

November 2015

A Spreadsheet Simulation to Teach Concepts of Sampling Distributions and the Central Limit Theorem

Mark H. Haney

Robert Morris University School of Business, haney@rmu.edu

Follow this and additional works at: <http://epublications.bond.edu.au/ejsie>



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Recommended Citation

Haney, Mark H. (2015) A Spreadsheet Simulation to Teach Concepts of Sampling Distributions and the Central Limit Theorem, *Spreadsheets in Education (eJSiE)*: Vol. 8: Iss. 3, Article 3.

Available at: <http://epublications.bond.edu.au/ejsie/vol8/iss3/3>

This Regular Article is brought to you by the Bond Business School at [ePublications@bond](mailto:EPublications@bond). It has been accepted for inclusion in *Spreadsheets in Education (eJSiE)* by an authorized administrator of [ePublications@bond](mailto:EPublications@bond). For more information, please contact [Bond University's Repository Coordinator](#).

A Spreadsheet Simulation to Teach Concepts of Sampling Distributions and the Central Limit Theorem

Abstract

This paper presents an interactive spreadsheet simulation model that may be used to help students understand the concept of sampling distributions and the implications of the central limit theorem for sampling distributions. The spreadsheet model simulates an approximation to a sampling distribution by taking 1,000 random samples from a population, calculating the mean of each sample, and then using percentage polygons to display the distribution of the sample means compared to the distribution of the population. A normal probability plot of the sample means is also created as a second tool for understanding the distribution of the sample means. The user may vary the size of the samples taken, and then observe the effects of sample size on the range and shape of the approximated sampling distribution. The spreadsheet model is built without macros or VBA programming, using only standard formulas and tools. The instructor may choose to build the model with students, or simply present it to them and lead them in experimenting with it, depending on the needs of the class.

Keywords

spreadsheet, simulation, sampling distribution, central limit theorem, statistics education

Distribution License



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

A Spreadsheet Simulation to Teach Concepts of Sampling Distributions and the Central Limit Theorem

Abstract

This paper presents an interactive spreadsheet simulation model that may be used to help students understand the concept of sampling distributions and the implications of the central limit theorem for sampling distributions. The spreadsheet model simulates an approximation to a sampling distribution by taking 1,000 random samples from a population, calculating the mean of each sample, and then using percentage polygons to display the distribution of the sample means compared to the distribution of the population. A normal probability plot of the sample means is also created as a second tool for understanding the distribution of the sample means. The user may vary the size of the samples taken, and then observe the effects of sample size on the range and shape of the approximated sampling distribution. The spreadsheet model is built without macros or VBA programming, using only standard formulas and tools. The instructor may choose to build the model with students, or simply present it to them and lead them in experimenting with it, depending on the needs of the class.

Keywords: sampling distribution, central limit theorem, simulation, spreadsheet, statistics education

1. Introduction

Students in introductory statistics courses often struggle to understand fundamental statistical concepts such as the nature of randomness and sampling distributions. Instead of focusing on understanding such fundamental concepts, many students concentrate on memorizing the mechanics of performing various calculations and statistical procedures. This approach, however, may not lead to effective learning, as students tend to forget the mechanics of calculations and procedures soon after the material is tested or the course ends [7]. A solid understanding of fundamental statistical concepts, sometimes termed “statistical literacy,” would be a far more enduring and valuable outcome of the statistics course, but how can such an understanding be achieved?

In order to improve student learning outcomes the American Statistical Association has published Guidelines for Assessment and Instruction in Statistical Education (GAISE). These guidelines recommend approaches to teaching and learning statistics that involve active learning of concepts rather than focus on mechanical calculations. Simulations, performed both manually and with the help of a computer, are recommended in the GAISE report as useful tools to involve students in a manner that helps them learn important statistical concepts [1]. Indeed, in recent years there has been much interest in the use of simulations as an aid to teaching and learning fundamental statistical concepts. For examples of the use of simulations to teach topics in statistics and econometrics see [3], [4], [5], [6], and [9]. For a review of the use of computer simulation methods to teach statistics see Mills, 2002 [10].

Spreadsheets are increasingly being used as tools for math and statistics education. For a review of the uses of spreadsheets in education see Baker and Sugden, 2003 [2]. Many introductory business statistics courses are now using spreadsheets, usually MS Excel, as the base environment to perform the statistical calculations in the course. Spreadsheets are familiar and accessible to business students, and they

provide a good environment in which to manipulate and visualize data. Spreadsheets also provide an effective “sandbox” for performing simulations that help students test out sampling scenarios and actively engage important statistical concepts. They allow students to enjoy the learning benefits of simulations without having to use custom programming or the complex simulation tools provided in advanced statistical packages.

