

An Exercise in the Central Limit Theorem

Tim Wise

February 21, 2015

Overview

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

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 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 , its sample mean and variance
- We compare those to the theoretical values given to us by the CLT
- We show that Z is normally distributed

Simulations

In this section, we 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 $1/lambda$.

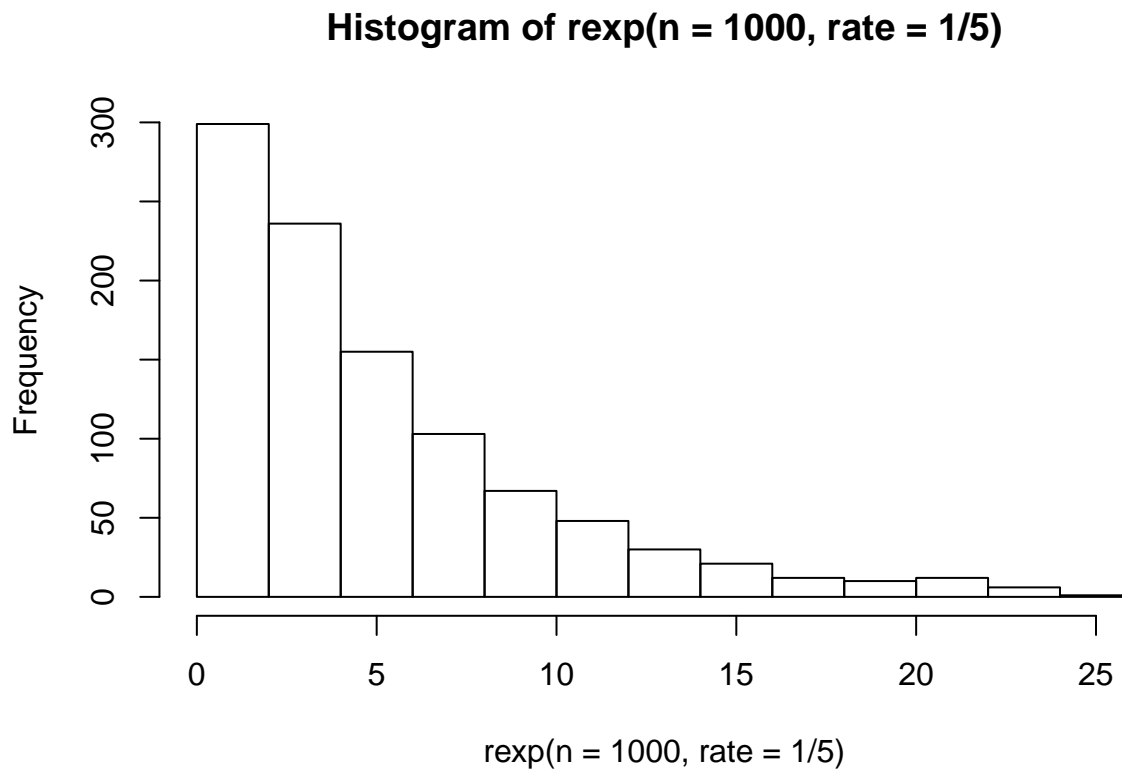
For this analysis, we'll use $lambda = 0.2$. Let's set $lambda$ and compute the mean, mu , and the standard deviation, $sigma$:

```
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))
```



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:

```
# 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 ...
```

