.

Process

For our purposes, the important points of the CLT, as summarized from the class notes, are as follows:

- Consider a population of random numbers from any distribution with mean μ and variance σ^2.
- The average of n random draws from this population is itself a random variable. This variable is commonly symbolized as $\bar{X}_n$.
- The mean of that random variable is the population mean, that is $E[\bar{X}_n] = \mu$.
- The variance of that random variable is $Var(\bar{X}_n) = \sigma^2/n$.
- The distribution of $\bar{X}_n$ is approximately standard normal, $N(\mu, \sigma^2/n)$.

In this exercise, we demonstrate those points of the CLT by doing the following:

- We start with a population of random numbers from the exponential distribution.
- We create a new random variable that is the average of 40 random exponential numbers. In CLT terms, this random variable would be named $\bar{X}_{40}$, but we will just call it Z.
- We generate 1000 values for Z.
- We look at the distribution of Z and compute its sample mean and variance.
- We compare the sample mean and variance to the theoretical values given to us by the CLT.
- We show that Z is normally distributed.

Simulations

First, let's create a random variable, Z, whose value is the average of 40 numbers drawn from an exponential distribution. We generate 1000 values for Z. It is the distribution of these values that we are interested in.

The exponential distribution

The exponential distribution is simulated in R with *rexp(n, lambda)* function where n is the number of exponential numbers to generate and $lambda$ is the rate parameter (e.g., 35 web hits per minute). The mean and standard deviation of the exponential distribution are both $1/lambda$.

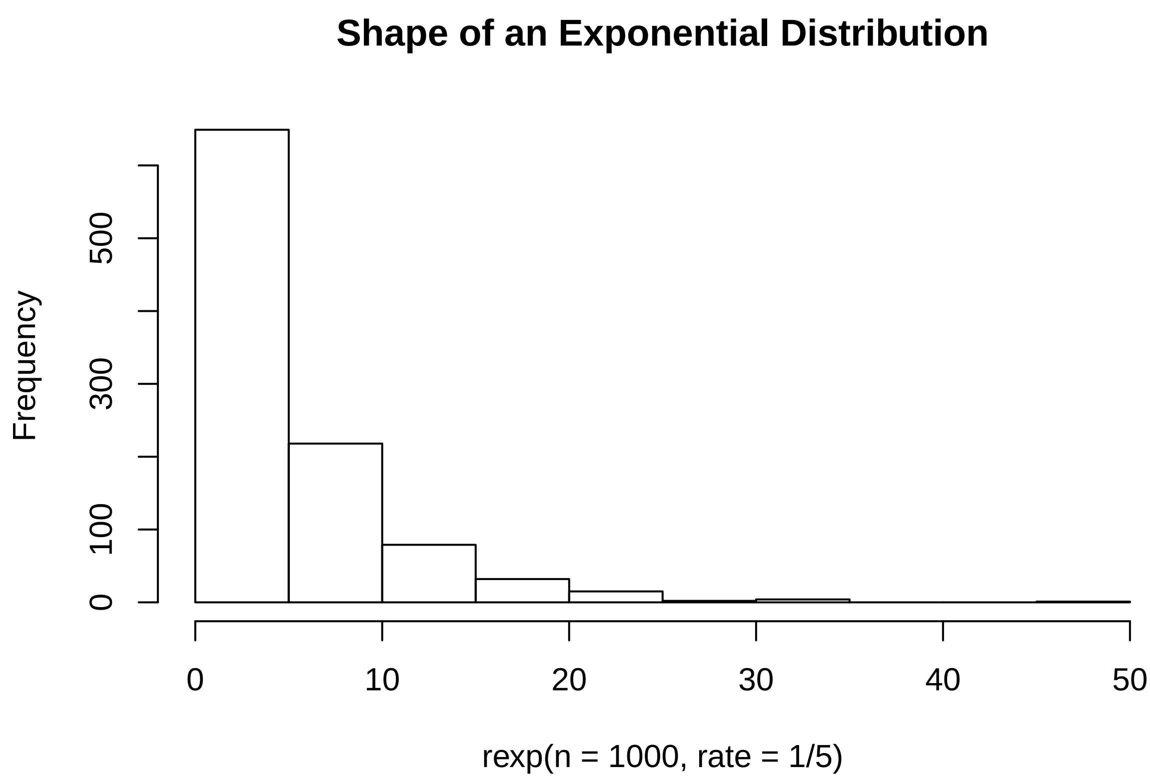
For this analysis, we'll use $lambda = 0.2$. Let's set $lambda$ and compute the mean, mu, and the standard deviation, $sigma$:

```
lambda <- 0.2
mu <- 1 / lambda
sigma <- 1 / lambda
c(lambda, mu, sigma)
```

```
## [1] 0.2 5.0 5.0
```

Just for grins, let's look at the shape of an exponential distribution with a mean of 5:

```
hist(rexp(n=1000, rate=1/5), main="Shape of an Exponential Distribution")
```



We see it's a highly skewed distribution, with a concentration of the values between 0 and the mean and a long tail extending above the mean. It looks most un-normal.

