Part I - Simulation Exercise

Michael Baldassaro

1/10/2018

Overview

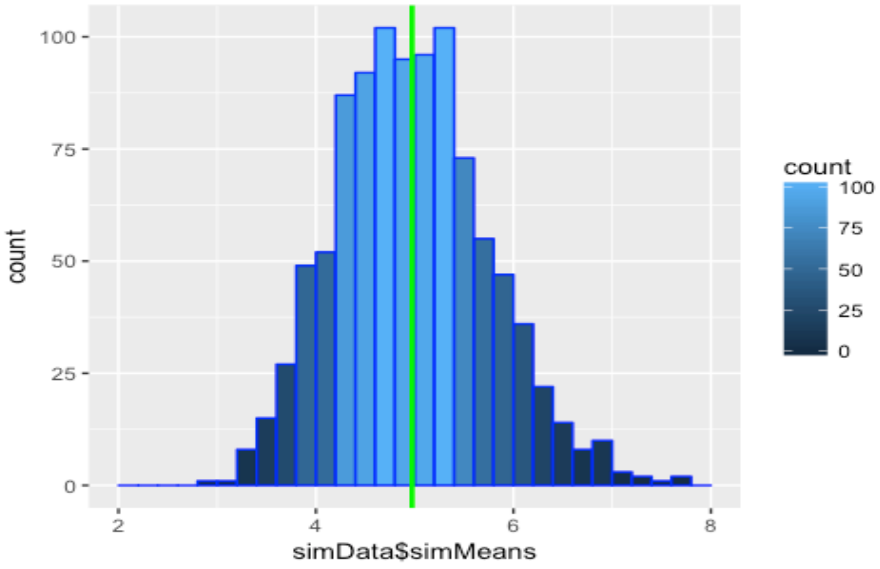
This project uses R to investigate the exponential distribution in comparison to the Central Limit Theorem. Specifically, this project investigates the distribution of averages of 40 exponentials over 1000 simulations where lambda = 0.2 is the rate parameter and the mean of the exponential distribution is 1 / lambda.

Simulation

We simulate the exponential distribution using `rexp(n,lambda)` and explore comparative distributions using the ggplot package. We will specifically explore the comparison of the center of the exponential distribution vis-a-vis the theoretical center of the distribution. We will also explore the variance of the exponential distribution vis-a-vis the theoretical variance.

First, we will install and load the ggplot library, seed for replication, and initialize our variables. Then we'll construct a 1000 by 40 matrix containing the 1000 observations from our 40 random simulations.

Next we'll calculate the means of each observation, create a dataframe of our matrix and mean data, and explore the basic features of the data using ggplot. The histogram displays the sampling distribution as well as the mean of the sampling distribution represented by a green line.



Sample Mean vs. Theoretical Mean

We calculate the mean of the sampling distribution using the `mean()` function and the theoretical mean using `1 / lambda`:

```
(simMean <- mean(simMeans))

## [1] 4.974239

(expMean <- 1 / lambda)

## [1] 5
```

The sampling distribution mean (4.974239) very well approximates the theoretical mean of the normal exponential distribution (5).

Sample Variance vs. Theoretical Variance

We calculate the variance of the simulated sampling data using the `var()` function and the variance of the theoretical exponential distribution using `(1 / lambda)^2 / n`:

```
(simVar <- var(simMeans))

## [1] 0.5949702

(expVar <- ((1 / lambda)^2 / n))

## [1] 0.625
```

The variance of the sampling distribution (0.595) is very close to the theoretical variance of the normal exponential distribution (0.625).

We will also calculate the standard deviation of the simulated sampling data using the `sd()` function and the standard deviation of the theoretical exponential distirbution using `1 / (lambda * sqrt(n))`:

```
(simSD <- sd(simMeans))

## [1] 0.7713431

(expSD <- 1 / (lambda * sqrt(n)))

## [1] 0.7905694
```

