Understanding Confidence Intervals and Hypothesis Testing Using Excel Data Table Simulation

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

Abstract: *Computer simulation methods have been used in upper level statistics classes for many years. Lately, many instructors are adopting computer simulation to introduce the concepts in the introductory level. Students in introductory statistics classes struggle to understand the basic concepts. Research has shown that the use of computer simulation methods as an alternative to traditional methods of books and lecture enhance the understanding of the concepts. Computer simulation using spreadsheets such as Excel allows students to experiment with data and to visualize the results. In this paper, we will describe how to use the simulation using Excel Data Tables facility and standard functions to teach confidence intervals and hypothesis testing in introductory statistics classes. We believe, by using this hands on approach, students get a better feel for these abstract concepts. Our preliminary assessment shows that this approach would enhance the student learning of the concepts.*

1. Introduction

The understanding of the statistical inferences concepts is critical in making accurate conclusions in research findings. Almost all college students will have to do some form of research and summarize their findings. Statistics courses give the necessary skills for these tasks. College students will take at least one statistics course to gain such critical skills. The introductory statistics is the first statistics course they normally take at college level. Fundamental statistical concepts such as sampling distributions, central limit theorem, confidence intervals, hypothesis testing, and *p*-values are vital in an introductory statistics course. Many students struggle to understand these concepts. The use of computer simulation methods to mimic the real life sampling or repeated sampling from a population helps to understand these concepts. Cobb [3] noted that incorporating computer simulation techniques to illustrate key concepts and to allow students to discover important principles themselves will enhance their knowledge. Almost all software packages offer ways to perform the simulation. Mills [7] has given a comprehensive review of literature of computer simulation methods used in all areas of statistics to help students understand difficult concepts. Butler et al. [1] 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. Excel provides ways to accomplish the same task without writing macros. Furthermore Excel is preferred due to the availability to students and the ease of presenting the data for the underlying statistical concepts in multiple rows and columns.

In this article, we describe how to use the Excel standard functions and the Data Table facility to generate different random samples from a population, compute confidence intervals, and perform hypothesis tests using repeated simulated sampling. Using this approach in the classroom, we are directly incorporating several recommended guidelines in the GAISE report [9] of American Statistical Association. The simulation and resampling have been used in teaching statistics for many years. Hagtvedt et al. [6] have developed simulation tools to compute multiple confidence intervals. Christie [2] has used the Excel Data Tables for estimating the population mean and the correlation. Winston [9] explained how to use Excel Data tables to simulate stock prices in asset allocation models. A valuable introduction to Excel Data Tables is given by Ecklund [4].

In next section, we give an introduction to Excel Data Tables and how to use them in generating different random samples. The next two sections will show how to use these simulation methods to teach the lessons of confidence intervals and hypothesis testing. Next, we illustrate how this approach meets the guidelines of the GAISE report. Finally, we give a preliminary comparison of traditional method of teaching and computer simulation approach, and our concluding remarks.

2. Excel Data Tables

Data Tables are part of a group of commands that are called what-if analysis tools in Excel. Data Table function allows a table of "what if" questions to be posed and answered simply in sensitivity analysis, and is useful in simulation. What-if analysis is the process of changing the values in cells to see how those changes will affect the outcome of formulas on the worksheet. We can use the Data Table function to compute the values of a test statistic for different random samples. The Data Table function can be accessed from menu bar Data > What IF Analysis > Data Tables in Excel 2007 and 2010.

To generate the values of a statistic for different samples using Data Table, first we calculate the value of the statistic using a random sample. This can be done using Excel random number generating functions for the appropriate population and other standard functions. This statistic value will be our original input value. This value (formula) will be put in the top cell of the right column of the Data Table. We set up our Data Table by selecting two columns and a certain number of rows depending on the number of values of the statistic we need to calculate. Leave the left column blank. The menu bar *Data > What IF Analysis > Data Tables* gives the Data Table dialog box. In this dialog box, leave Row input cell blank and type an empty cell reference that has no part of this Data Table setup for the Column input cell. Excel generates a new random sample and computes the value of the statistic for each substitution of this empty input cell and fills the table. Copying the formula down in the output cells does not work in this case. If we do this manually, we need to recalculate the statistic by repeated sampling by pressing the F9 key and recording these values in a column. The Data Table function provides a convenient way of generating values of the statistic for different samples.