Random variable Z, the average of 40 exponential numbers

Now, we generate the average of 40 random exponential numbers 1000 times. We store those values in a vector Z, which is our random variable of interest:

```r
# set the seed so others can reproduce our results
set.seed(123456789)
n <- 40
Z <- NULL
for (i in 1:1000) {
  Z <- c(Z, mean(rexp(n, lambda)))
}
str(Z)
```

```
##  num [1:1000] 4.43 4.4 4.6 3.25 3.95 ...
```

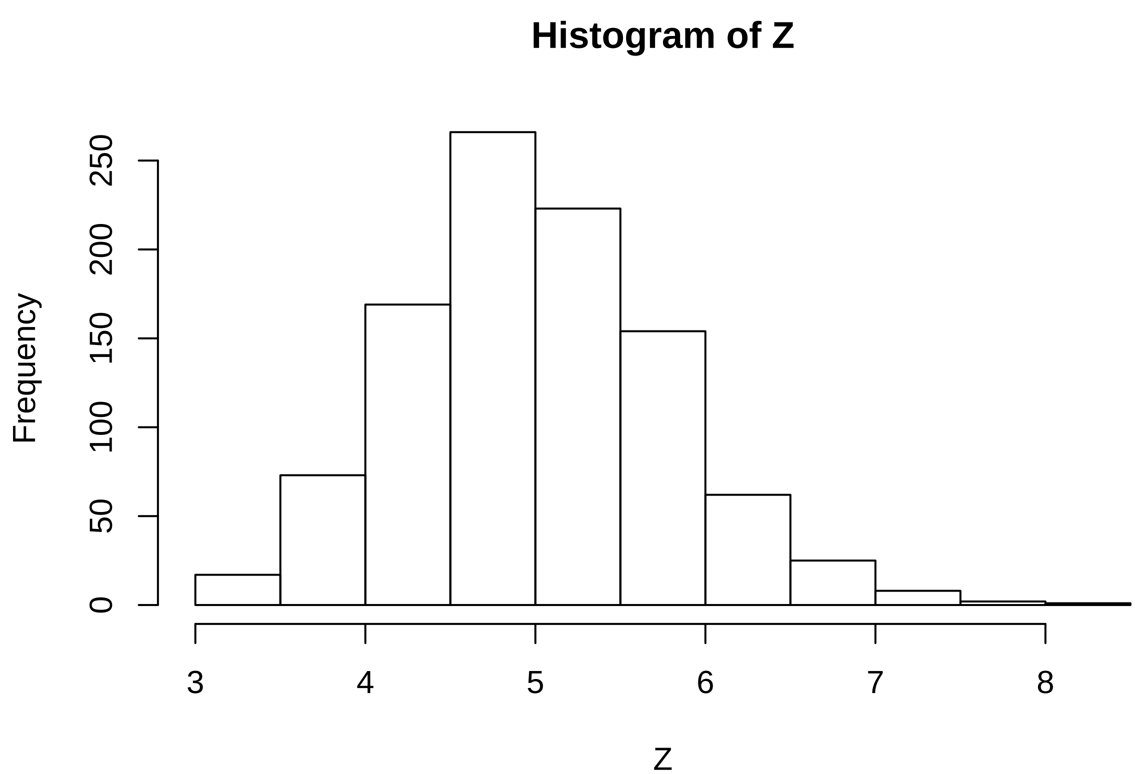
Sample Mean versus Theoretical Mean

The CLT states **the distribution of the average** of independent identically distributed (iid) variables **is normally distributed around the population mean**.

For our example, that implies the values of Z should be distributed around the mean of our exponential distribution, which is $mu = 5$. Let's see if that's true.

First, let's visualize Z, looking at the distribution of values:

```r
hist(Z)
```



The distribution of Z, shown in the histogram above, certainly looks centered around 5.

Now let's compute the sample mean of Z and compare with mu:

```r
theoretical_mean <- mu
sample_mean      <- mean(Z)

round(c(theoretical_mean, sample_mean), 2)
```

```
## [1] 5.00 5.01
```

We can see that the two compare very closely, supporting the CLT claim.

Sample Variance versus Theoretical Variance

The CLT says the theoretical variance of Z is σ^2/n, where $sigma$ is the standard deviation of our exponential distribution and n is the number of exponential numbers we draw for each set.

Let's compare the theoretical variance of Z to the sample variance:

```r
theoretical_variance <- sigma^2/n
sample_variance      <- var(Z)

round(c(theoretical_variance, sample_variance), 2)
```

```
## [1] 0.62 0.60
```

These two compare closely also, supporting the CLT claim.

