Central Limit Theorem and the Exponential Distribution: An Exploration Using Simulation in R

Jen Becker

July 23, 2015

In this paper, I will use R's exponential distribution to illustrate the predictions of the Central Limit Theorem, namely that the distribution of sample means from a population is nearly normal and centered around the population mean, even if the population distribution is not normal. Also, that the sample variance approximates the population's variance.

To simulate the exponential distribution in R, we use rexp(n, lambda) with n=40 and lambda=0.2. Data from a single run will generally resemble an exponential distribution. If we plot the data from 1,000 simulations of 40 exponentials, we see a much more accurate depiction of the exponential distribution.

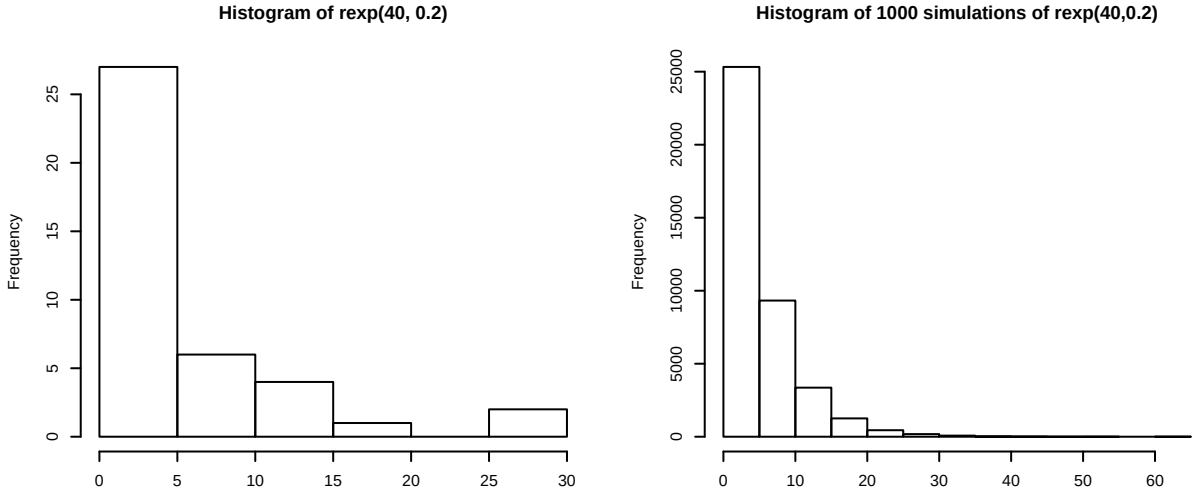

Figure 1: See appendix for supporting R code

R code for simulation of exponential distribution

To run our analysis, we collect a large number of exponential distributions. For each of 1,000 runs, we generate 40 numbers with an exponential distribution and calculate the mean and variance of that sample. On each run, we also calculate and store a running mean of means – averaging together the means we've collected so far. (See appendix for exact R code.)

Mean

The theoretical mean of an exponential distribution is 1/lambda, or 5 in this case. The mean of the sample means from our 1,000 simulations of 40 exponentials is

```
mean(simedmeans)
```

```
## [1] 4.999702
```

This is only off by 0.00596%.

Variance

The theoretical variation of this exponential distribution is 1/(lambda^2), or 25 in this case. The mean of the sample variations from our 1,000 simulations of 40 exponentials is

```
mean(simedvars)
```

```
## [1] 25.37287
```

This is only off by 1.49%.

Standard Error

The distribution of the mean of our samples of 40 exponentials have a variation that can be described through the standard devation. This is also the standard error for the sample mean. It should closely approximate the result of the formula

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

Our sample means have a standard devation of

```
sd(simedmeans)
```

```
## [1] 0.8020251
```

which is very nearly the calcuated value of the standard error.

$$\text{SE} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{40}} = 0.7905694$$

The variance of the sample means is significantly smaller than the theoretical variance for the exponential distribution.

```
var(simedmeans); 1/(lambda^2)
```

```
## [1] 0.6432442
```

```
## [1] 25
```

