EXCEL SIMULATION AS A TOOL IN TEACHING SAMPLING DISTRIBUTIONS IN INTRODUCTORY STATISTICS

Leslie Chandrakantha
Department of Mathematics and Computer Science
John Jay College of Criminal Justice of CUNY, New York, USA
lchandra@jjay.cuny.edu

Many instructors are adopting computer simulation to introduce key concepts which many students in introductory statistics classes struggle to understand. Research has shown that the use of computer simulation methods as an alternative to traditional methods of books and lecture enhance conceptual understanding. Computer simulation using spreadsheets such as Excel allows students to experiment with data and to visualize the results. The concepts of sampling distributions are essential for the understanding of later inferential statistics topics such as confidence intervals and hypothesis testing. In this paper, we will describe how to use simulation using the Excel Data Tables facility and standard functions to teach sampling distributions in introductory statistics classes. This approach gives a better understanding of the Central Limit Theorem which describes the sampling distribution. Our preliminary assessment shows that this approach would enhance the understanding of the concepts.

INTRODUCTION

Statistics courses are increasingly becoming an integral part of all levels of higher education curriculum today. The Association of Advance Collegiate Schools of Business (AACSB) requires that undergraduate and graduate students exhibit proficiency in “statistical methods”. Statistics skills are necessary for conducting research, analyzing data, and making correct conclusions. Many students have difficulties in understanding basic statistical concepts such as sampling distributions, confidence intervals, and hypothesis testing. These concepts are critical in gaining necessary statistical skills. Research has shown that the use of computer simulation methods (CSM) in teaching and learning as opposed to traditional methods of books and lecture provide better understanding of concepts. Cobb (1994) noted that incorporating computer simulation techniques to illustrate key concepts and to allow students to discover important principles themselves will enhance their knowledge. Bell (2000) has shown that the use of Excel together with realistic examples can contribute greatly to students understanding of statistics. These methods allow students to experiment with data themselves and to visualize concepts. Mills (2002) has given a comprehensive review of literature of computer simulation methods used in all areas of statistics to help students understand difficult concepts. Butler, Rothery & Roy (2003) has developed macros that can run on Minitab environment for resampling methods in teaching statistics. Many introductory statistics students do not have the necessary skills to write or implement macros to perform these tasks. The EXCEL Data Table facility and standard functions can be used to accomplish the same task without writing macros. The Data Table is a tool that can perform many calculations at once, similar to calculating the value of a statistic based on many simulated random samples. Christie (2004) has used the Excel Data Tables for estimating population mean and correlation. A valuable introduction to Data Tables is given in Ecklund (2009).

In this paper, we describe how to use Excel standard functions and the Data Table facility to simulate random sampling from a population to generate the sampling distribution of the mean and to understand the Central Limit Theorem. The main advantages of Excel are that it is available to students and it presents the situation in multiple rows and columns. However, we are aware of the limitations of Excel. Many believe Excel is not a statistical data analysis package. Statistical methods such as creating boxplots and dotplots are not available in Excel. Excel does not handle missing data properly. But for our lessons of generating random numbers, calculating sample means, simulations, and creating histograms, Excel is adequate. In coming sections, we give an overview of sampling distributions, simulation of sampling distributions, meeting GAISE (2005) report recommendations, comparison of two teaching methods, and concluding remarks.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

OVERVIEW OF SAMPLING DISTRIBUTION OF MEAN

The gateway to statistical inferences is the sampling distributions. It is essential to gain a good understanding about the concepts of sampling distributions for students in introductory statistics classes. At the beginning of the class, we give the definition of the sampling distribution of the mean, introduce the properties, and explain the connection between sampling distributions and the central limit theorem. The sampling distribution of the mean is the probability distribution of sample mean based on all possible simple random samples of the same size from the same population. The sampling distribution of the mean has the following properties:

- The mean of all sample means is equal to the population mean.
- The standard deviation of the sample means (known as the standard error) is equal to the population standard deviation divided by square root of the sample size.
- Sample means are more normal than individual observations.

The central limit theorem explains the shape of the sampling distribution. This theorem tells that for a population of any distribution, the distribution of the sample mean approaches a normal distribution as the sample size increases. The larger the sample size, the better the approximation. Based on this theorem, we can use the normal distribution for inferences about the mean for larger sample sizes, even if the original population is not normally distributed. Many students are using this fact without understanding the underlying concept. Simulation allows students to visualize this fact.