Distribution of Z compared to Standard Normal

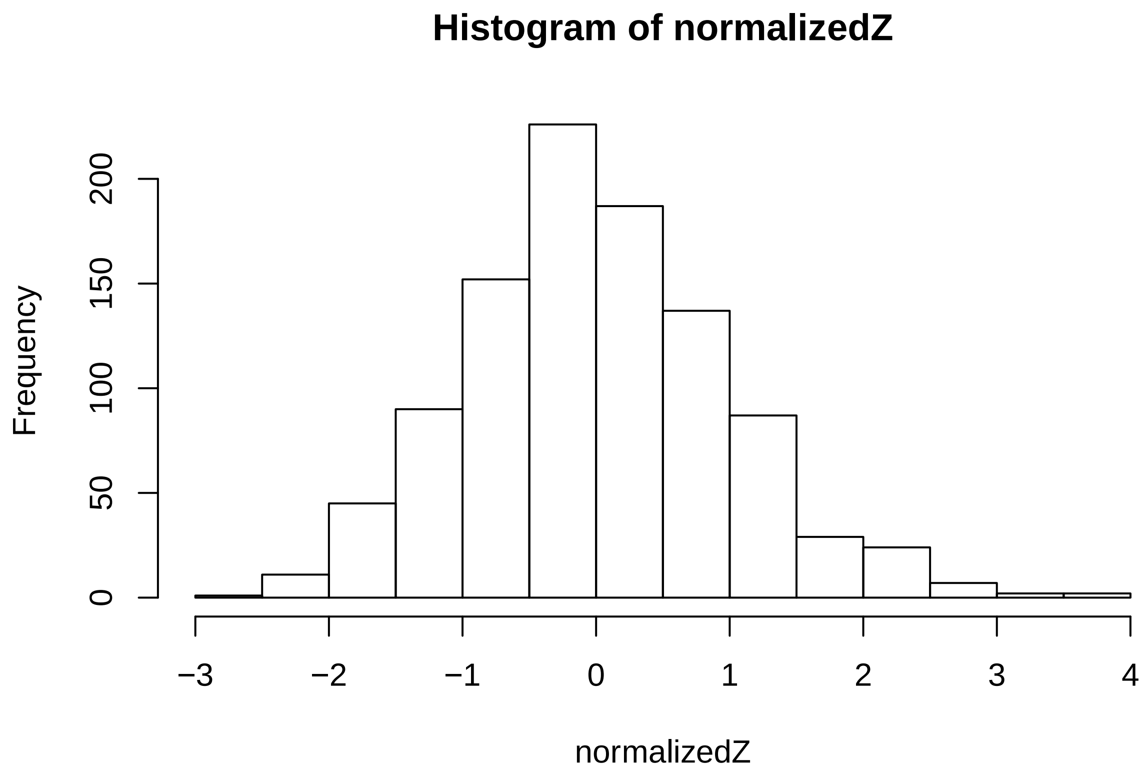
Finally, the CLT states that, properly normalized, the distribution of Z is standard normal. A **standard normal distribution** has a mean and median of 0 and a standard deviation of 1. Let's standardize Z and see if it compares favorably to the standard normal measures.

We normalize the values of Z by subtracting mu, the theoretical mean of Z, and dividing by the theoretical standard deviation of Z (the square root of the variance of Z):

$$\frac{Z - \mu}{\sigma/\sqrt{n}}$$

Let's normalize the Z values and plot their distribution:

```r
normalizedZ <- (Z - mu)/(sigma / sqrt(n))
hist(normalizedZ)
```



The values certainly look normally distributed, centered around 0.

Now, let's compute the mean, median, and standard deviation of the normalized Z values and compare them to the standard normal measures. Again, the standard normal distribution has these measures:

```
standard_mean   <- 0
standard_median <- 0
standard_dev    <- 1

round(c(standard_mean, standard_median, standard_dev), 4)
```

```
## [1] 0 0 1
```

Computing the measures for our normalized Z values:

```
normalizedZ_mean   <- mean(normalizedZ)
normalizedZ_median <- median(normalizedZ)
normalizedZ_dev    <- sd(normalizedZ)

round(c(normalizedZ_mean, normalizedZ_median, normalizedZ_dev), 4)
```

```
## [1]  0.0069 -0.0571  0.9806
```

We see the mean and deviation of the normalized Z are comparable (<2 %) to those of the standard normal distribution, however, the sample median is suspicious, off by almost 6%. So we conclude that Z is *approximately* normally distributed.

Summary

In this exercise, we did a simulation to demonstrate the principles of the Central Limit Theorem. For a sample distribution of the average of 40 exponential numbers, we showed that:

- The sample mean was comparable to the theoretical mean
- The sample variance was comparable to the theoretical variance, and
- The sample values are approximately normally distributed