Statistical Inference Course Project: Part 1

Emma Ideal

June 12, 2015

Overview

In this analysis, we will run simulations to compare the exponential distribution with the Central Limit Theorem. The Central Limit Theorem indicates that given a sufficiently large sample size, the distribution of averages of the samples will be normally distributed, centered at the population (theoretical) mean and with a standard deviation equal to the standard error of the mean.

Simulations

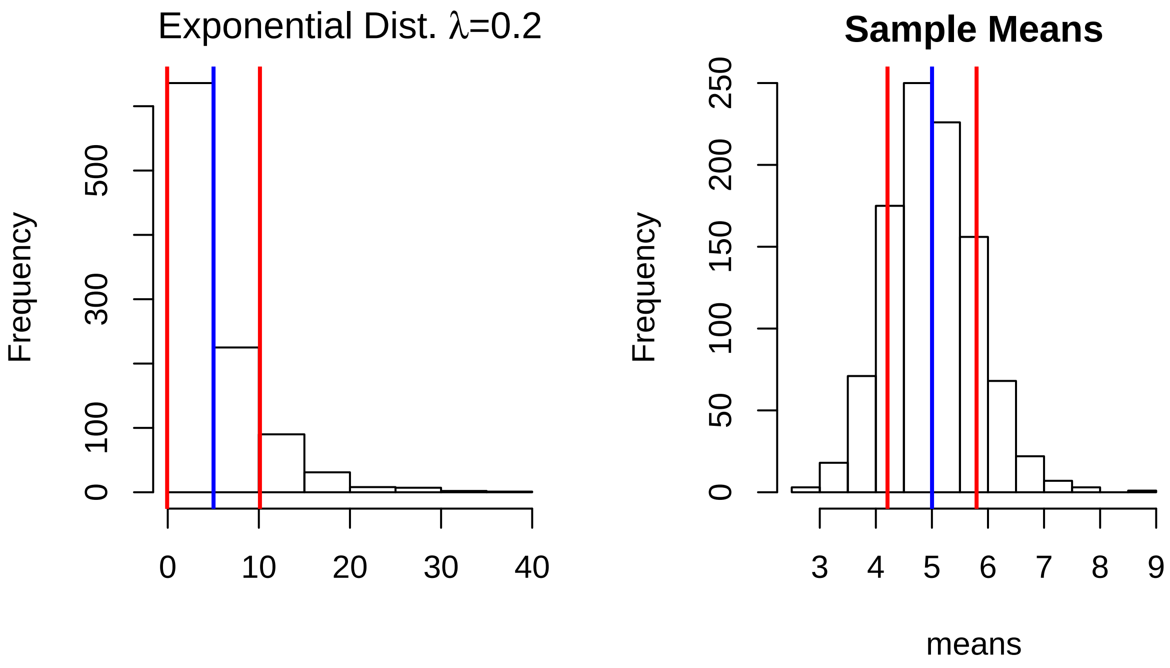
Here, we will run 1000 simulations of 40 randomly generated exponentials with lambda = 0.2.

Sample Mean vs. Theoretical Mean

We first simulate 1000 random exponentials and then generate 1000 means of 40 random exponentials. The results are plotted on the next page.

```r
# 2 panel plot
par(mfrow=c(1,2), mar=c(9,5,2,1))
set.seed(63)
# Simulate 1000 random exponentials, lambda = 0.2
sim <- rexp(1000,0.2)
mn <- mean(sim)
sd <- sd(sim)
# Histogram the random exponentials
hist(sim, main=expression(paste('Exponential Dist. ', lambda, '=0.2')),
     xlab='')
# Add vertical lines at the mean and 1 standard deviation marks
abline(v = mn, col = 'blue', lwd=2)
abline(v = mn + sd, col = 'red', lwd=2)
abline(v = mn - sd, col ='red', lwd=2)

means <- NULL
# Generate 1000 means of 40 random exponentials
for (i in 1:1000){
        set.seed(i)
        means <- c(means, mean(rexp(40,0.2)))
        }
# Histogram the means and draw vertical lines at the center and 1 standard deviation marks
hist(means, main='Sample Means')
abline(v = mean(means), col = 'blue', lwd = 2)
abline(v = mean(means) + sd(means), col = 'red', lwd = 2)
abline(v = mean(means) - sd(means), col = 'red', lwd = 2)
```



Vertical blue lines have been placed at the means of the two distributions, while red lines mark one standard deviation away from the mean. The theoretical mean, marked in blue on the left plot, and the mean of the sample means, marked in blue on the right plot are:

```
# Theoretical mean
mn
```

```
## [1] 5.026507
```

```
# Average of the sample means
mean(means)
```

```
## [1] 5.002327
```

The theoretical mean is equal to

$$1/\lambda = 1/0.2 = 5,$$

and this is precisely what we find. The average of the sample means should converge to this theoretical mean, given a large enough sample size, and in fact, we find that the distribution of sample means is centered on 5.

Sample Variance vs. Theoretical Variance

The theoretical variance of an exponential distribution, given a large simulation size, is equal to

$$(1/\lambda)^2 = (1/0.2)^2 = 25.$$

We can compute this theoretical variance in our simulation:

```
# Theoretical variance
set.seed(63)
th_var <- var(rexp(1000,0.2))
th_var
```

```
## [1] 25.92878
```

However, the variance of our sample means distribution is expected to be the theoretical variance divided by the sample size (n = 40):

```r
# Expected variance
th_var/40
```

```
## [1] 0.6482194
```

Compare this to the variance of the sample means distribution, and find that it is quite close to the expected variance:

```r
# Simulated variance
var(means)
```

```
## [1] 0.6308244
```

Distribution of Sample Means: Is it ~ Normal?

To verify that the distribution of sample means is roughly normal, we can compute, for example, how much of the distribution lies between one standard deviation of the mean (we expect this to be ~68%):

```r
# Define the 1 standard deviation marks
one_below <- mean(means) - sd(means)
one_above <- mean(means) + sd(means)

# Compute the fraction of the distribution within 1 std dev from the mean, compare to 0.68
sum(means > one_below & means < one_above)/1000
```

```
## [1] 0.675
```

We can then compute, for example, the fraction of the distribution lying below and above 2 standard deviations from the mean. We expect both of these numbers to be roughly equal to 0.025.

```r
two_below <- mean(means) - 2*sd(means)
two_above <- mean(means) + 2*sd(means)

sum(means < two_below)/1000
```

```
## [1] 0.017
```

```r
sum(means > two_above)/1000
```

```
## [1] 0.026
```