Simulation Exercises to Reinforce the Foundations of Statistical Thinking in Online Classes

Simcha Pollack, Ph.D.

St. John's University

Tobin College of Business

Queens, NY, 11439

pollacks@stjohns.edu

0. ABSTRACT

Numerical literacy is a core requirement for functioning well in the global economy. Statistical thinking is a major component of that core but many students, especially those in online courses, find the abstract concepts of statistics difficult to assimilate. This paper describes a unified set of simulation exercises that assist in the teaching of introductory statistics. These simulations are done in parallel with the course material. This series of exercises graphically demonstrate descriptive and inferential statistical concepts in a way that, in the opinion of surveyed students, makes for better understanding of these abstract ideas.

The dominant formula in statistics is the mean. Since means are calculated based on a sample and samples are obtained from a population, we begin by defining a population whose parameter values are known and easily accessible: the discrete uniform distribution from 0 to 9. This distribution is readily obtained by employing the random number generator within Excel. Through repeated sampling from this population the student learns in concrete terms the meaning of such subtle notions as: the Central Limit Theorem; probability; confidence interval estimation; hypothesis testing and alpha and beta. In addition, the student is given an opportunity to hone their computer and numerical skills. A description of the exercises and a student's sample project follows.

Many students find the concepts discussed in introductory statistics to be abstract and difficult. In particular the section of the course called inferential statistics gives much trouble to the student who is not mathematically sophisticated. In the course of this simulation project the student is able to verify that the predictions of the standard statistical formulas correspond to the results obtained from repeated sampling. This educational tool is specifically designed to explain the basic logic of inferential statistics and has been successfully tested over time. Most students have favorable thing to say about their learning experience.

Monte Carlo simulation is a useful tool for teaching and learning statistics. The basics of descriptive statistics (e.g., mean, standard deviation, distributions, histograms) and introductory inferential statistics (e.g., confidence intervals, hypothesis testing) are all taught using a unified framework and an incremental and cumulative technique.

1. INTRODUCTION:

This paper describes the use of a variety of collaborative Monte Carlo simulation project to assist in the teaching of introductory statistical concepts. These simulations are carried out as part of a project that students do in stages (in parallel with the course material) and hand in periodically during the course of the term. The project described below consists of a series of simulation exercises that together graphically demonstrate descriptive and inferential statistical concepts in a way that makes for better understanding of these abstract ideas.

The primary formula in statistics is the mean. Since all means are calculated based on a sample and all samples are obtained from a population, we begin by defining a population whose parameter values are known and easily accessible: the discrete uniform distribution from 0 to 9. This distribution is readily obtained by picking numbers from a random number table, using a randomly generated number within Excel, by manually picking numbers from a hat or by spinning a roulette wheel. Through repeated sampling from this population the student learns in concrete terms the meaning of such subtle notions as: the Central Limit Theorem; probability; interval estimation; hypothesis testing and alpha and beta. In addition, the student is given an opportunity to hone their computer skills.

2. DETAILS OF THE SIMULATION PROJECT:

COMPUTER SIMULATION ASSIGNMENT: – Sampling distribution of the mean.

Simulation assignments and random numbers

Throughout the term you will be asked to do several simulation assignments. To do these projects you must generate random numbers in Excel using the Randbetween function.

There are 8 parts to this assignment. Please label them in your report to me as Part 0, Part 2, etc.

Part 0:

Characterize the population.

Using Excel’s Randbetween (0,9) function, generate 100 random numbers between 0 and 9, approximating the discrete uniform distribution.

Part 1:

Using Excel’s Randbetween (0,9) function, generate 200 samples of five random numbers between 0 and 9, calculate the mean of each sample. Show me the list of the 200 means. Typically, they should look like: 4.8, 3.6, 4.4, 6.0, etc.

Part 2:

Using Excel, calculate the overall mean of the 200 sample means (the average of the averages). This should be around 4.5.

Part 3:

Using Excel, calculate the standard error of the mean (SEM) (i.e. the standard deviation of the 200 sample means). We established in the previous simulation that the population average is 4.5 and the standard deviation of the population is 2.87.

Since the SEM= ‹ $\bar{\mathbf{X}}$ = σ/√n. and n=5, the SEM therefore is 1.28. Thus, the standard deviation of the 200 sample means should be approximately 1.28.

Part 4:

Using Excel, make the histogram of the 200 sample means (sampling distribution of the mean) (use interval size 1, i.e., 0-1, 1-2, 2-3, …8-9). According to the Central Limit Theorem a bell shaped curve should appear.

Part 5:

Discuss the intuitive logic of the Central Limit Theorem. Discuss the implications of part 4 in this context.

Part 6:

Use 2 methods to find P ($\bar{x}$ >6.3), (with n=5 as in Parts 1-4): First the z-method of chapter 7. Do it again by simply counting how many of your 200 $\bar{x}$ were above 6.3.

Part 7:

Discuss the standard error of the mean. Explain clearly the reasons why there is an “n” in the bottom of the formula. Do this by repeating Part 1 using an n of 10, 25 and 100.