This paper introduces an interactive spreadsheet simulation model for investigating sampling distributions and the central limit theorem. Instead of introducing a sampling distribution as an abstract concept, the simulation model allows students to learn by doing and discovering. Through the process of building a model that samples from a population many times, calculates the mean for each sample, and then displays characteristics of the sample means, students learn what sampling distributions are and how they are affected by the size of the samples and the distribution of the population from which the samples are drawn. When students encounter more theoretical discussions of sampling distributions they will be able to tie them back to the concrete experience of building and manipulating the simulation model.

The simulation model is built using standard Excel tools and functions, and does not require the use of macros, VBA programming, or proprietary add-ins. This helps make the model more accessible for beginning statistics students, in particular students of business. Although there are numerous applets and other animations available that instructors can use to demonstrate principles of sampling distributions and the central limit theorem, students may learn and understand the concepts better when they build the model themselves and then use the model to actively experiment and test their assumptions. While acknowledging that there is little empirical research testing the effectiveness of this approach, Mills [10] notes that many researchers of statistics education recommend an active approach to learning statistical concepts through computer simulation models, and that such an approach is in line with constructivist theories of learning.

In addition to helping increase student understanding of statistical concepts, building or studying the spreadsheet model is also a useful way for students to improve their spreadsheet skills, an important outcome for business students. It should be noted that in courses in which the development of spreadsheet skills is not an important goal, or in which students already have strong spreadsheet skills, a similar simulation model could be built more quickly by using available add-ins, such as the Analytic Solver Platform from Frontline Systems, Inc.

The following sections present the simulation model, how it may be used in class for exploration and learning, and how to build it. Depending on the goals and available time in the course in which the model is used, instructors may choose to build it together with students, assign parts of the model construction as assignments, or simply present the model to students and use it to explore properties of sampling distributions and the central limit theorem.

2. Introduction to the simulation model

The spreadsheet model simulates repeated sampling from a population to generate an approximation to a sampling distribution. It is divided into two worksheets. The user interface is the “Interface” worksheet, which provides a cell in which the user may enter the desired sample size, and which displays graphs that help students visualize the distribution of the population and the sample means. A second worksheet, named “Calculations” is where the calculations that perform the sampling and the simulation are implemented.

Figure 1 illustrates the “Interface” worksheet. This worksheet serves as the model’s interactive interface. The user enters the desired sample size into the yellow input cell and then presses the F9 key to cause all formulas in the workbook to recalculate. This recalculation causes the random sampling from the population to re-occur 1,000 times and the sample mean to be recalculated 1,000 times. The distribution of the resulting set of sample means is represented in a percentage polygons graph, on which the percentage polygon for the population is also displayed so that the distribution of the sample means may be compared to the distribution of the population. A normal probability plot of the sample means is also displayed as a second tool to help understand their distribution. Because sampling 1,000 times approximates a sampling distribution, building and using the model allows students to learn by discovery some basic properties of sampling distributions and the central limit theorem. In addition, the user can investigate relationships among the population mean and standard deviation, the expected mean and standard deviation (standard error) of the sampling distribution, and the actual mean and standard deviation of the set of 1,000 samples, which approximates the sampling distribution.

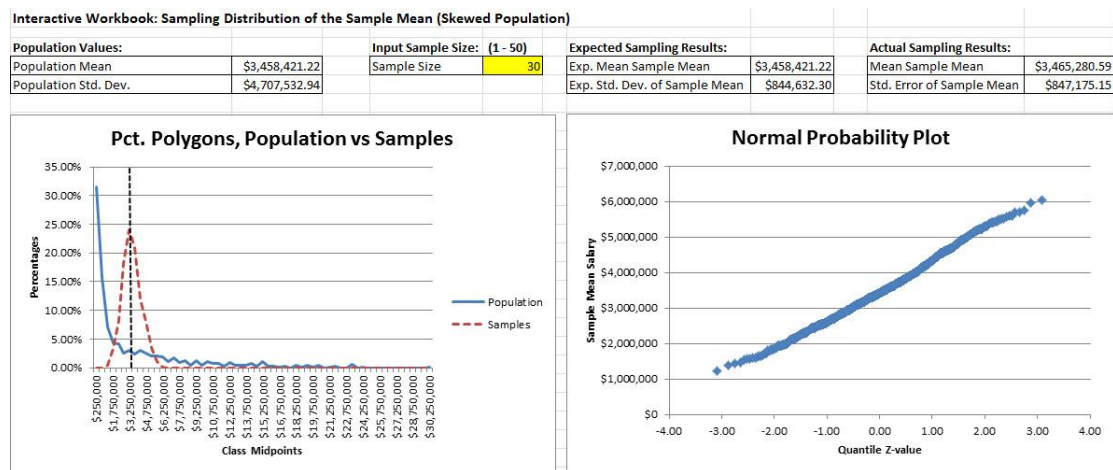


Figure 1: “Interface” worksheet

Figure 2 illustrates the “Calculations” worksheet. Note that the “Calculations” worksheet has over 1,000 rows. Figure 2 shows the top part of the worksheet to demonstrate its structure. The left side of the “Calculations” worksheet holds the population data. Then, to the right of the population data is a section in which one random sample is drawn from the population and its mean calculated. To the right of that is the simulation data table that causes 1,000 such samples to be drawn and their means recorded. The quantile z-values next to the sample means are used to create the normal probability plot on the “Interface” worksheet. Finally, the section to the

far right creates a distribution table by counting how many of the population values and sample mean values fall within various ranges, and then expressing those counts as percentages. This table is used as the base for the percentage polygons that are shown on the “Interface” worksheet. The model and the techniques used to build it will be described in more detail in sections 5-7.

