Central Limit Theorem Simulation

Andrew Witherspoon

11/11/2018

The Central Limit Theorem is an important theorem in statistics, which states, that a distribution of means of independent independently distributed random variables, will take on the shape of a normal distribution if our sample size big enough. This is extremely useful, as so many of our statistical tools rely on the assumption of normality.

Let's start by looking at an exponential distribution, which is decidedly not a normal shape. Here is a look at the shape of exponential distribution, with a rate (lambda) of 0.2:

```
x <- seq(0, 50, length.out=1000)
par(mar=c(2, 2, .75, .1))
plot(x, dexp(x, rate=0.2), type ="l")
```
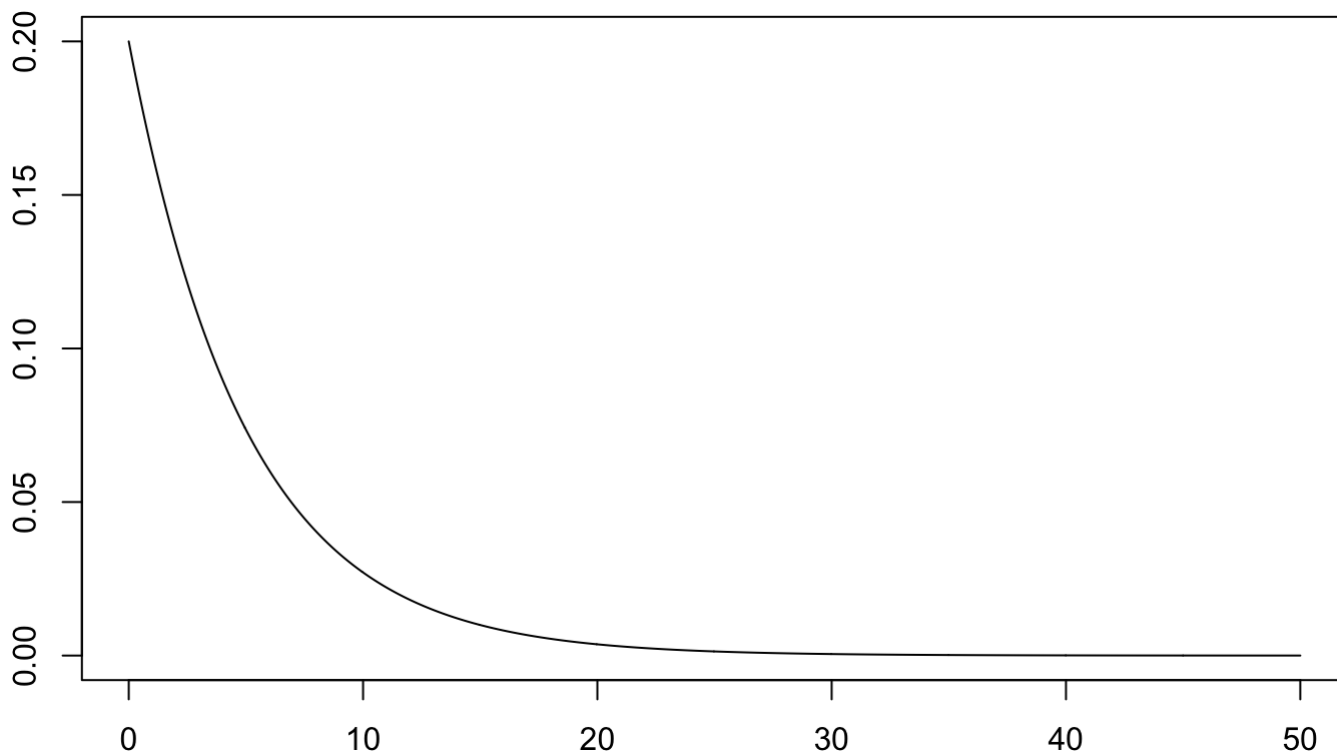

If The Central Limit Theorem is correct, then if we take a random sample from the above distribution, take the mean of that random sample, subtract off the population mean (which is 1/lambda), divide by the standard error of the estimate, and do this n times, the distribution of *these* sample means will be approximately normal for large values of n.