3. **RESULTS**

Part 0: The population

3

1

0

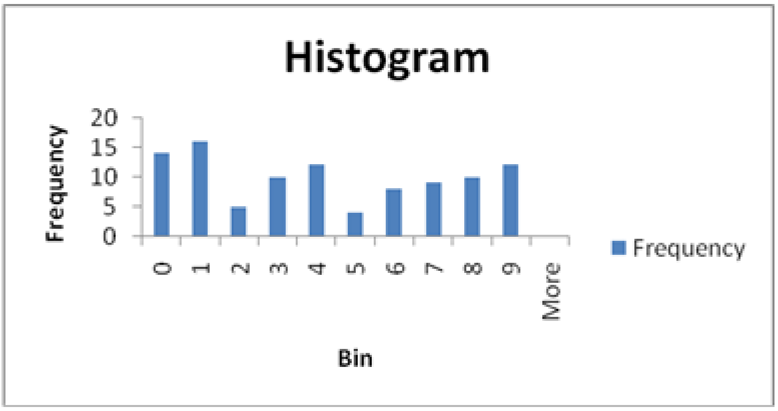
…

4

2

9

100 random numbers	
Mean	**4.23** (vs. ideal 4.5)
Standard Deviation	3.13 (vs. ideal 2.87)



Uniform distribution

Part 1: 200 means

4	0	6	3	7	4
1	5	5	8	3	4.4
8	5	2	5	5	5
4	6	0	1	4	3
2	9	0	5	7	4.6
0	0	5	0	5	2
…	…	…	…	…	…
9	3	5	7	6	6
0	5	3	3	0	2.2
7	2	2	9	0	4
8	8	4	8	5	6.6
9	9	4	3	7	6.4
1	9	1	8	7	5.2
					4.576
					1.267381

Part 2:

Mean=sum/200= 4.58

Part 3:

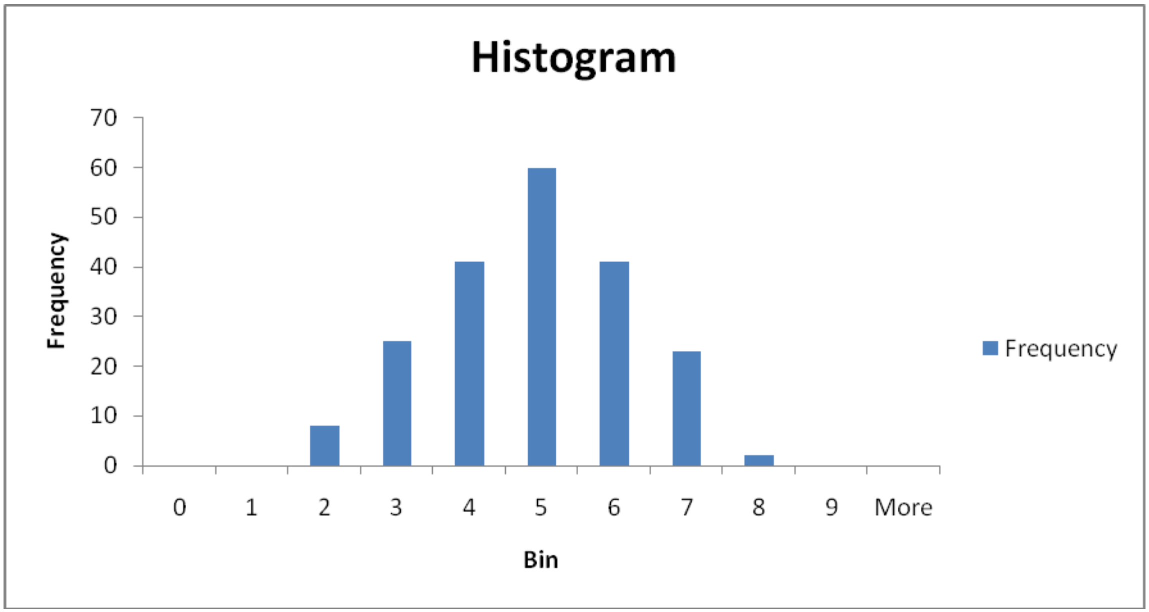
the SEM=$\sigma_{\bar{X}}$ = σ/√n

= 2.87/ 2.24

= 1.28

The 'empirical' standard deviation of the 200 means is 1.27

Part 4:



Why do we 'expect' to see a "bell shape" curve?

Part 5:

Discuss the intuitive logic of the Central Limit Theorem. Discuss the implications of part 4 in this context.

Central limit theorem predicts mathematically that it has to be a bell shaped curve under certain conditions. These conditions being the sample size has to be large enough for the sampling distribution of the mean to be approximately normally distributed. The CLT allows you to make inferences about the population mean without having to know the specific shape of the population distribution. As we can see from the last simulation, as we increased the sample size, the histogram looks more and more like a bell shaped graph.

Part 6 :

P ($\bar{X}$ >6.3),=

Theoretical:
z=(6.3-4.5)/1.28
= 1.41
=> probability=.9207

=1- .9207
=.079→ 7.9%

Counting the number of sample means that are over 6.3
=17/200= 8.5%

Part 7

The standard error of the mean, tells us how the sample means vary from the population mean. The standard error of the mean is equal to the standard deviation in the population divided by the square root of the sample size, n. So if the sample size (n) increases the standard error of the mean decreases by the square root of the sample size. In other words taking a larger sample, results in less variability from the population.

We see this by repeating the simulation using, the second time around, a sample of 10, then 25 and finally 100.

With a sample of 100, almost every mean is close to the population value of 4.4, so there is very little variability among them. Indeed, the SEM=2.87/√100=0.29.

4. **CONCLUSIONS**

Many students find the concepts discussed in introductory statistics to be abstract and difficult. In particular that section of the course called inferential statistics gives much trouble to the student who is not mathematically sophisticated. In the course of this project the student is able to verify that the predictions of the standard statistical formulas correspond to the results obtained from repeated sampling. This educational tool is specifically designed to explain the basic logic of inferential statistics and has been successfully tested over time. Most students have favorable thing to say about their learning experience.

In this presentation we will also explain why Monte Carlo simulation is a useful tool for teaching and learning statistics. The basics of descriptive statistics (e.g., mean, standard deviation, distributions, histograms) and introductory inferential statistics (e.g., confidence intervals, hypothesis testing) are all taught using a unified framework and an incremental and cumulative technique.