3. Simulation of Confidence Intervals

The topic of confidence intervals of parameter estimation is a difficult lesson to teach in introductory statistics classes. Confidence intervals give the most likely range of the unknown population parameter. In this discussion, we only consider the creating and interpreting confidence interval for mean (μ) assuming population standard deviation (σ) is known. Beginning of the class, we introduce the background and define the confidence interval for the mean as $\bar{X} \pm Z^* \sigma/\sqrt{n}$, where Z* is the value of the standard normal curve with area C between critical points –Z* and Z* and *n* is the sample size. *The confidence level C is the probability that the confidence interval actually does contain the population mean μ, assuming the estimation process is repeated a large number of times*, Moore [8]. Students have major difficulty in understanding this last statement. Many students misunderstand this statement as that majority of individual values are in this interval. It is important in this lesson that students understand in repeated sampling form a population, C percent of intervals (say 95%) would capture the true unknown mean. In using the traditional way of teaching, we only consider one sample and calculate one interval. This leads them to believe the wrong interpretation of the interval that there is a 95% chance that this interval will have the true mean.

A computer simulation method using Excel will allow students to understand the true meaning of the confidence interval. After introducing the basic facts about confidence intervals to the class, we will calculate 95% confidence intervals for the population mean. First, we will show how to generate one random sample from the normal distribution with an assumed mean (say 100) and assumed standard deviation (say 10) of a convenient size (say 30). That is μ = 100, σ = 10, and *n* = 30. The Excel formula = ***norminv(rand(), 100, 10)*** is typed in a cell and copied using the fill handle to cells below in the column to generate a random sample. In the class room, everybody has computers and students follow our instructions and create their own samples. Then we calculate the mean of this sample using the ***average*** function in Excel, and use this as the top value to generate another 999 sample means using Excel Data Table. Now we have a table with 1000 sample means. For each sample mean, the confidence interval bounds are calculated using the formula $\bar{X} \pm Z^* \sigma/\sqrt{n}$. The Excel formulas for these are= ***E2 - 1.96*10/sqrt(30)*** and **= *E2 + 1.96*10/sqrt(30)*.** The cell ***E2*** contains the sample mean of the first cell of the Data Table. Then we copy these formulas to generate the rest of the confidence interval bounds. Finally, the proportion of intervals contain the true mean μ of 100 is calculated using ***IF*** and ***COUNTIF*** functions. If the interval contains the true mean, we assign one and otherwise zero using **= *IF(AND(100>=G2,100<=H2),1,0)*** formula and then compute the proportion of ones using **= *COUNTIF(J2:J1001,1)/1000*** formula. This proportion should be closer to 95% which is the assumed confidence level C. Since students are doing these steps themselves, they visualize each step and get a clear understanding of the meaning of confidence intervals. *Figure 3.1* shows a portion of the spreadsheet implementation of this simulation.

	A	B	C	D	E	F	G	H	I	J	K
1		N(100, 10) Sample of n = 30		Sample Means (Data Table)			95% CI Lower Bound	95% CI Upper Bound		Does CI contain the mean? (1-yes, 0-no)	Proportion of samples contained the true mean
2		117.8964765			98.948		95.36955032	102.5264584		1	0.95
3		105.1381437			100.207		96.62850182	103.7854099		1	
4		102.1722608			97.30691		93.72845208	100.8853602		1	
5		111.6343039			98.18426		94.6058025	101.7627106		1	
6		109.4543404			96.62267		93.0442167	100.2011248		1	
7		94.90576309			98.79253		95.21408078	102.3709889		1	
8		105.056665			97.60669		94.02824072	101.1851488		1	
9		91.97505739			99.66016		96.08170938	103.2386175		1	
10		97.70580237			97.34671		93.76825241	100.9251605		1	
11		97.98155704			101.0588		97.48029972	104.6372078		1	
12		96.76448788			98.84559		95.26713865	102.4240467		1	
13		98.94094087			96.0825		92.50404966	99.66095774		0	
14		90.20739352			99.77196		96.19350545	103.3504135		1	
15		108.5321596			99.38756		95.80910821	102.9660163		1	
16		106.1843228			98.72426		95.1458084	102.3027165		1	
17		74.35549658			99.32286		95.74440286	102.9013109		1	
18		112.7895912			101.9586		98.38015195	105.53706		1	

Figure 3.1 Confidence Intervals Calculation Spreadsheet

Notice that 95% of the confidence intervals do contain the true mean. If we generate another set of samples and compute the confidence intervals by pressing the **F9** key, we will find the new proportion which should be 95% or closer to it. Changing the confidence levels (to 90% or 99%) by just changing the confidence coefficient Z* to appropriate values in the formulas will help students to understand the meaning of confidence intervals. This lesson can be easily done in one class period of 75 minutes.