Population Data		Drawing one Sample		Simulation of 1,000 Samples			Calculation of Percentages for Polygons						
MLB Salaries	Random #	Sample mean:	\$3,544,130.00	Trial #	Sample Mean	quantile z-value	Percentage Distribution Table						
							At least	Less Than	midpoint	Pop. count	Pop. pct.	Sample mean count	Sample mean pct.
\$12,350,000	0.42732535			1	\$2,292,687	-1.45							
\$10,000,000	0.73647696	Sample:		2	\$2,364,577	-1.35	\$0	\$500,000	\$250,000	267	31.49%	0	0.00%
\$7,833,333	0.71072376	1	\$5,400,000	3	\$2,594,992	-1.01	\$500,000	\$1,000,000	\$750,000	134	15.80%	0	0.00%
\$7,416,666	0.73782552	2	\$4,750,000	4	\$3,526,926	0.13	\$1,000,000	\$1,500,000	\$1,250,000	60	7.08%	4	0.40%
\$6,150,000	0.81714346	3	\$23,000,000	5	\$3,318,391	-0.13	\$1,500,000	\$2,000,000	\$1,750,000	37	4.36%	34	3.40%
\$5,200,000	0.85364167	4	\$5,500,000	6	\$2,208,280	-1.56	\$2,000,000	\$2,500,000	\$2,250,000	35	4.13%	83	8.30%
\$4,750,000	0.60093445	5	\$502,500	7	\$4,864,054	1.51	\$2,500,000	\$3,000,000	\$2,750,000	22	2.59%	181	18.10%
\$3,600,000	0.49703109	6	\$494,500	8	\$2,965,003	-0.57	\$3,000,000	\$3,500,000	\$3,250,000	25	2.95%	240	24.00%
\$3,072,000	0.04351305	7	\$490,600	9	\$2,726,217	-0.84	\$3,500,000	\$4,000,000	\$3,750,000	20	2.36%	210	21.00%
\$2,625,000	0.63324171	8	\$6,845,000	10	\$2,847,174	-0.71	\$4,000,000	\$4,500,000	\$4,250,000	25	2.95%	120	12.00%
\$1,500,000	0.19795477	9	\$750,000	11	\$4,433,069	1.08	\$4,500,000	\$5,000,000	\$4,750,000	22	2.59%	81	8.10%
\$1,450,000	0.53291192	10	\$481,300	12	\$1,949,394	-1.88	\$5,000,000	\$5,500,000	\$5,250,000	17	2.00%	35	3.50%
\$1,350,000	0.43668176	11	\$487,500	13	\$3,196,508	-0.28	\$5,500,000	\$6,000,000	\$5,750,000	18	2.12%	11	1.10%
\$1,300,000	0.00965241	12	\$1,537,500	14	\$4,057,815	0.74	\$6,000,000	\$6,500,000	\$6,250,000	16	1.89%	1	0.10%
\$1,250,000	0.39836856	13	\$483,000	15	\$3,751,399	0.41	\$6,500,000	\$7,000,000	\$6,750,000	9	1.06%	0	0.00%
\$1,000,000	0.29015060	14	\$493,500	16	\$4,449,871	1.10	\$7,000,000	\$7,500,000	\$7,250,000	15	1.77%	0	0.00%
\$825,000	0.03103568	15	\$6,000,000	17	\$3,887,737	0.55	\$7,500,000	\$8,000,000	\$7,750,000	8	0.94%	0	0.00%

Figure 2: “Calculations” worksheet

3. Using the model to explore sampling distributions

We build the model in class and use it to explore sampling distributions before going into a theoretical discussion of sampling distributions and their properties. The goal is for students to discover the main characteristics of sampling distributions on their own, rather than simply being told what they are. The key points we hope that students learn from this discovery exercise are the following:

- A sampling distribution is the distribution of a sample statistic that results when a large number of samples are taken from a population and the sample statistic calculated for each one
- The mean sample mean is a good estimator of the population mean
- With very small sample sizes the sampling distribution resembles the distribution of the population
- With a normally distributed population, as the sample size increases the sampling distribution remains approximately normally distributed, and the standard error of the sampling distribution decreases
- Even with a highly skewed population, as the sample size increases the sampling distribution moves closer to a normal distribution (central limit theorem), and the standard error of the sampling distribution decreases

The first point is reinforced to students as they build or study the simulation model. They discover the last several points by testing the model with a skewed population and a normally distributed population, with a range of sample sizes. For example, figure 3 illustrates the results when sample size is set to 1 and the population is right-skewed. The polygon graph shows that the distribution of the sample means is almost identical to the distribution of the population. The normal probability plot shows that the sample means are not even close to being normally distributed; the

closer a normal probability plot is to being a straight line, the closer the data is to being normally distributed. The standard error of the sample means is almost the same as the standard deviation of the population.

