Nicholas Stanford
Statistical Inference
Data Science Specialization
Johns Hopkins University
Coursera
May 25, 2017

Course Project Part 1: A Simulation Exercise

1. Overview

The goal of this part of the assignment is to study the distribution of a sample mean. Letting n be the size of the sample, according to the CLT this distribution is roughly normal for large n. Often $n = 30$ is considered the cutoff for invoking the CLT, so taking $n = 40$ here will certainly suffice. We will simulate data for $1,000$ observations of the sample mean, compute the mean and variance of these observations, and compare those values to the theoretical means and variances of the distributions from which the observations are drawn.

2. Simulations

To create the simulated data and associated graphics, I used the following R code:

```
#Fix a randomization.
> set.seed(0)

#Create an empty vector that will contain samples from a distribution.
> mns=NULL

#Populate the vector with 1000 entries that are the means of samples
 of 40 random variables from the distribution exp(.2).
> for(i in 1:1000) mns=c(mns,mean(rexp(40,.2)))

#Compute the mean of this new sample of sample means.
> means(mns)
[1] 4.989678

#Compute its variance as well.
> var(mns)
[1] 0.6181582

#Create a histogram of it with 50 different bars.
> hist(mns,breaks=50)

#Overlay the histogram with a normal curve with the same mean and variance.
> hist(mns,prob=TRUE,breaks=50,main="Normal Curve over Histogram of mns")
> curve(dnorm(x,mean=mean(mns),sd=sd(mns)),add=TRUE)
```

3. Sample Mean versus Theoretical Mean

To analyze the theoretic foundations of this problem, let X_i be a set of iid random variables such that

$$X_i \sim exp(.2).$$

It is known that $E[X_i] = \frac{1}{.2} = 5$. Each entry in `mns` is an iid random variable Y_j defined by

$$Y_j = \frac{1}{40}\sum_{i=1}^{40} X_i.$$

Using linearity of expectation one then computes that

$$E[Y_j] = E\left[\frac{1}{40}\sum_{i=1}^{40} X_i\right] = \frac{1}{40}\sum_{i=1}^{40} E[X_i] = \frac{1}{40}\sum_{i=1}^{40} 5 = 5.$$

Thus the sample mean for the Y_i and the theoretical mean of the distribution from which the Y_i are drawn agree, as

$$4.989678 \approx 5.$$

4. Sample Variance versus Theoretical Variance

It is similarly known that $Var[X_i] = \frac{1}{.2^2} = 25$. It therefore follows from properties of variance that

$$Var[Y_j] = Var\left[\frac{1}{40}\sum_{i=1}^{40} X_i\right] = \frac{1}{40^2}\sum_{i=1}^{40} Var[X_i] = \frac{1}{40^2}\sum_{i=1}^{40} 25 = \frac{25}{40} = \frac{5}{8} = .625.$$

This theoretical variance of the distribution of the Y_i agrees with the sample variance, as

