Statistical Inference Project Part 1: A Simulation

Andrew Chellinsky

August 20, 2016

Overview

This part of the Statistical Inference course project will use an exponential distribution simulation to explore three questions:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

After setting up the simulation, this report will look at each question in turn.

Simulation Setup

Per the instructions, let lambda = 0.2 and n = 40 for all simulations:

```
lambda <- 0.2

n <- 40
```

Additionally, there will be 1000 simulations. For reproducibility, a seed will be set.

```
numsims <- 1000

set.seed(331)
```

Because the questions concern the mean of each simulation, the following code will calculate the mean of each exponential distribution and store the 1000 simulations in a long vector.

```
simmeans <- replicate(numsims, mean(rexp(n, lambda)))
```

To verify this worked, let's look at the head, summary, and length of the `simmeans` vector.

```
head(simmeans)
```

```
## [1] 4.866161 3.706305 4.453645 6.249144 4.052733 6.264665
```

```
summary(simmeans)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.516   4.451   4.972   4.995   5.509   8.069
```

```
length(simmeans)
```

```
## [1] 1000
```

Sample Mean and Theoretical Mean Comparison

The first step is to show the sample mean and compare it to the theoretical mean of the distribution.

He is the mean of the sample:

```
mean(simmeans)
```

```
## [1] 4.994617
```

The theoretical mean, according to the instructions is 1/lambda for an exponential distribution. That means this simulation's theoretical mean is:

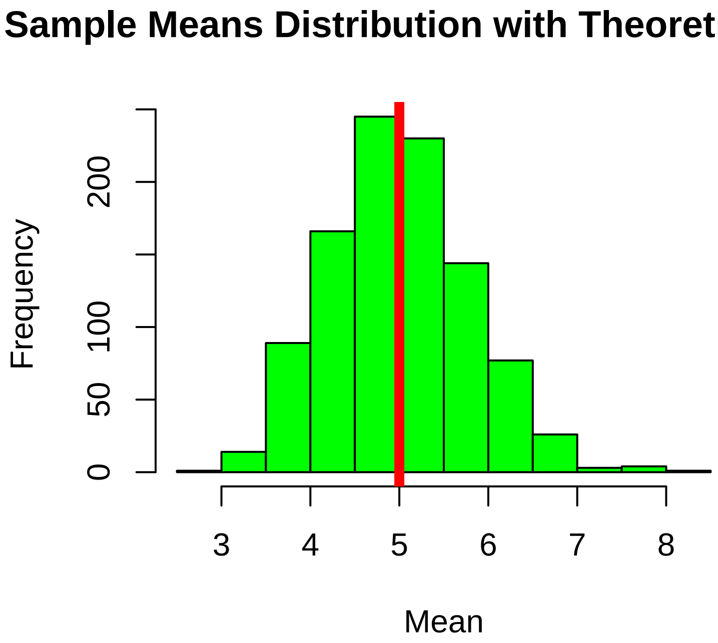
```
1/lambda
```

```
## [1] 5
```

The simulated and theoretical means are very close. To see how close they are visually, let's replot the histogram with a vertical line at the theoretical mean.

```
par(pin = c(3, 2))
hist(simmeans, xlab = "Mean", main = "Sample Means Distribution with Theoretical Mean", col = "green", 

abline(v = 1/lambda, col = "red", lwd = 5)
```



Sample Variance and Theoretical Variance Comparison

The second step is to show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

The sample's variance of the means is calculated similar to the mean:

```
var(simmeans)
```

```
## [1] 0.6139188
```

Also similar to the mean, the theoretical variance for an exponential distribution is 1/lambda^2. This is:

```
1/lambda^2
```

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

Note that this theoretical variance does not match the variance of the means. This is because the classical central limit theorem indicates that a large enough `n` gives a distribution where the variance approaches sigmaˆ2/n instead of sigmaˆ2. In our example, that would give a theoretical variance of:

```r
1/lambda^2/n
```

```
## [1] 0.625
```

This is a much closer result to the sample variance.

Distribution Approximates Normal

The third step is to show that the distribution is approximately normal.

