Statistical Inference: Comparing Simulated Exponentially Distributed Data to the Theoretical Exponential Distribution

Keith S Messina

August 21, 2015

Overview

To provide an example of the Central Limit Theorem (CLT) in action, we will look at simulations of data using the R random number-generating exponential function, `rexp()`, and compare it to what we expect to see from a theoretical standpoint. We will first show how the mean of this data approaches the theoretical mean, while simultaneously showing that the mean is better approximated with larger numbers of simulations.Then we will also look at the distribution of the means of multiple sets of data and compare its variability to the theoretical variability. Finally, we take a look at the distribution of the averages of the data and how it compares to the prediction of the CLT.

Parameters

In the comparison, we will be looking at the exponential distribution with a lambda of `.2`. The mean of the exponential distribution is 1/lambda, so in this case we are looking at a theoretical mean of `5`. The standard deviation of the exponential distribution is also given as 1/lambda, or `5` in this case. I took the data simulation number to be `1000` and the sample sizes of my data sets to be `40`. The seed of the simulations was set to `367` for reproducibility.

Simulations

The first simulation I ran was a a simple generation of 1000 data points using the `rexp()` function.

```r
simulated.data <- data.frame(count = 1:sim.size, exp = c(rexp(sim.size, rate = lambda)))
```

To show how the simulated means approached the theoretical mean over the course of many simulations, I took the data simulated in the first simulation and evaluated the cumulative mean of the data, saving that to a new data frame, `cummean.sim`.

```r
cummean.sim <- NULL
cummean.sim <- data.frame(cumsum(simulated.data)/1:sim.size)
cummean.sim$count <- cummean.sim$count * 2
```

Next, to show the variance of the data compared to the theoretical, I ran a simulation of `40` random exponentially distributed numbers and took `1000` iterations of this. I then found the Standard Error for the samples to show that the Standard Error is approximately normally distributed around the theoretical variance of `25`, $(1/\text{lambda})^2$, as predicted by the CLT. This shows the amount of variation in the data as compared to the theoretical variation.

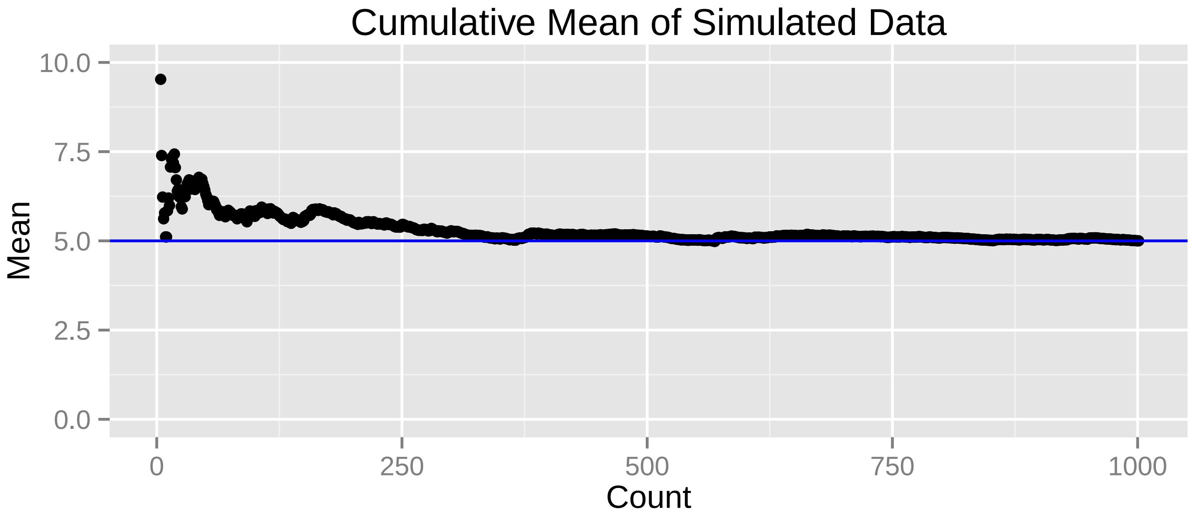
```
var.samples <- NULL
for (i in 1 : sim.size){
  sample.data <- rexp(sim.size, rate = lambda)
  var.samples <- c(var.samples, sd(sample.data)^2)
}
var.samples <- as.data.frame(var.samples)
```

Finally, to show the sample mean distribution, I ran a simulation of 40 random exponentially distributed numbers and took 1000 iterations of this. The mean of each iteration was saved to a vector and then a histogram of this vector was plotted to show the distribution of the means of the simulations.

```
mean.samples <- NULL
for (i in 1 : sim.size){
  sample.data <- rexp(sim.size, rate = lambda)
  mean.samples <- c(mean.samples, mean(sample.data))
}
mean.samples <- as.data.frame(mean.samples)
```