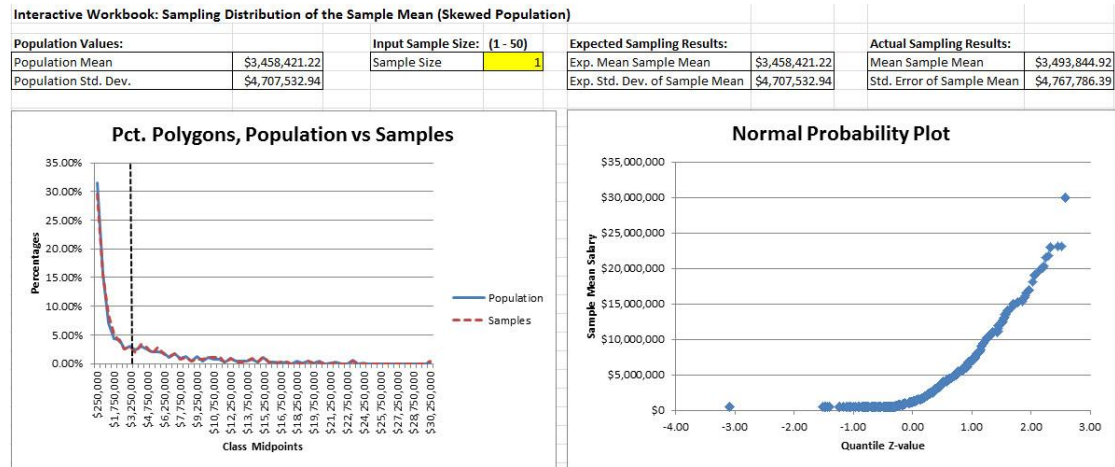


Figure 3: Sampling from a right-skewed population with sample size 1

When the sample size is set to 5, as in figure 4, the distribution of the sample means starts to resemble a bell-shaped normal distribution, and the normal probability plot becomes straighter. Moreover, the standard error of the sample means becomes smaller relative to the standard deviation of the population.

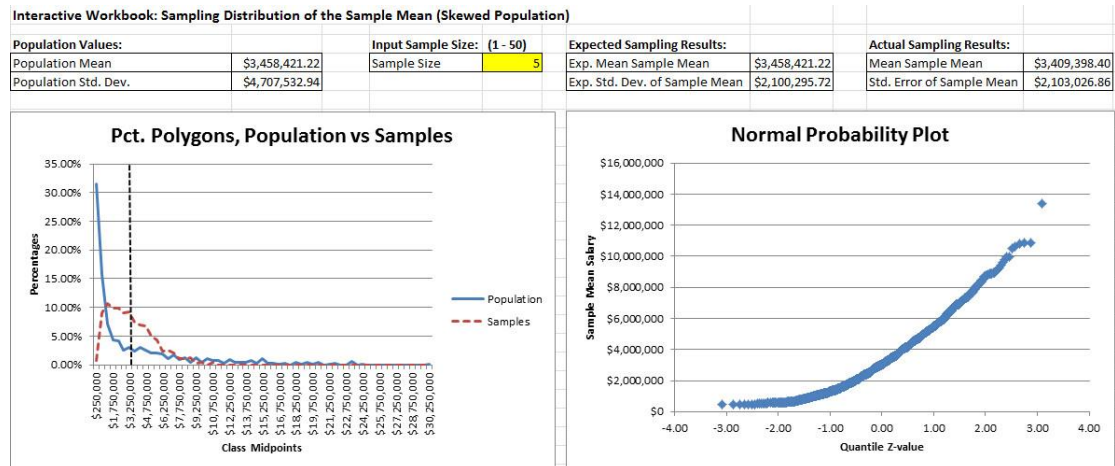


Figure 4: Sampling from a right-skewed population with sample size 5

Students continue to try different sample sizes. When they reach a sample size of 30 the polygons and normal probability plot suggest that the distribution of the sample means is now approximately normal. Figure 1 at the beginning of the section shows the polygons and normal probability plot for a sample size of 30.