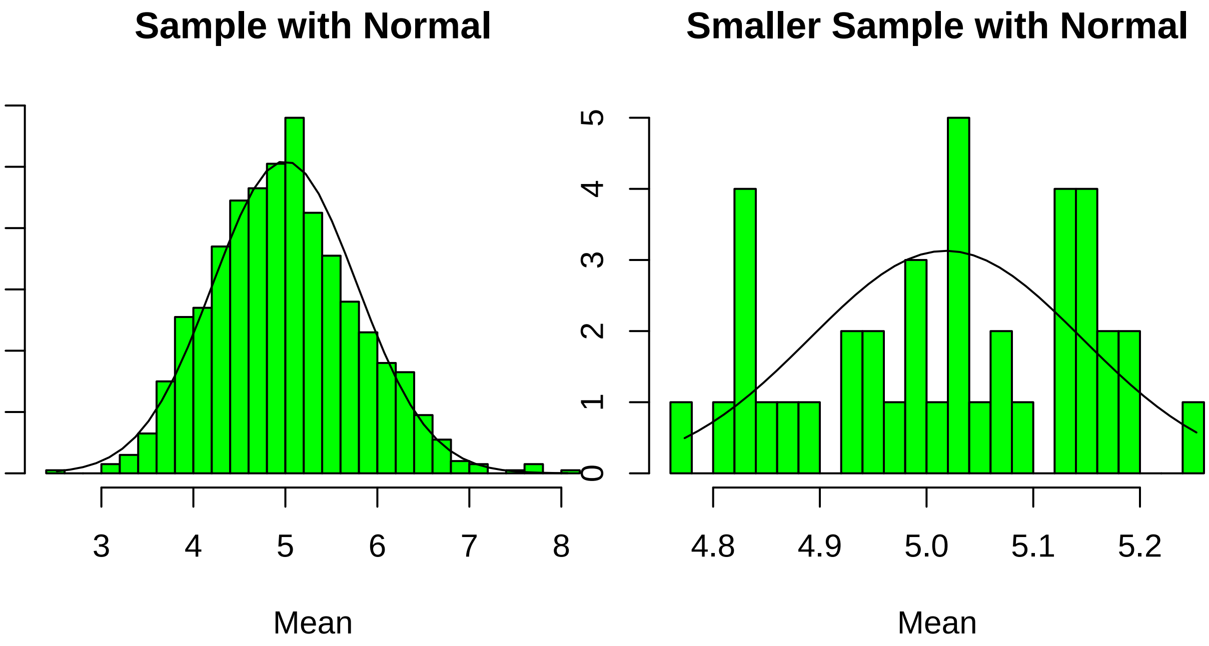
To do this, let's plot two histograms. The first is a replot of the histogram from earlier in this project. However, this time we will include smaller bins to give more definition to the plot and overlay a normal distribution line to compare the two visually.

To show how the simulation helped approximate a normal distribution, let's repeat the histogram and overlay. However, this time we will use 40 random exponential distributions with `n` equal to 1000 as the simulation.

```r
par(mfrow = c(1, 2), pin = c(3, 2))

h <- hist(simmeans, xlab = "Mean", main = "Sample with Normal", col = "green", breaks = 25)
xfit <- seq(min(simmeans), max(simmeans), length = 40)
yfit <- dnorm(xfit, mean = mean(simmeans), sd = sd(simmeans))
yfit <- yfit * diff(h$mids[1:2]) * length(simmeans)
lines(xfit, yfit)

small_sample <- replicate(40, mean(rexp(1000, lambda)))
hist(small_sample, xlab = "Mean", main = "Smaller Sample with Normal", col = "green", breaks = 25)
xfit <- seq(min(small_sample), max(small_sample), length = 40)
yfit <- dnorm(xfit, mean = mean(small_sample), sd = sd(small_sample))
lines(xfit, yfit)
```



Notice how the simulation presented a better approximation of the normal distribution. In other words, a large collection of random exponential distributions is a better approximation than a collection of 40 random exponentials.

Appendix

One note on the sample versus theoretical variance deserves some further exploration. The sample and theoretical variance were not too terribly close. To demonstrate the central limit theorem works as expected, let's rerun it with a larger `n` for the simulation.

To demonstrate that the estimation improves with a larger `n`, let's rerun the simulation and replace `n` with 1000. Then, we'll compare the new sample's variance with the theoretical variance.

```
simmeans_bign <- replicate(numsims, mean(rexp(1000, lambda)))
var(simmeans_bign)
```

```
## [1] 0.0244544
```

```
1/lambda^2/1000
```

```
## [1] 0.025
```

As you can see, the central limit theorem helps obtain a closer sample and theoretical variance here.