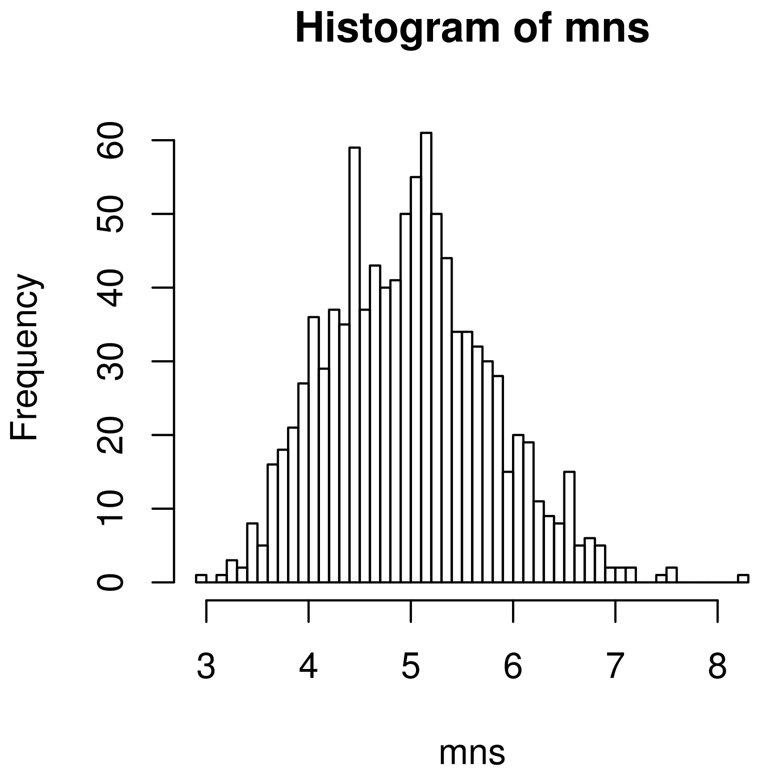
$$.6181582 \approx .625.$$

5. Distribution

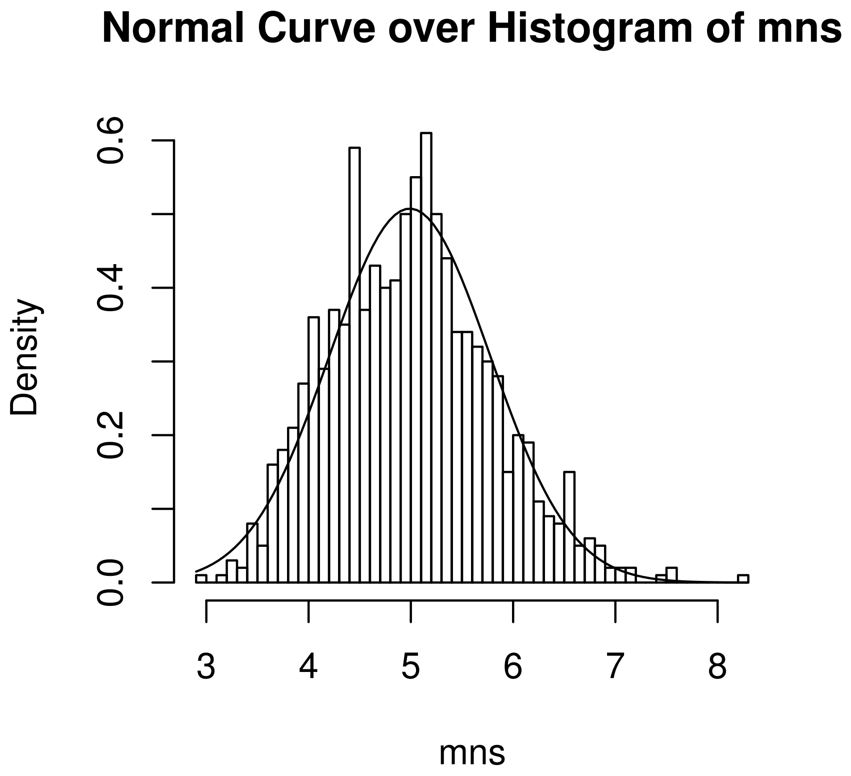
As seen in the second figure on the following page, the graph of a normal distribution with mean and variance equal to those of the distribution of the Y_j roughly fits the histogram of `mns`. I chose to use 50 subdivisions to make the shape of the distribution of the Y_j more apparent: too few subdivisions and it is hard for any shape to emerge, and too many makes the plot messy.

The correspondence between the sample distribution and the associated normal curve of course in not perfect, and this is in large part due to $n = 40$ being a relatively small value of n. A much stronger fit would appear for larger values of n, and this would primarily be due to the lower variance in sample means with larger sample sizes.

It is also of course impossible for a normal curve to fit this histogram perfectly because the normal curve has positive density over all reals, whereas the support of an exponential distribution is only the positive integers.



Histogram of mns



Normal Curve over Histogram of mns