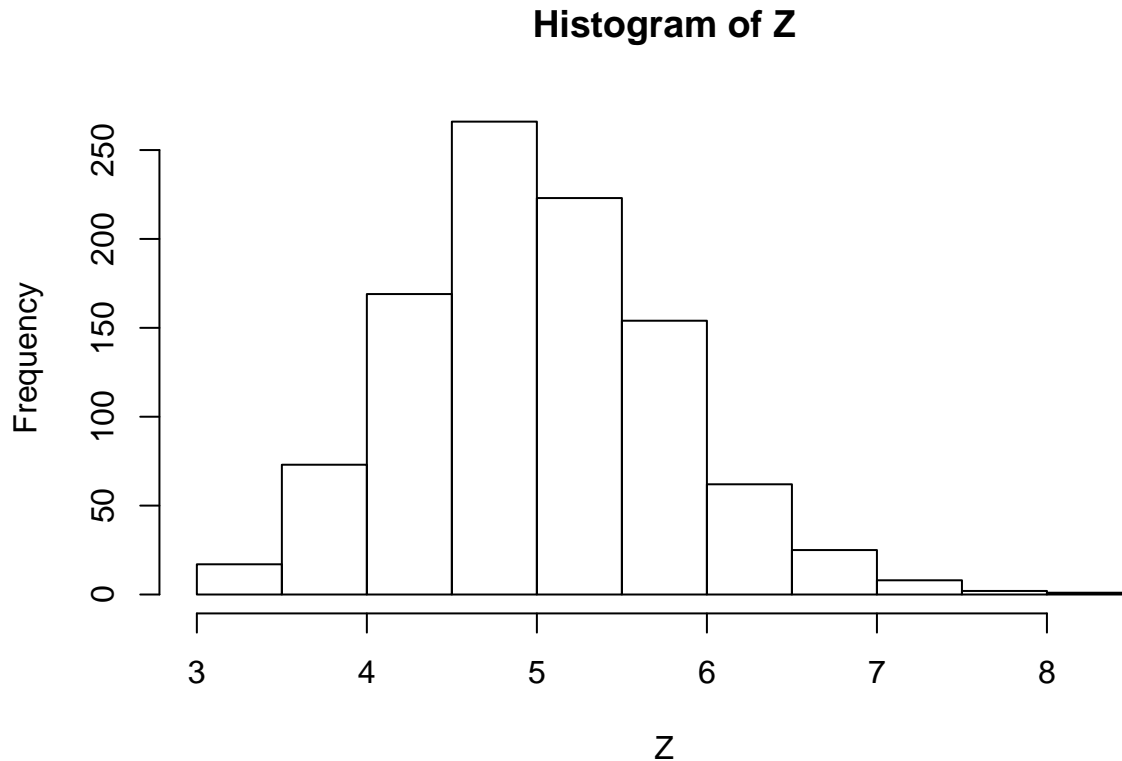
Sample Mean versus Theoretical Mean

The CLT states **the distribution of averages** 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, $\mu = 5$. Let's see if that's true.

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

```
hist(Z)
```



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

The CLT says that the theoretical mean of Z is μ . So, let's compute the sample mean of Z and compare with μ :

```
round(c(mu, mean(Z)), 2)
```

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

We can see that the two compare very closely.

Sample Variance versus Theoretical Variance

The CLT says the theoretical variance of Z is σ^2/n .

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

```
round(c(sigma^2/n, var(Z)), 2)
```

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

These two compare closely also.

Distribution of Z compared to Standard Normal

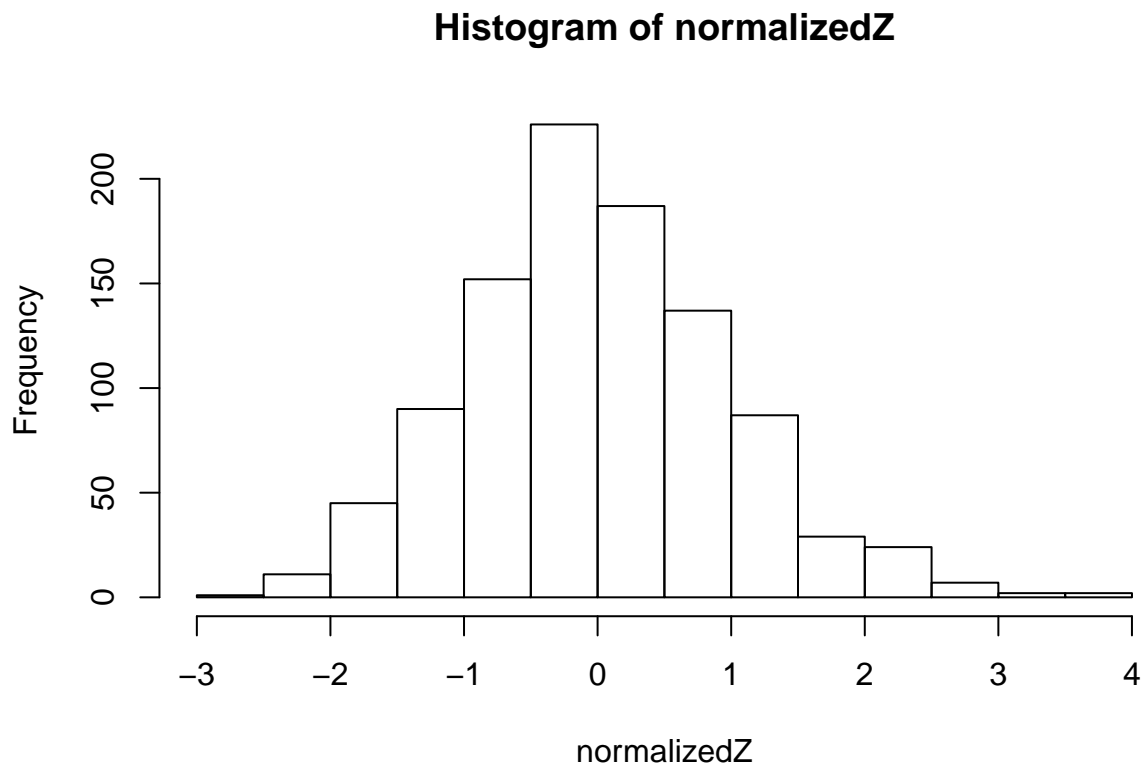
Finally, the CLT states that, properly normalized, the distribution of Z is standard normal. A standard normal distribution has a mean and median of 0 and a standard deviation of 1.

We normalize the values of Z by subtracting μ , 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 plot the distribution of normalized Z values:

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



Computing the mean, median, and standard deviation of the normalized Z values, we see they are comparable to those of the standard normal distribution: mean = 0, median = 0, standard deviation = 1:

```
round(c(mean(normalizedZ), median(normalizedZ), sd(normalizedZ)), 4)
```

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