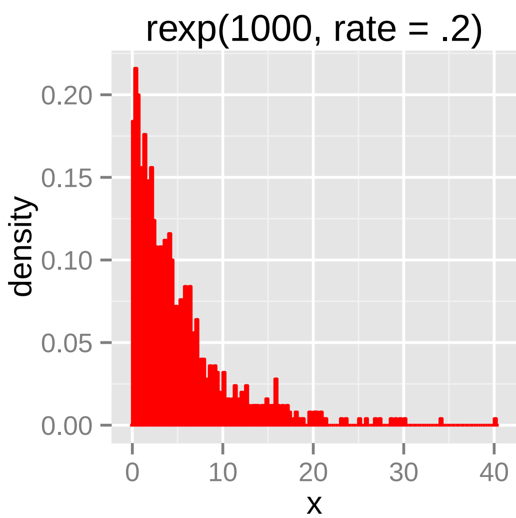
Inferential Statistics Assignment 1 Part 2

treepruner

September 24, 2015

Overview:

This is the project for the statistical inference class. It illustrates via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. This is a plot an exponential function:



Create a Simulation

The simulation is based upon one in the RMD markdown file for section 07. The original code was simulated rolling a die. Here is the heart of that simulation:

| apply(matrix(sample(1 : 6, nosim * 10, replace = TRUE), nosim), 1, cfunc, 1 0)

The portion of the code related to the die was replaced by the rexp(1000, rate = lambda). The function and the portion of the code relating to changing the sample size was also removed. This is the simulation code I used:

```
nosim <- 1000
dat <- data.frame(x = matrix(sample(rexp(1000, rate = lambda), nosim * 40, replace = TRUE), nosim))
```

There were 1000 samples taken from 40 different simulations. The resulting data frame had 1000 rows and 40 columns. The mean was calculated for each row and stored in the last variable with:

```
dat$row_mean = rowMeans(dat)
```

Sample Mean versus Theoretical Mean:

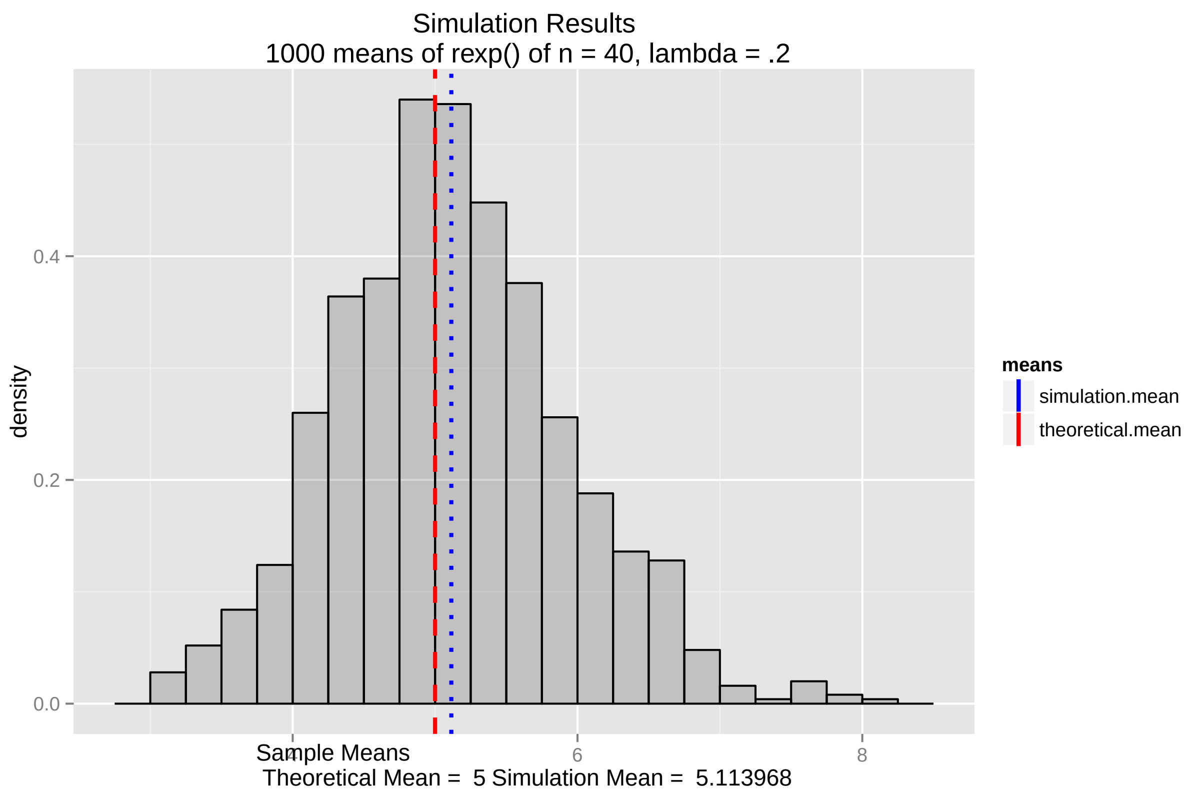
The theoretical mean = 1 / lambda. In our case that is 1/.2 = 5.

The simulation.mean is the mean of those 40 means.

```r
simulation.mean <- mean(dat$row_mean)
```

```
## [1] 5.113968
```

This graph plots the distribution and overlays the theoretical and simulation means. The means are nearly the same, which is to be expected because the law of large numbers says the sample mean of iid samples is consistent for the population mean. A dotted line was selected to display the theoretical mean in order for them both to be visible.