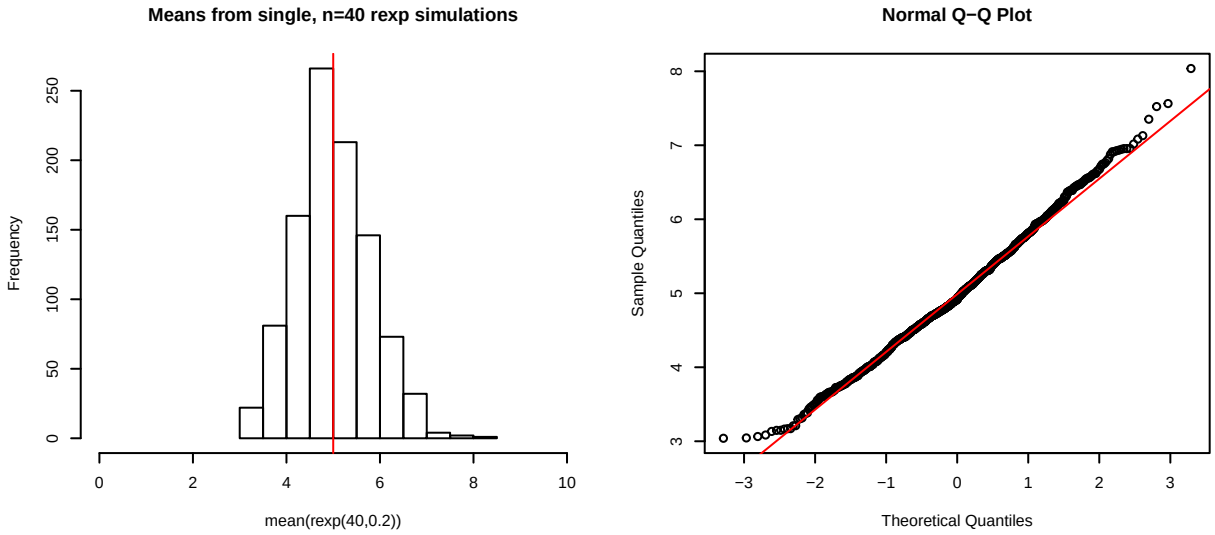

Figure 2: See appendix for supporting R code

Central Limit Theorem

As predicted by the Central Limit Theorem, the distribution of sample means from the simulations closely follows a normal distribution. The distribution is centered at the theoretical mean, 5 (shown in the figure with a vertical red line). The Q-Q plot shows that the values closely follow the expected normal distribution (shown in the figure with a red line).

While the mean of a single, n=40 simulation may vary from the population mean, as we gather the mean from more simulations and average them together, the value approaches the population mean (shown in the figure with a horizontal red line).

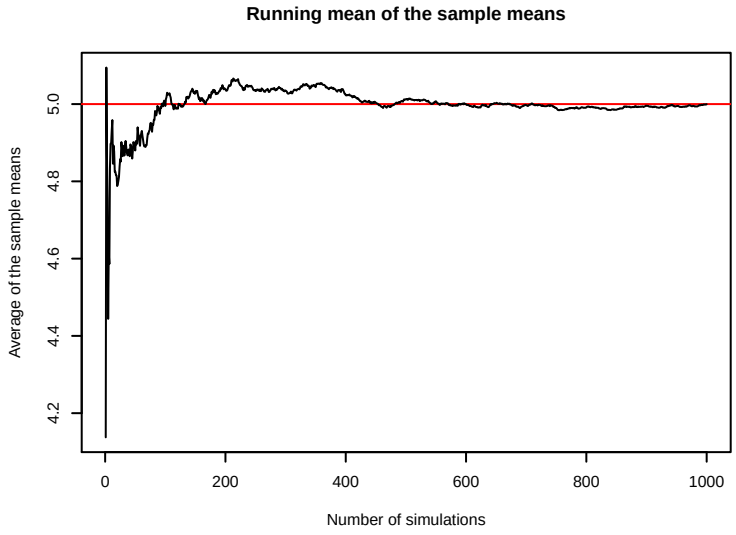

Figure 3: See appendix for supporting R code

Appendix

R Code for simulations

```
## Run 1,000 simulations with n=40.

set.seed(100)         # For reproducibility
lambda <- 0.2         # Lambda value for rexp
n <- 40               # Number of values to generate by rexp
simedvars <- c()      # Variance of each simulation
simedmeans <- c()     # Mean of each simulation
runningmean <- c()    # Running mean of the means of each simulation

for (i in 1:1000) {
    thisrun <- rexp(n, lambda)  # Run the simulation, n=40, lambda=0.2
    simedvars <- c(simedvars, var(thisrun))
    simedmeans <- c(simedmeans, mean(thisrun))
    runningmean <- c(runningmean, mean(simedmeans))
}
```

Figure 1:

```
par(mfrow=c(1,2), cex=.5)
## Histogram of 40 exponential values
hist(rexp(40,0.2), xlab="")

## Run 1,000 simulations with n=40
set.seed(100)
alldata <- c()
for (i in 1:1000) {
    thisrun <- rexp(40, 0.2)
    alldata <- c(alldata, thisrun)
}

## Histogram of 1,000 * 40 exponential values.
hist(alldata, main="Histogram of 1000 simulations of rexp(40,0.2)", xlab="")
```

Figure 2:

```
## Graph the means for each run with a normal distribution curve
par(mfrow=c(1,2), cex=0.5)
hist(simedmeans, main="Means from single, n=40 rexp simulations",
     xlab="mean(rexp(40,0.2))", xlim=c(0, 2*mean(simedmeans)))
abline(v=1/0.2, col="red")
curve(dnorm(x, mean=mean(simedmeans), sd=sd(simedmeans))*550, add=TRUE, col="blue")

qqnorm(simedmeans)
qqline(simedmeans)
```

Figure 3:

```
par(cex=.5)
## Graph the mean of the sample means as it changes as more simulations are run
## with a line representing the theoretical mean.
plot(x=1:length(runningmean), y=runningmean, pch=20, type="n",
     xlab="Number of simulations", ylab="Average of the sample means",
     main="Running mean of the sample means")
abline(h=1/0.2, col="red")
lines(x=1:length(runningmean), y=runningmean)
```