Statistical Inference - Bootstrapping and the Central Limit Theorem

Jonathan Hill, MPA

Wednesday, April 22, 2015

Overview

This project demonstrates the similarity between bootstrapped data using R's rexp() function, which generates random deviates from an exponential distribution with the rate lambda, and the theoretical distribution of an exponential equation, 1/lambda where lambda = 0.2.

In order to illustrate this relationship, I ran 1000 simulations with sample sizes of 40 and analyzed the characteristics of that simulation as compared to their theoretical values.

First I set the seed to 1 so that the randomly generated numbers in my simulations can be reproduced:

```
# Set the seed to 1  #
set.seed(1)
```

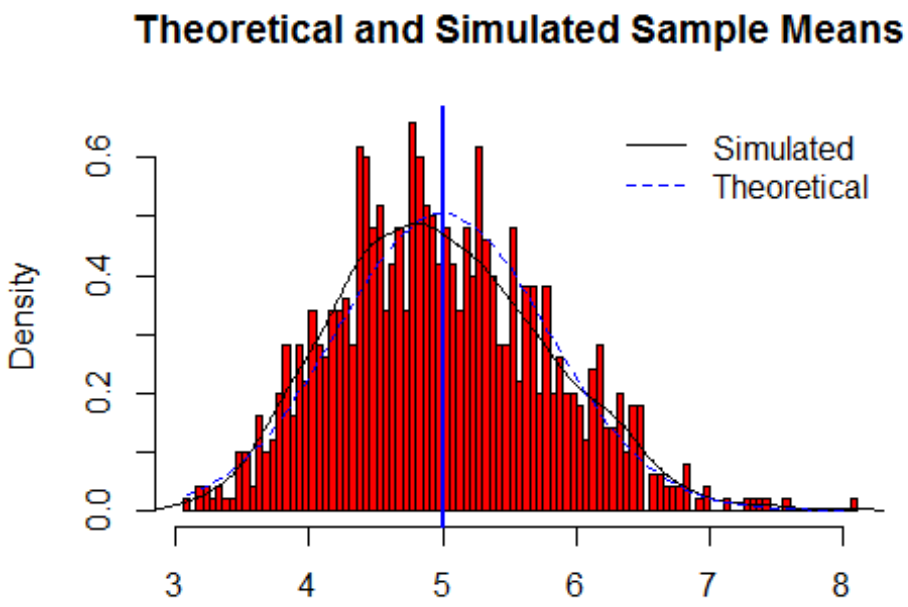
Simulations

```
# Variables #
lambda <- .2
n <- 1000
sample.size <- 40

# Simulate the data #
sim <- matrix(rexp(n*sample.size, lambda), n, sample.size)

# Calculate the mean of each simulation #
sim.dat <- rowMeans(sim)
```

Sample Mean versus Theoretical Mean

The theoretical mean of the distribution, 1/lambda, is 5. However, the mean of the simulated data is 4.99.

This is a histogram of the simulated data with a blue line representing the theoretical mean and density curves for both distributions:



As you can see, they are practically the same. There are some slight differences between the density curves, but the characteristics of the theoretical and simulated distrubtions could be assumed to be the same for most practical applications.

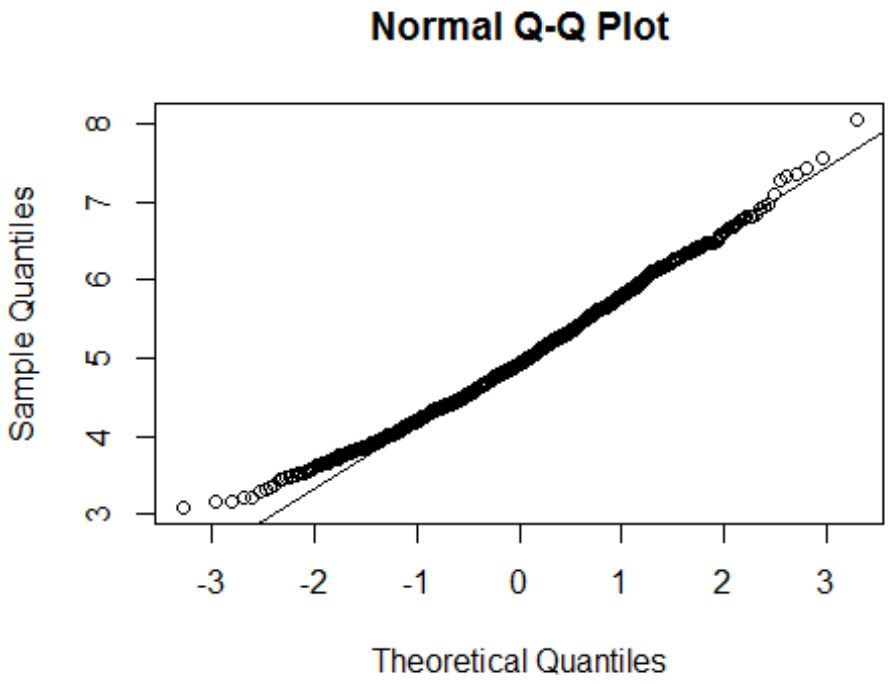
Sample Variance versus Theoretical Variance

```
# standard deviation of simulated data #
sim.sd <- sd(sim.dat)

# variance of simulated data #
sim.var <- sim.sd^2

# calculated standard deviation #
theoretical.sd <- (1/lambda)/sqrt(sample.size)

# calculated variance #
theoretical.var <- ((1/lambda)*(1/sqrt(sample.size)))^2
```

The standard deviation of the simulated data is 0.786, while the theoretical standard deviation is 0.791. The respective variances are 0.618 and 0.625. These differences are slightly larger than the differences between the sample means and theoretical mean, but they are still very similar.

Distribution

Here is a qqplot of the distribution:



Most of the points lie along the normal line. There are some slight deviations, but they are not significantly large. The simulated data are normally distributed with mean 4.99.