Sample Variance versus Theoretical Variance:

The theoretical standard deviation for this population is 1 / lambda which is 5.

The theoretical variance is (1 / lambda)^2 which is 25.

Per our TA Brian Canute, the expected standard deviation within the set of 1000 sample means is predicted by the sampling standard error. This is the standard deviation / sqrt(sample size). In our case this is 5/sqrt(40) or 0.7905694

The standard deviation of our simulation is

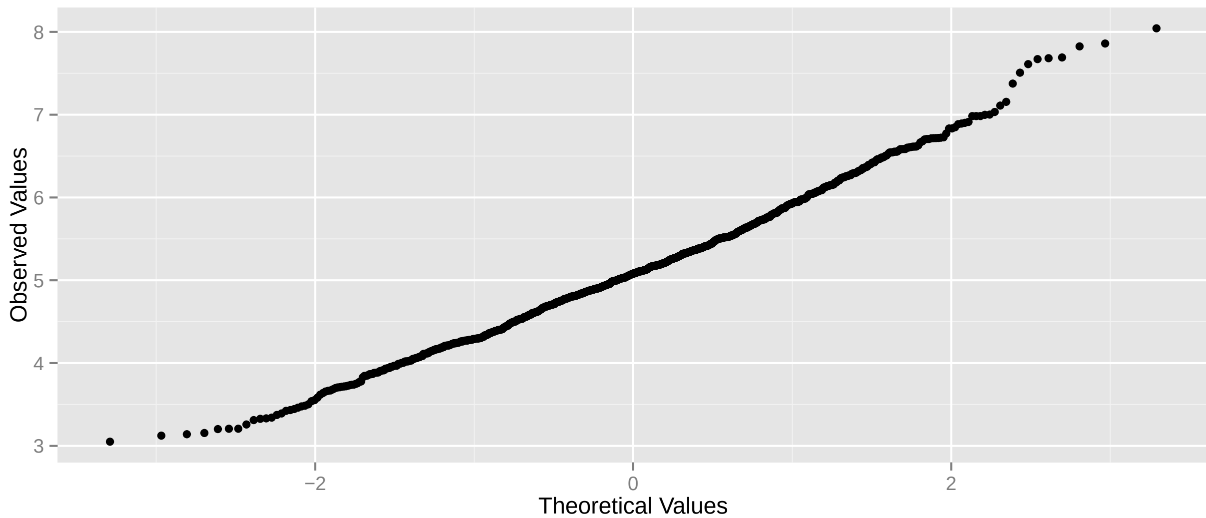
```
## [1] 0.8112755
```

The simulation is behaving as expected.

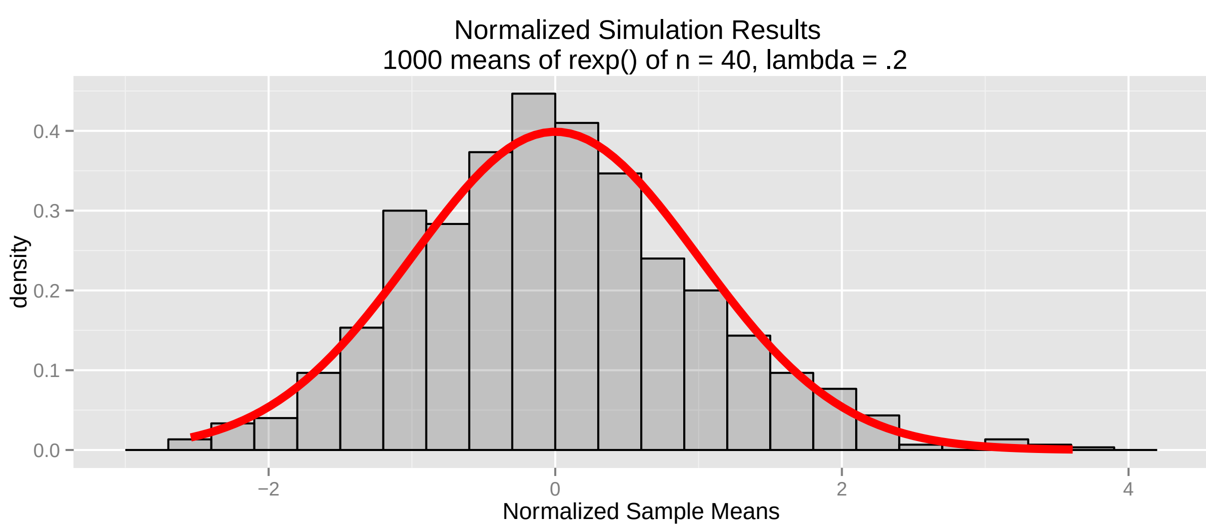
Distribution:

The distribution of the sample means looks completely different than the original plot. This is because the sample means vary around the population mean. From our online textbook, "The Central Limit Theorem states that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases." So, if you collect enough sample means, the histogram of them will look like a normal distribution.

The means of the samples can be compared to a normal distribution with a Q-Q plot. The Q-Q plot would be a straight line if it was perfectly normal.



Normalizing the simulation results by subtracting the simulation mean and dividing by the simulation standard deviation allows us to overlay normal distributon and visualize the differences better:



The distribution of normalized sample means approximates a normal distribution.