SIMULATION OF SAMPLING DISTRIBUTION OF MEAN

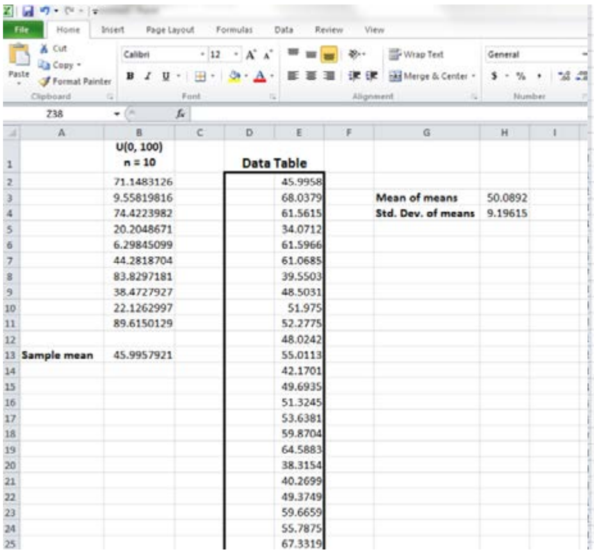
For this part of the lesson, we use the instructor’s computer and the projector to show the simulation of the sampling distribution to the class. The students have computers in the classroom so they follow our instructions and generate their own. We consider different population distributions and different sample sizes to observe the effects of the sample size and the shape of the original distribution on the sampling distribution of the mean. The uniform, chi-square, and normal populations and sample sizes of 10, 25, and 50 are considered. These three populations have uniform, skewed, and bell shapes so each student visualizes the fact that as the sample size increases the sampling distribution approximates a normal distribution for different original shapes. This is one of the main points that we want to convince our students as a result of this lesson. Students were taught these three populations, their shapes, and parameters in previous lessons so we just briefly explain them to refresh their memory and introduce the Excel formulas to generate random variants form these three populations. *Table 1* gives the characteristics of the samples and the Excel formulas.

Table 1: Sample characteristics

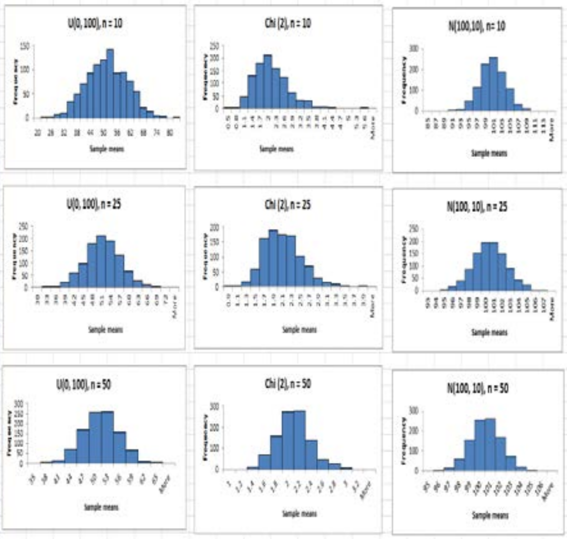
Distribution	Sample Sizes	Mean	Std Deviation	Excel Function
Uniform (0, 100)	10, 25, 50	50	28.87	= 100*rand( )
Normal (100, 10)	10, 25, 50	100	10	= norm.inv(rand(), 100, 10)
Chi-square (2)	10, 25, 50	2	2	= chisq.inv(rand(), 2)

Figure 1 shows a part of the spreadsheet implementation of generating random samples of 10 Uniform random numbers and computing the sampling distribution of the mean. In the cell B2, we write the formula **= 100*rand()** and copy this to cells B3:B11 to generate a random sample. In cell B13, we write the formula **= average(B2:B11)** to compute the sample mean. Then we use this value in cell B13 as the base value to create a Data Table to generate 1000 random samples. Our Data Table has the cell range D2:E1001. In the top right cell E2, write the formula **=B11** and leave the left column blank. Select the cell range D2:E1001, and use the commands sequence **Data > What-If Analysis > Data Table**. This will create the Data Table dialog box. Leave the Row input cell blank and type an empty cell reference (say **F1**) which is not part of this process for the Column input cell and click OK. This will fill the right column of the table with 1000 sample

means from different samples. Pressing the **F9** key will recalculate the table of different set of sample means.



	A	B	C	D	E	F	G	H	I
1		U(0, 100) n = 10		Data Table					
2		71.1483126			45.9958				
3		9.55819816			68.0379		Mean of means	50.0892	
4		74.4223982			61.5615		Std. Dev. of means	9.19615	
5		20.2048671			34.0712				
6		6.29845099			61.5966				
7		44.2818704			61.0685				
8		83.8297181			39.5503				
9		38.4727927			48.5031				
10		22.1262997			51.975				
11		89.6150129			52.2775				
12					48.0242				
13	Sample mean	45.9957921			55.0113				
14					42.1701				
15					49.6935				
16					51.3245				
17					53.6381				
18					59.8704				
19					64.5883				
20					38.3154				
21					40.2699				
22					49.3749				
23					59.6659				
24					55.7875				
25					67.3319				