4. Simulation of Hypothesis Testing

Hypothesis testing is another key topic of inferences in statistics classes and yet it is one of the abstract concepts to understand. A sound knowledge about terms such as null and alternative hypotheses, significance level, and *p*-value is essential in performing a hypothesis test and making the correct decision. Since all statistical software calculate *p*-values, more and more instructors are using the *p*-value approach to make decisions. Many students do not have a good understanding about the *p*-value and they blindly use the rejection criterion that "if the *p*-value is less than the significance level, rejects the null hypothesis". The definition of the *p*-value is *the probability, assuming the null hypothesis is true, that the test statistics would take a value as extreme or more extreme than that actually observed*, Moore [8].

Computer simulation methods can give a better understanding of the *p*-value because it generates many samples assuming null hypothesis and then one can find the proportion of the test statistic values that are extreme or more extreme than the actual value observed from sample data.

Erickson [5] has used Fathom Dynamic Data Software to simulate flipping coins in teaching hypothesis testing concepts and found out that it makes difficult concepts more visible and understandable. Now we describe how to use Excel Data Tables to simulate hypothesis testing and compute the corresponding *p*-value. At the beginning of the class, we introduce the key components of hypothesis tests for mean μ assuming population standard deviation σ is known. Then we use an example to set up the null and alternative hypothesis, find the value of the statistic based on the actual sample, generate many random samples assuming null hypothesis, compute the test statistic values, and compute the *p*-value by finding the proportion of values exceeding the actual value. If the *p*-value is too small (less than the significance level, normally 0.05), we have evidence against the null hypothesis. The following example is discussed in the class.

EXAMPLE: One sample z test. Data for this example are the reading times, Moore [8].
Does the use of fancy type fonts slow down the reading of text on a computer screen? Adults can read four paragraphs of text in an average time of 22 seconds in the common Times New Roman font. 25 adults were asked to read this text in the ornate font named Gigi. Here are their times:
23.2, 21.2, 28.9, 27.7, 23.4, 27.3, 16.1, 22.6, 25.6, 32.6, 23.9, 26.8, 18.9, 27.8, 21.4, 30.7, 21.5, 30.6, 31.5, 24.6, 23.0, 28.6, 24.4, 28.1, 18.4.
Suppose that reading times are normally distributed with σ = 6 seconds. Is there good evidence that the mean reading time for Gigi fonts is greater than 22 seconds? In other words, is μ greater than 22 seconds for Gigi fonts?

The null and alternative hypotheses are: H_0: $\mu = 22$ seconds and H_1: $\mu > 22$ seconds. The test statistics used for this test is $(\bar{x} - \mu)/\frac{\sigma}{\sqrt{n}}$ and that follows the standard normal distribution. The value of the test statistics based on the above sample assuming null hypothesis is 2.627. To find the *p*-value using simulation, we show how to generate one sample of 25 values from normal distribution with a mean of 22 and a standard deviation of 6 and calculate the value of the test statistic. Then use this value as the top value of the Data Table to generate 999 more random samples from the same distribution and calculate value of the test statistic. Now the Data Table has 1000 values of the test statistic. Then we use the Excel formula = ***COUNTIF(G2:G1001, ">= 2.627")/1000*** to find the proportion of those values which exceed the actual value observed (2.627). This will allow students to understand that if the null hypothesis is true, and if we repeat the process large number of times, this is the chance that the value of the statistic is extreme or more extreme. Students will realize that if this chance is too small, the null hypothesis is unlikely. Since students are also doing this in the classroom, they clearly experience the process themselves. *Figure 4.1* shows a portion of the spreadsheet implementation of this simulation.

	A	B	C	D	E	F	G	H
1	N(22,6) Sample of n= 25						Value of Test Statistic (Data Table)	
2	14.85214246						-0.06122	
3	17.00157901		Sample Mean	21.92653			0.334412	
4	20.44798115		Value of the Test Statistic	-0.06122			0.599874	
5	20.88775936						1.634011	
6	28.6507339						1.788618	
7	32.88832421		p-value	0.004			-1.44148	
8	17.79692507						1.107673	
9	20.10462593						0.632176	
10	12.71772439						0.843168	
11	18.12509896						-0.85154	
12	21.38762783						-0.52319	
13	24.59459382						1.243032	
14	24.4754052						0.526319	
15	27.16582715						-0.49738	
16	22.17604272						-1.25612	
17	12.77279913						0.804571	
18	20.41276222						-0.25105	
19	29.42669933						0.80136	
20	16.11601878						0.275278	
21	29.25147382						-1.05186	
22	15.93632039						1.220461	
23	28.68634823						-1.68135	
24	29.61035276						-0.09151	
25	20.56958745						0.123601	
26	22.10859323						0.850515	
27							0.473803	

Figure 4.1 *p*-value Calculation Spreadsheet

This *p*-value is clearly equal to the exact value calculated from standard normal distribution {P(Z > 2.627) = 0.004}. Rather than using the normal distribution table to find the *p*-value, this method explains the process better. The empirical distribution of the test statistic is also tabulated in the class and it shows the values are approximately standard normally distributed as it should be. *Figure 4.2* shows the histogram.