The next several sections describe how to build the simulation model in Excel. At our university the model is used in the elementary business statistics course, which is a required course for all business students. One of the objectives of this course is for students to develop spreadsheet skills, and building the model contributes to that goal. We build the model near the middle of the semester. At that point in the course students have already learned the Excel techniques and functions used to build several sub-parts of this model, such as how to construct a frequency/percentage distribution table and then build a polygon based on the table, and how to create a

normal probability plot. Nevertheless, building the model requires the use of several fairly sophisticated Excel functions and techniques, such as dynamic ranges built with the INDIRECT function and concatenation, the RANK function, the RAND function, the INDEX function, and the data table tool. For students, the investment in Excel skills is not wasted. We use similar techniques later in the course to investigate hypothesis testing via spreadsheet simulation, and the Excel skills may come in handy in other courses students take, as well as in their future careers.

4. Data for the simulation model

The first step in building the model is to select data to serve as the population. When we build the model in class we typically use datasets that are already familiar to the students from investigations they have performed earlier in the semester. The example demonstrated in this paper uses player salaries from the 2012 Major League Baseball (MLB) season. Students in the class are already familiar with this dataset because we use it to illustrate various concepts in descriptive statistics, such as measures of central tendency (mean, median), measures of variation (standard deviation, coefficient of variation), and representations of distribution shape (skewness, 5-number summary, boxplots, histograms). This is a good dataset for demonstrating the central limit theorem because it is strongly right-skewed. Students are usually surprised that the simulated sampling distribution is approximately normally distributed at larger sample sizes.

To demonstrate sampling from a population that is approximately normally distributed we use 2012 MLB batting average data. Students are familiar with this dataset because they have used it earlier in the course to practice tests of normality. By doing tests on this dataset students discover that batting averages for all players with one or more at bats deviate significantly from the normal distribution, but batting averages for players with 100 or more at bats are very close to being normally distributed. To build the sampling distribution model we use batting averages for all players with at least 100 at bats.

5. Drawing one sample and calculating its mean

Before simulating drawing a large number of samples we first need to implement drawing one random sample from the population. Techniques for sampling with replacement and sampling without replacement in Excel will both be illustrated here. Although a theoretical sampling distribution is the result of sampling with replacement, we prefer to use sampling without replacement to generate the approximation of the sampling distribution. This is because we are only approximating the sampling distribution, and sampling without replacement provides a reasonable approximation for larger populations. Moreover, when students think of acquiring a random sample they picture a situation in which no item or person in the population is selected more than once. For this reason, using sampling without replacement makes the scenario more natural and concrete for students, and thus easier for them to understand.

We will also parameterize the sampling procedure so that the user of the spreadsheet can enter the sample size into an input cell. For this example we will allow the user to enter sample sizes from 1 to 50. Figure 5 shows the formulas that are used on the “Calculations” worksheet to draw one sample at random, without replacement, from the population data, of a size determined by the parameter input cell on the “Interface” worksheet. Note that, as with other figures in this paper, the figure only includes the top part of the worksheet. In figure 5 column B contains all 848 MLB player salaries. Column C contains 848 random numbers. Column E contains integers from 1 to 50, to number the values chosen randomly to be potentially included in the sample.

	A	B	C	D	E	F	G	H	I	J	K	L
1		Population Data			Drawing one Sample							
2												
3		MLB Salaries	Random Number		Sample mean:	\$2,712,496						
4		\$12,350,000	0.64886992									
5		\$10,000,000	0.21969832									
6		\$7,833,333	0.63940364		Sample:							
7		\$7,416,666	0.45390416		1	\$1,500,000						
8		\$6,150,000	0.85233540		2	\$1,200,000						
9		\$5,200,000	0.05957220		3	\$486,900						
10		\$4,750,000	0.91098182		4	\$490,000						

Figure 5: Formulas used to draw one random sample, without replacement

In order to draw the samples without replacement we first generate a list of random numbers using the =RAND() function. This list of random numbers should have the same number of values as the population. The formula in cell F6, which implements random sampling without replacement, uses the RANK function to rank the random value in cell C4 among all the random numbers in the range C4:C851. Because of the way the RANK function works, each of the random values can have only one rank in the list. Because there are 848 random values in the list, the possible values that the RANK function may return range from 1 to 848. The INDEX function in the formula takes the value returned by the RANK function and uses it to select the value in the population based on its position in the list of population values. This formula is copied down so that 50 random values are pulled from the population. For sampling with replacement, the RANK function can simply be replaced by RANDBETWEEN(1,848) to generate the index value.

The sample mean calculation formula in cell F3 in Figure 3 averages the first n values in the sample, where n is the sample size parameter entered by the user. In this example the sample size is input in cell F4 of the “Interface” worksheet. The formula in cell F3 constructs a range dynamically, and then that range serves as input to the AVERAGE function. The dynamic range begins at F6, the first value pulled from the population, and then extends downward so that the number of values included in the range matches the sample size parameter entered by the user. The INDIRECT function is a function that returns a range. The range is built dynamically by concatenating “F6:F” with 5 plus the number entered by the user for the sample size. If the user enters 1 for the sample size, for example, the range becomes F6:F6, and only the first randomly pulled number is included in the average. If the user enters 2 for the sample size, then the range evaluates to F6:F7 and the first two randomly pulled numbers are included in the sample, and so on.