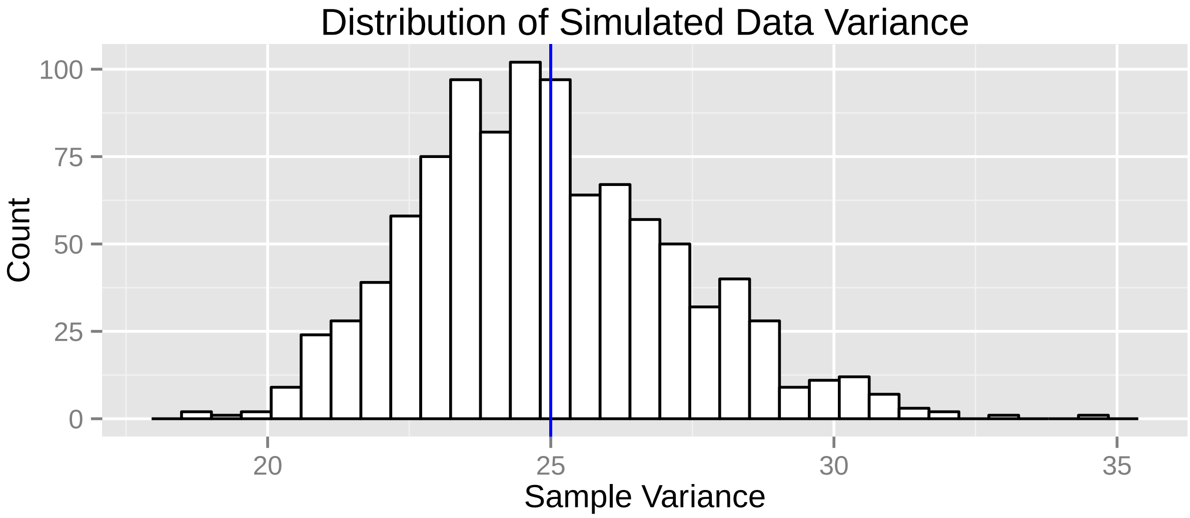
Sample Mean versus Theoretical Mean

To see how the mean of the sample approaches the mean of the population, we look at the cumulative mean of an ever-increasing sample size as it approaches the population size. We can assume the population of this data is the theoretical exponential distribution, in which case the population mean is the theoretical mean of the exponential distribution of 1/lambda or 5 in this case. The blue line is at the theoretical value of 5.



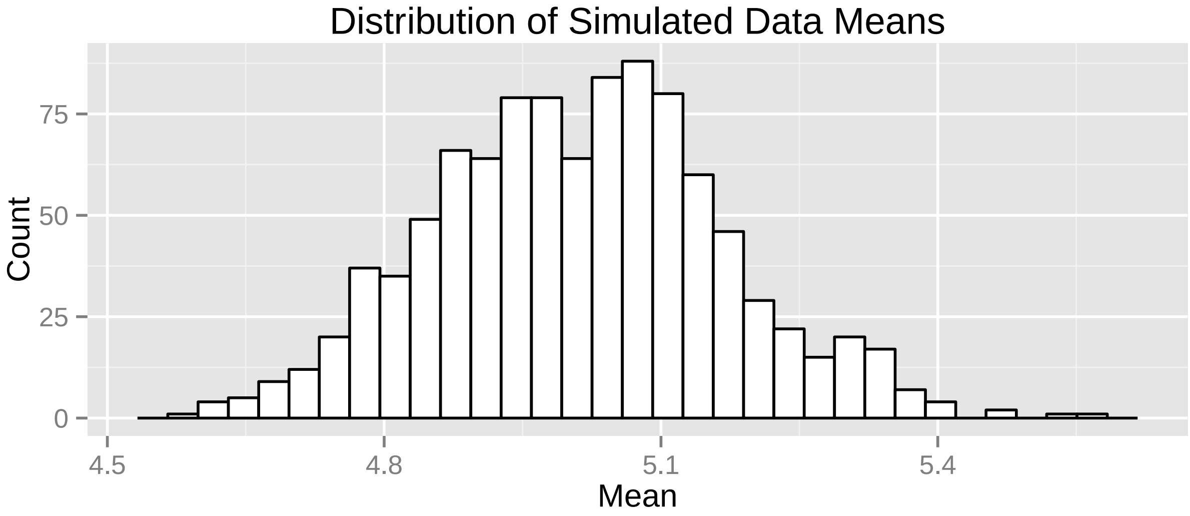
Sample Variance versus Theoretical Variance

If we take a look at the variance of samples of 40 observations using the Standard Error, we can see that their distribution is approaching a normal distribution. This distribution should be normally distributed around the theoretical variance,25, as predicted by the CLT.



Distribution

If we take samples with size `40` of the theoretical exponentially distributed population, we should find that their means are distributed normally about `0` according to the CLT.


Distribution of Simulated Data Means
Count
Mean