To put the above paragraph into a formula notation,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate} - \text{Mean of the Estimate}}{\text{Standard Error of the Estimate}}$$

Let's run a simulation to see if this is correct. We will take a 1000 samples of 40 random numbers from the above exponential distribution:

```
lambda = .2 #this is the rate from the above distibution
n = 40 #sample size
nSamps = 1000 #the number of samples

samps <- matrix(, nrow = nSamps, ncol = n)
for(i in 1:nSamps){
        samps[i,] <- rexp(n, rate = lambda)
}
```

Now we can take the mean of each of our 1000 samples to get the sample mean:

```
sampleMeans <- apply(samps, 1, mean)
```

Now let's take the CLT formula piece by piece. Subtract the population mean (1/lambda) from the sample mean:

$$\bar{X}_n - \mu$$

```
CLTtop <- (sampleMeans - 1/lambda)
```

Next divide the sample standard deviation (also 1/lambda) by the square root of the sample size, n (this is the standard error of the mean):

$$\sigma/\sqrt{n}$$

```
CLTbottom <- (1/lambda)/sqrt(n)
```

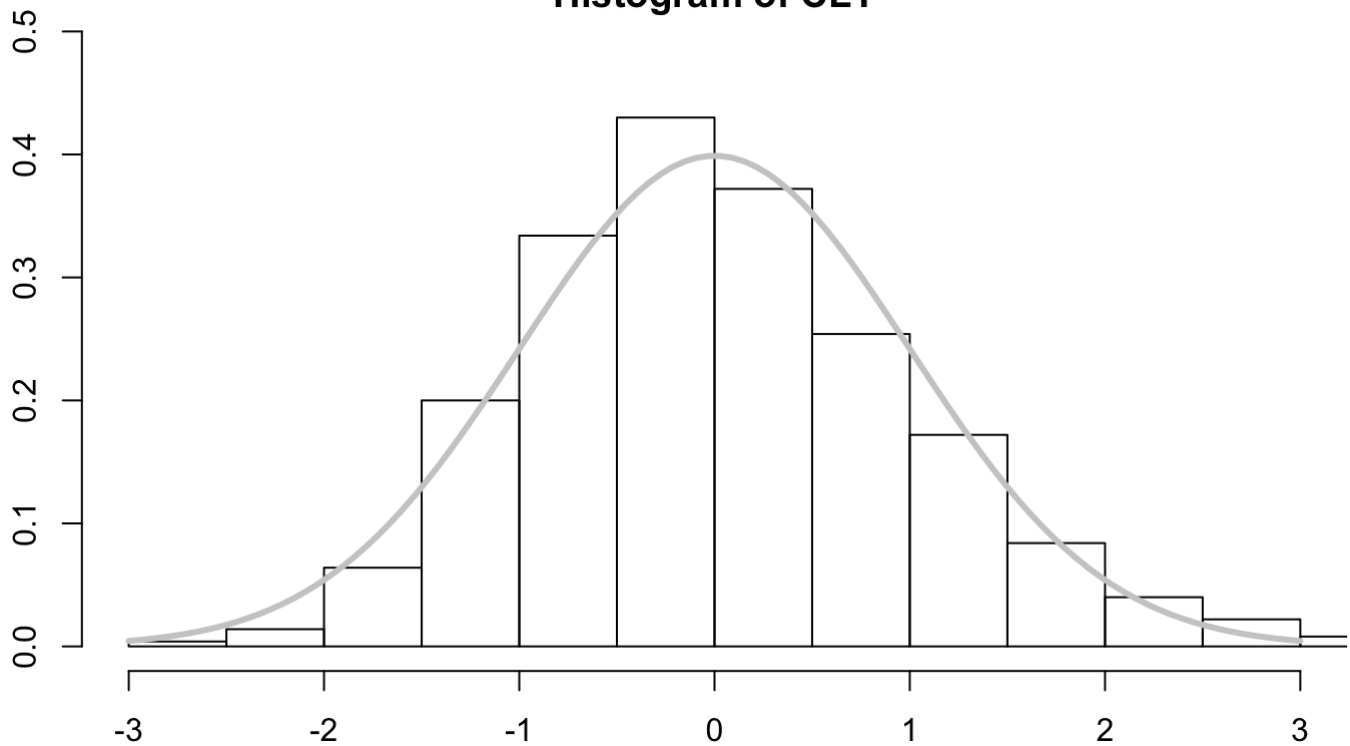Now divide former by the latter:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

```
CLT <- CLTtop / CLTbottom
```

And finally we will plot *that* distribution as a histogram:

```
par(mar=c(2, 2, .75, .1))
hist(CLT, freq = FALSE, xlim= c(-3,3), ylim = c(0,.5))
lines(seq(-3, 3, length = 100), dnorm(seq(-3, 3, length = 100)), col = grey(.8), lwd = 3
)
```

Histogram of CLT

We've overlayed a normal distribution density plot to show that our histogram is a very close fit to a standard normal. The Central Limit Theorem works!

To further illustrate, the mean of a standard normal is 0, and the variance is 1. Let's compare that to the mean and variance of our CLT calculation from our simulation samples:

```
mean(CLT)
```

```
## [1] 0.03515649
```

```
var(CLT)
```

```
## [1] 0.9727332
```

Looks like we're pretty close! With even larger values of n, these would be get even closer.