6. Simulating multiple samples to approximate the sampling distribution

To approximate the sampling distribution we use Excel’s data table tool to draw multiple samples from the population, calculate the sample mean for each one, and record that sample mean. This is a non-standard use of the data table tool, so before describing it we should discuss the typical use of the data table tool – sensitivity analysis. The tool is used to automatically fill in a table that displays what the results of a calculation would be as one or two inputs to the calculation are varied. For example, if you have a spreadsheet that calculates a loan payment, you could create a data table to show what the payment would be for a range of different interest rates or principal amounts. Figure 6 illustrates a simple payment calculator that uses the PMT function to calculate a payment based on three inputs: loan principal, period of the loan, and interest rate. To the right of the calculation is the structure for a data table that will display how the payment changes as the interest rate varies.

	A	B	C	D	E	F	G
1							
2		Payment Calculator			Sensitivity Analysis		
3							
4					Interest Rate	Payment	
5		Loan Principal	\$15,000.00				
6		Loan Period (Yrs)	3		4.50%		
7		Interest Rate	5.00%		4.60%		
8					4.70%		
9					4.80%		
10		Monthly Payment	\$449.56		4.90%		
11					5.00%		
12					5.10%		
13					5.20%		
14					5.30%		
15					5.40%		
16					5.50%		
17							

Figure 6: Structure of a data table

To fill the table a formula that calculates the result or a reference to the formula that calculates the result should be placed in cell F5, the cell immediately above where the first result will be placed. In this example cell F5 contains the formula =C10. Next, the entire data table area is selected, from one row above the first input, down to the bottom of the table. In this example the range to be selected is E5:F16. Once the data table area is selected go to the data table tool in the ribbon under Data Tab → What-If Analysis → Data Table. After you click on “Data Table” you will be presented with a dialog in which to indicate where the inputs to be varied should be plugged into the spreadsheet. In the payment calculator example the inputs are in a column, and they should be plugged into cell C7, where the formula for the payment finds the interest rate, so we enter “C7” for “Column input cell.” “Row input cell” is left blank, since we don’t have any inputs in the top row of the table. Figure 7 shows the stage in the process of building the data table where the data table area is selected, the data table tool has been opened, and the user has entered the “Column input cell.”

	A	B	C	D	E	F	G	H	I	J
1										
2		Payment Calculator				Sensitivity Analysis				
3										
4					Interest Rate	Payment				
5		Loan Principal	\$15,000.00			\$449.56				
6		Loan Period (Yrs)	3		4.50%					
7		Interest Rate	5.00%		4.60%					
8					4.70%					
9					4.80%					
10		Monthly Payment	\$449.56		4.90%					
11					5.00%					
12					5.10%					
13					5.20%					
14					5.30%					
15					5.40%					
16					5.50%					
17										

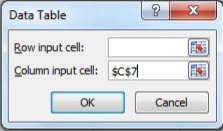


Figure 7: Process of filling in the data table

When we click “OK” the table is filled with the loan payments corresponding to each of the interest rate inputs. Figure 8 shows what the completed data table looks like. It now displays the monthly payments for the entire range of interest rate values. The formula in cell F5 is essential to making the data table work, so it may not be removed, but its result is not actually part of the data table. To make the data table more presentable we can format cell F5 to use white text, so that it doesn’t show up. Another option is to use a custom format to cause the cell to display a text label regardless of the underlying value in the cell.

	A	B	C	D	E	F	G	
1								
2		Payment Calculator				Sensitivity Analysis		
3								
4					Interest Rate	Payment		
5		Loan Principal	\$15,000.00			\$449.56		
6		Loan Period (Yrs)	3		4.50%	\$446.20		
7		Interest Rate	5.00%		4.60%	\$446.87		
8					4.70%	\$447.55		
9					4.80%	\$448.22		
10		Monthly Payment	\$449.56		4.90%	\$448.89		
11					5.00%	\$449.56		
12					5.10%	\$450.24		
13					5.20%	\$450.91		
14					5.30%	\$451.59		
15					5.40%	\$452.26		
16					5.50%	\$452.94		
17								

Figure 8: Completed data table

The way the data table works is that the first input value, 4.50%, is placed in cell C7 (the “Column input cell”), the formulas are recalculated, and then the result from cell C10 is placed to the right of the first input value, in cell F6. Then the second input value, 4.60%, is placed into cell C7, formulas are recalculated, and the result from cell C10 is placed to the right of the second input value, in cell F7, and so on.