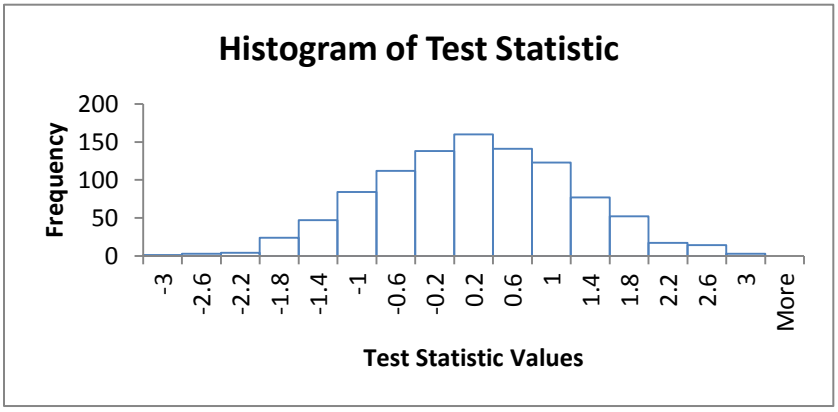

Figure 4.2 Histogram of the Test Statistics Values

5. Meeting GAISE Report Recommendations

The activity of teaching sampling distributions using computer simulation meets three of the six recommendations suggested in the GAISE report [9] 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 the active learning in the classroom since students are experimenting and obtaining the results by simulating the sampling distribution using Excel software. The following recommendations are met from this approach:

a) **Stress conceptual understanding rather than mere knowledge of procedures**. Sampling distribution is the key to understand the topics of statistical inferences such as confidence intervals and hypothesis testing. 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.
b) **Foster active learning in the classroom.**
While we show how to do it, students conduct the simulation and generate the sampling distribution themselves. They can verify the properties of the sampling distribution of the mean and the test statistics using the simulated results. Regularly we ask questions during the lesson to see whether students understand the concepts.
c) **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 confidence intervals, values of test statistics and corresponding p-values to understand the underline concepts of these topics. This will visually illustrate the abstract concepts and enhance the conceptual understanding.

6. Comparison of Two Methods

We have taught two introductory statistics sections last semester, one using computer simulation methods with Excel described in this paper and the other by the traditional method of not using simulation. Both classes have the same course content, same exams, quizzes, and assignments. *Table 6.1* shows the final exam scores statistics:

Table 6.1 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 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 performs 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.

7. Conclusions

Many students have difficulties understanding introductory statistics concepts such as confidence intervals and hypothesis testing. Statistics instructors are always searching for new and efficient teaching methods to improve statistics instruction in hopes of enhancing student learning. Computer simulation methods as teaching tools are considered to be effective methods. We have demonstrated the use of simulation using Excel and Data Tables in teaching these topics. This is a very useful way to visualize the sampling distribution, confidence intervals, and to comprehend the *p*-value. Preliminary comparison of the two methods showed better outcomes using computer simulation methods. These simulation methods are acceptable to students with varying backgrounds of mathematics.

References

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

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

[3] Cobb, P. (1994). Where is the Mind? Constructivist and Sociocultural Perspectives on Mathematical Development. *Educational Researcher*, 23, 13-20.

[4] Ecklund, P. (2009). Introduction to Excel 2007 Data tables and Data Table Exercises. https://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm.

[5] Erickson, T. (2007). Using Simulation to Learn About Inferences. *International Conference on Teaching Statistics*, 7, 1-6.

[6] Hagtvedt, R., Jones, G. T., & Jones, K. (2008). Teaching Confidence Intervals Using Simulation, *Teaching Statistics*, 30 (2), 53-56.

[7] Mills, J. D. (2002). Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. *Journal of Statistics Education (Online)*, 10 (1). http://www.amstat.org/publications/jse/v10n1/mills.html.

[8] Moore, D. S. (1996). *Essential Statistics.* New York, USA: W. H. Freeman & Company.

[9] Winston, W. L. (2007). Excel *2007, Data Analysis and Business Modeling*: Microsoft Press.