In the cells **H3** and **H4** of the spreadsheet, we calculate the mean and standard deviation of the 1000 sample means. The 1000 sample means are considered as the sampling distribution of the mean to verify the validity of the properties the mean and standard deviation (standard error) of the sampling distribution. To study the shape of the sampling distribution, we create a histogram of the 1000 sample means. The steps of creating histograms using Excel Data Analysis tools have already been introduced in a previous lesson so student have skills for this part of the lesson. Similarly, we generate the sampling distributions and histograms for all the cases we have considered in this lesson. *Figure 2* shows histograms for all the cases. These histograms allow students to understand the meaning of the central limit theorem. Students will be able to visualize that as the sample size increases, the shape of the distribution is becoming more normal and the sample means are less variable. At the end of the lesson, students are able to understand the properties of the sampling distribution discussed in the class.

MEETING THE GAISE REPORT RECOMMENDATIONS

This activity of teaching sampling distributions using computer simulation methods meets three of the six recommendations suggested in the GAISE (2005) report of the American Statistical Association. The purpose of this report is to lay the foundation to help students achieve a goal of being sound statistically literate citizens who can apply the concepts well and think statistically. This report has revolutionized the way we teach introductory statistics. Our approach is somewhat closer to active learning in the classroom since students are experimenting and obtaining results by simulating the sampling distribution using Excel software. The following recommendations are met from our approach:

- *Stress conceptual understanding rather than mere knowledge of procedures:* This activity in the classroom stresses the importance of the understanding of randomness of the sampling distribution and that helps to learn the meaning of confidence intervals and *p*-values in hypothesis tests in future lessons.
- *Foster active learning in the classroom*: Students conduct the simulation and generate the sampling distribution themselves. They can verify the properties of the sampling distribution from the simulated results. The instructor asks questions to verify the students' understanding.
- *Use technology for developing conceptual understanding and analyzing data:* Students are using classroom computers and Excel software for this activity. They generate random samples, compute sample means and create histograms to visualize the concepts.

COMPARISON OF TWO TEACHING METHODS

We have selected two introductory statistics sections with the same instructor, one was taught using computer simulation methods (CSM) with Excel described in this paper and the other with traditional method of not using simulation. The assignment of students to each section was done by the registrar's office using normal registration process of first come first serve basis. Both classes have the same course content, same exams, quizzes, and assignments. *Table 2* shows the final exam statistics.

Table 2: Exam Score Statistics

Class	n	Mean	Median	Std. Dev.
CSM used	25	76.20	77	15.12
Traditional method used	23	68.61	68	14.34

The two sample t-test was performed using Excel to test the hypothesis that the CSM class performs better on average than the traditional method class. The *p*-value produced by Excel was 0.0408 which indicates that CSM class performed significantly better at 0.05 level of significance. We have to caution that these sample sizes are not large enough to make a firm judgment on the conclusion. We plan to use larger sample sizes in the future classes to make a formal assessment.

CONCLUSION

In this paper, we have demonstrated how to teach introductory statistics students the concepts of sampling distributions and the central limit theorem using Excel Data Tables and simulation. In particular, learning statistics concepts can be aided by doing rather than by reading or listening. This computer simulation method is an active learning approach which uses physical activities to demonstrate abstract concepts. These methods can also be used in more difficult lessons such as interval estimation and hypothesis testing that are taught after sampling distributions and the central limit theorem. Having introduced simulation methods early in the course would definitely help the instructor to teach effectively and the students to achieve better outcomes. Our preliminary assessment has shown that computer simulation approach would enhance the student learning of concepts.

REFERENCES

AACSB (2013). *Association to Advance Collegiate Schools of Business 2013 Accreditation Standards*. www.aacsb.edu/accreditation/business/standards/2013/learning-and-teaching/standard9.asp

Bell, P. C. (2000). Teaching Business Statistics with Microsoft Excel. *INFORMS Transactions on Education*, *1*(1), 18-26.

Butler, A., Rothery, P., & Roy, D. (2003). Minitab Macros for Resampling Methods. *Teaching Statistics*, *25*(1), 22-25.

Christie, D. (2004). Resampling with Excel. *Teaching Statistics*, *26*(1), 9-14.

Cobb, P. (1994). Where is the Mind? Constructivist and sociocultural perspectives on mathematical development. *Educational Researcher*, *23*, 13-20.

Ecklund, P. (2009). Introduction to Excel 2007 Data Tables and Data Table Exercises. faculty.fuqua.duke.edu/~pecklund/ExcelReview/Excel%202007%20Data%20Table%20Notes.pdf

GAISE (2005). *Guidelines for Assessment and Instruction in Statistics Education Report*. Alexandria, VA: American Statistical Association. www.amstat.org/education/gaise/

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education,* *10*(1), ww.amstat.org/publications/jse/v10n1/mills.html