If the formulas to be recalculated include random components the data table function may be used to implement a Monte Carlo simulation. This is the way we use it in building the sampling distribution model; we use it to draw many random samples from the population. When using the data table tool to perform Monte Carlo simulations in this way it is recommended to first go to Excel Options and in the formulas area set “Workbook calculation” to “Automatic except for data tables.”

When this is set you will need to recalculate formulas for the data table manually with the F9 key. This prevents the spreadsheet from responding slowly as the data table continually recalculates every time something is changed on the worksheet.

When building the data table for the sampling distribution simulation we put trial numbers into a column, instead of inputs, where each trial represents a separate random sample. The output we want to see is the sample mean, so we put a reference to the sample mean calculation one cell above where the first sample mean will be placed. When we use the data table tool the “Column input cell” should be set to an empty cell, since the trial numbers in the column do not actually play a role in the calculation of the sample mean. We are only using the data table to force formula recalculation, which forces a different sample to be drawn, and then to record the sample mean for each sample. Figure 9 shows the top of the portion of the “Calculations” tab where multiple samples are simulated.

E	F	G	H	I	J	K	L	M
Drawing one Sample			Simulation of 1,000 Samples					
Sample mean:	\$1,913,585.00		Trial #	Sample Mean	=F3			
			1	\$2,005,418				
Sample:			2	\$3,463,325				
1	\$521,000		3	\$2,508,755				
2	\$2,100,000		4	\$3,761,502				
3	\$487,500		5	\$4,413,041				
4	\$560,000		6	\$3,581,085				
5	\$5,500,000		7	\$3,164,566				
6	\$482,500		8	\$3,256,920				
7	\$700,000		9	\$3,791,623				
8	\$8,250,000		10	\$2,467,129				
9	\$525,000		11	\$4,239,080				
10	\$13,000,000		12	\$2,920,600				

Figure 9: Using a data table to store means for 1,000 random samples

Cell I3 contains the formula =F3, but it is formatted with a custom format to display the column label (“Sample Mean”) regardless of the value in the cell. For this example we chose to draw 1,000 random samples, so the range H4:H1003 has trial numbers, integers from 1 to 1,000 (Use the Fill → Series tool under the Home tab to easily create the integers from 1 to 1,000 in the column). To fill in the data table with 1,000 sample means select the range H3:I1003 and then go to the data table tool. Leave “Row input cell” blank, set the “Column input cell” to an empty cell, and then press “OK.” If you set the calculation options to “Automatic except for data tables” you will also need to press the F9 key after creating the data table to force all of its cells to recalculate.

The decision of how many samples to draw is influenced by a tradeoff between calculation time and accuracy of the results. A larger number of samples results in a longer time required for the calculations, but also results in a smoother percentage polygon for the sample means and less difference between the calculated mean and standard error of the sample means and the theoretical mean and standard error of the sampling distribution. On my 5-year-old Dell Latitude laptop there is a delay of around 1-2 seconds after F9 is pushed before the numbers and graphs on the “Interface” worksheet adjust. With 10,000 samples that delay sometimes exceeds 10 seconds. (Note that the calculations are much faster if the sampling is done with replacement.)

7. Examining characteristics of the approximated sampling distribution

Once the simulation has calculated and displayed the sample means, characteristics of the sampling means may be investigated in several ways. A primary approach is to create a distribution table of the population values and the sample means, and then use the percentages in the distribution table to create percentage polygons of the distributions of the population and the sample means. The class ranges for the distribution table may be set up in adjacent columns, with the ranges including values that are at least as large as the low point of the range, but less than the high point of the range. The formula to perform the counting of the number of values within the range can then be entered just once at the top of the table and then copied down to perform all the counts. In our example we use COUNTIF functions, with the criterion built dynamically from the values in the “At Least” and “Less Than” columns using the concatenation operator, &. Figure 10 shows the formula in context. The range \$B\$7:\$B\$854 refers to the range that contains the 848 population values, and the formula counts how many of the population values are greater than or equal to the low point of the class range, minus the number of population values that are greater than or equal to the high point of the class range. A similar formula is used for the count of sample means that fall into each range.

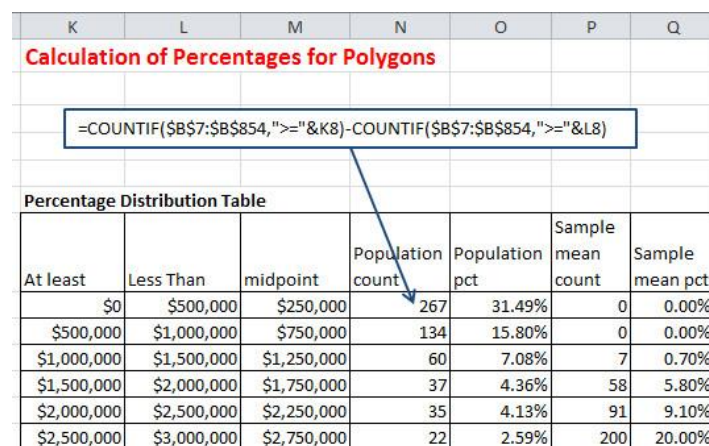


Figure 10: Formula for counting values in ranges

These counts may also be calculated with the COUNTIFS function or with the FREQUENCY function. The FREQUENCY function is entered as an array function and uses Excel’s concept of bins to delineate the upper and lower bounds of each range.