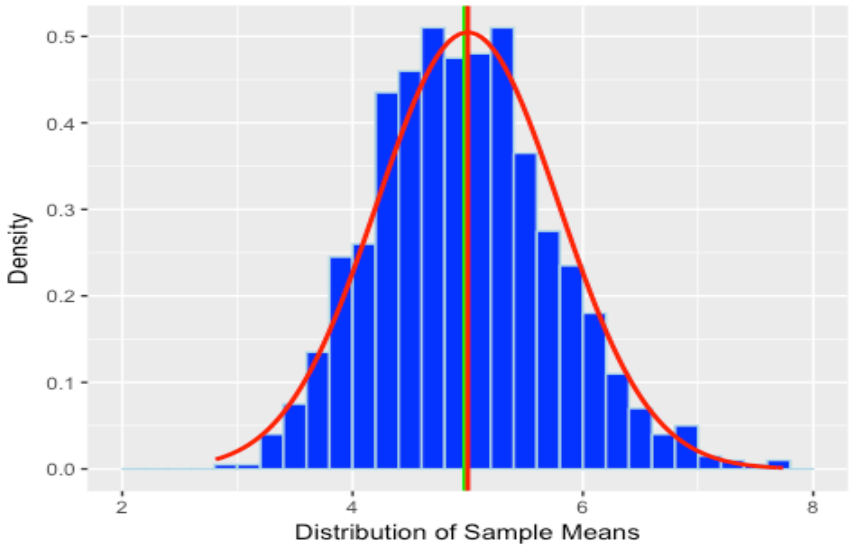
As expected these values are also quite close.

Distribution

Now we can illustrate the sample mean in comparison to the theoretical mean of the normal exponential distribution using ggplot. Specifically, we will display a

histogram of the simulated sample distribution with an approximate normal distribution curve.

```
ggplot(simData, aes(x=simData$simMeans)) +
geom_histogram(breaks=seq(2,8, by=0.2), aes(y=..density..),
color="lightblue", fill="blue") +
geom_vline(aes(xintercept=mean(simData$simMeans)), color="green",
size=1) + geom_vline(aes(xintercept=expMean), color="red", size=1) +
stat_function(fun=dnorm, args=list(mean=expMean, sd=expSD),
color="red", size=1) + labs(x="Distribution of Sample Means",
y="Density")
```



The red curve is the normal distribution curve generated from the theoretical exponential mean and standard deviation. The sampling distribution and mean of the sampling distribution, represented by the green line, closely approximate the normal distribution and theoretical mean.

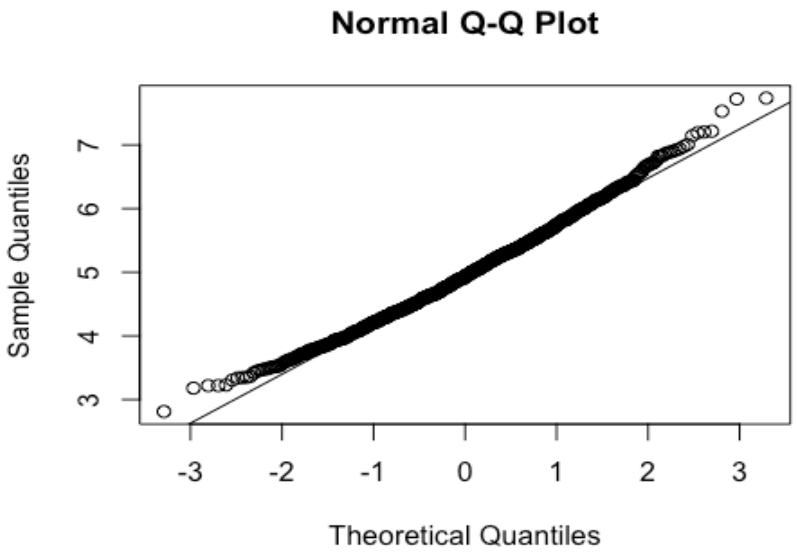
We can calculate a confidence interval for both the sample mean and theoretical exponential mean using the 95 CI z-score value (1.96):

```
(simCI <- simMean + c(-1, 1) * 1.96 * sqrt(simVar) / sqrt(n))

## [1] 4.735197 5.213280

(expCI <- expMean + c(-1, 1) * 1.96 * sqrt(expVar) / sqrt(n))

## [1] 4.755 5.245
```

The upper and lower bounds of the confidence intervals for both the mean of the sampling distribution and theoretical exponential distribution are very similar.

Lastly, we can use a Quantile-Quantile (Q-Q) Plot to check the validity of the distributional assumption. The `qqnorm()` and `qqline()` functions handily generate a Q-Q Plot for comparison:

```r
qqnorm(simMeans)
qqline(simMeans)
```



As illustrated, the distribution of sample and theoretical quantiles is closely aligned.