As a second tool to evaluate the distribution of the sample means we also create a normal probability plot (quantile-quantile plot) beside the percentage polygons graph. Normal probability plots are used to assess the normality of a dataset. The closer the plot is to being a straight line the closer the dataset is to being normally distributed. To build the plot, in the column next to the sample means calculate the quantile z-value for each sample mean. Then, construct a scatter plot with the sample means as the y-values and the quantile z-values as the x-values. The formula for calculating the quantile z-values is shown in equation 1.

$$=NORM.S.INV((RANK(I4,I4:I1003,1)/1001)) \tag{1}$$

This formula calculates the quantile for each of the sample means by dividing each mean's rank among the set of means (sorted in increasing order) by $n+1$, where n is the number of sample means. Then, the inverse standard normal function NORM.S.INV is used to calculate the z-value that corresponds to the quantile.

Finally, to help students understand the relationship between the mean sample mean and the population mean, and between the standard error (standard deviation of sample means) and population standard deviation, we calculate the expected and actual mean sample mean and standard error of the sample mean and display them above the normal probability plot. The expected mean sample mean is equal to the population mean, and the expected standard error of the sample mean when sampling without replacement from a finite population is given by equation 2, where σ is the population standard deviation, N is the population size, and n is the sample size. In this equation the common expression for standard error of the mean when sampling with replacement or sampling without replacement from an infinite population is multiplied by a finite population correction factor because we are sampling without replacement from a finite population [8].

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}} \quad (2)$$

8. Conclusion

The specific example introduced in this paper is a spreadsheet simulation model of a sampling distribution. It has been used in an introductory business statistics course to help students gain a solid understanding of the concepts of sampling distributions and the central limit theorem through active exploration of those concepts, while also improving their spreadsheet skills. The techniques used to build the model and the active discovery approach to learning which it enables may also be applied to a wide range of topics in an introductory statistics class. For example, the concept of sampling error can be illustrated by the natural variation in statistics calculated from different random samples of the same size drawn from the same population. The concept of confidence intervals may be illustrated by drawing thousands of random samples from a population, calculating a confidence interval from each sample, and then demonstrating that the percentage of confidence intervals that contain the population mean is approximately equal to the confidence level used to create the confidence intervals. The techniques may also be used to investigate concepts related to hypothesis testing, such as the probability of type I or type II error, the effect of sample size on power, or the meaning of a p-value.

References

- [1] American Statistical Association. (2005). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: Author.

- [2] Baker, John and Sugden, Stephen J. (2003). Spreadsheets in education –the first 25 years. *Spreadsheets in Education* (eJSiE): Vol. 1: Iss. 1, Article 2. Available at: <http://epublications.bond.edu.au/ejsie/vol1/iss1/2>
- [3] Barr, Graham D. and Scott, Leanne. (2011). Teaching statistics in a spreadsheet environment using simulation. *Spreadsheets in Education* (eJSiE): Vol. 4: Iss. 3, Article 2. Available at: <http://epublications.bond.edu.au/ejsie/vol4/iss3/2>
- [4] Black, Thomas R. (1999). Simulations on spreadsheets for complex concepts: teaching statistical power as an example. *International Journal of Mathematical Education in Science and Technology*: Vol. 30, Number 4, pp. 473-481.
- [5] Chandrakantha, Leslie. (2014). Visualizing and understanding confidence intervals and hypothesis testing using Excel simulation. *The Electronic Journal of Mathematics and Technology*: Vol. 8, Number 3, pp. 212-221.
- [6] Craft, Kim R. (2003). Using spreadsheets to conduct Monte Carlo experiments for teaching introductory econometrics. *Southern Economic Journal*: Vol. 69, Number 3, pp. 726-735.
- [7] Hesterberg, T. C. (1998). Simulation and bootstrapping for teaching statistics. In *American Statistical Association Proceedings of the Section on Statistical Education*. Alexandria, VA: American Statistical Association, pp. 44-52.
- [8] Jaggia, Sanjiv and Kelly, Alison (2013). *Business Statistics: Communicating with Numbers*. McGraw-Hill/Irwin.
- [9] Johnson, Arvid C. and Drougas, Anne M. (2004). Illustrating type I and type II errors via spreadsheet simulation in the business statistics course. *Decision Sciences Journal of Innovative Education*: Vol. 2, Number 1, pp. 89-95.
- [10] Mills, Jamie D. (2002). Using computer simulation methods to teach statistics: a review of the literature. *Journal of Statistics Education*: Vol. 10, Number 1.