COMPILING TOWARD AN UNDERSTANDING OF SAMPLING DISTRIBUTIONS

Ryan Phelps, Stephen F. Austin State University
Emiliano Giudici, Stephen F. Austin State University
William Long, Stephen F. Austin State University

ABSTRACT

This paper presents a novel independent learning experience and details evidence of its effectiveness. The learning experience presented here empowers students to better understand the sampling distribution identity and the Central Limit Theorem. This improved understanding is thought to be derived from the active role that students play in assembling sampling distributions. The valuable hands-on experience requires negligible effort on the part of the instructor. The assignment is ready to use and also completely customizable. Additionally, the student output generates a convenient plagiarism check resulting from time seeded random sampling. The assignment was designed for use within a business curriculum. Access to Excel and a basic understanding of Excel operations are sufficient for student success. Results suggest that this experience improved student understanding of the sampling distribution identity.

Keywords: Sampling distribution, Statistics education research, Central Limit Theorem, Active learning, Excel, Online learning, Discovery learning

INTRODUCTION

There are several excellent resources aimed at helping students to understand sampling distributions and their properties. Some of the more popular, freely-available resources are reviewed below. The learning experience detailed here provides a novel approach to the important and challenging topic. The guiding theory for our approach is that student effort applied to the abstract process of compiling sampling distributions will yield improved understanding. The assignment and all supporting materials are available online at: http://faculty.sfasu.edu/phelpsrt/bsd/. To utilize the assignment simply point students to the link after covering sampling distributions and the Central Limit Theorem.

The recent trend has been a move toward applets and the powerful demonstrations that they yield. McDaniel and Green (2012) present a walk through the discoveries made along this vein of the literature. In summary, applets can improve student outcomes in the area of sampling variation and these results are stronger when students are given guidance to focus on the more relevant results. Also, Larwin and Larwin (2011) demonstrate the benefit of utilizing this type of tool within a business statistics course. More recently, many of these applets have suffered from reduced browser compatibility.

While the reviewed applets generate sampling distributions from sampling parameters using software, our assignment utilizes guided student effort and a structured Excel work environment. When approaching the concept of sampling distributions, well prepared students have a grasp of probability distributions. Also, random samples are not particularly abstract or difficult to comprehend. Students seem to lose track of the process when progressing from the randomly generated means toward a sampling distribution. The tactile experience afforded by our assignment coupled with the visual presentation of the results makes the abstract construct more concrete.

Hakeem (2001) demonstrates that active-learning can increase both student engagement and understanding in business statistics. The learning experience detailed here provides a blending of the applied nature of small population exercises with the rewarding visual presentations afforded by more sophisticated applets. Also, online delivery provides guided active-learning and student interaction with technology outside of the classroom environment. While technology in the classroom may be desirable, it is not always feasible and can be a distraction as shown in Martin (2011).

LITERATURE REVIEW

There is a plethora of literature decrying the difficulty and importance of understanding the sampling distribution identity (Becker and Greene 2001; Chance, del Mas, and Garfield 2004; Lane and Tang 2000; McDaniel and Green 2012). An understanding of the underpinnings of statistical inference becomes even more important in advanced applications of statistics. According to Kennedy (2001) “[S]tudents must stop viewing econometrics through a mathematical lens and start viewing it through the sampling distribution lens” (p. 113). The sentiment of this early piece is certainly in line with that of more recent developments as he goes on to add: “Although lecturers and textbook authors like to think otherwise, brilliant expositions seldom cause students fully to understand - such understanding comes through working out problems based on the concept to be learned” (p. 114).

Exponential improvements in computers and their use in statistical applications and pedagogy have been redefining our craft (Chance, del Mas, and Garfield 2004). These changes have made a host of new tools available including computer simulation methods (CSMs). These tools seem to have led to improvements in statistics instruction (Lane and Tang 2000; Mills 2002; Mills 2004). Many of these resources are freely available and have components aimed at improving understanding of the sampling distribution identity. For convenience, Table 1 details differences between some of the more popular options and the approach in this paper (Ours).

The Rossman/Chance applet collection (RCAC) provides a warehouse of statistical demonstrations (Rossman and Chance, n.d.). While this valuable resource does not have built in assignments, it serves as a powerful engine for statistical discovery. Also, publically available student exercises are in the development process (McDaniel and Green, 2012). Some nice features of these applets include the use of finite populations and the reporting of the histogram of the last random sample drawn. Also, their sampling distribution simulations include applications for proportions, confidence intervals, and regression predictive equations. One would be remiss not to review this resource due to its value and accessibility.

The remaining reviewed CSMs are: Web Interface for Statistics Education (WISE) from Claremont Graduate University (n.d.), the sampling distributions section of Rice Virtual Lab in Statistics (RVLS) (Lane, 1997) and Tools for Teaching and Assessing Statistical Inference (TTASI) (Garfield, delMas and Chance, 2000). The applets that drive these learning experiences are similar in nature. Also, the learning experiences themselves are more limited in scope when compared to (RCAC). The WISE learning environment provides the user with an integrated learning experience. The applet is embedded in the online question set giving it the most simple startup process of the group. It also displays the answers which can be an attractive nuisance for non-motivated students. The RVLS is more focused on sample statistics. It is unique in that it generates sampling distributions for several estimators. It has a short exercise that requires a little navigation to find. While unavailable as of 9/30/19, the TTASI experience is different in that it provides both a pre- and post-test and a more involved activity powered by the applet. It is also notable that the activity asks students to form expectations of outcomes prior to finding the solutions, as does the Wise applet. The use of business application problems is also a nice touch. The TTASI learning environment requires the acquisition and use of additional software; as a result, startup is a little cumbersome. Also, the student activity requires that labels be printed for each student or student group.

Table 1: Sampling Distribution Educational Resource Feature Chart

	Sampling Distribution Educational Resource				
	Ours	RCAC	RVLS	WISE	TTASI
Students Compile Sampling Distribution	Yes				
Builds Experience with Excel	Yes				
Provides Physical List of Sample Means	Yes				
Allows for the Use of Finite Population	Yes	Yes			
Allows for Custom Population	Yes	Yes	Yes		
Simple Startup	Yes	Yes	Yes	Yes	
Provides Paired Exercises	Yes		Yes	Yes	Yes

The learning experience detailed here is not an applet and does not cover a broad set of statistical knowledge. The assignment is focused on improving students' understanding of the sampling distribution of the sample mean and its properties. This topic was chosen because of its high level of importance and difficulty. Additionally, there is a lack of good solutions in standard texts. The learning experience detailed here is so simple to utilize that it may be adopted by faculty who have not been bold enough to utilize more comprehensive resources. Also, positive results may encourage those who are unfamiliar with the above resources to invest the time needed to become familiar with them. This experience fills gaps in the current environment by applying active learning to an abstract process. It is also novel in that it requires the student to undergo involved processes in Excel. Exposure to some of the functionality of VBA within Excel is an additional benefit. Student feedback suggests that the assignment is both challenging and enjoyable. The remainder of the paper describes the assignment and provides evidence of its effectiveness including student responses.

THE ASSIGNMENT

The assignment materials include an instruction document, an Excel work environment and an Excel-tutorial video. They are all are available at: http://faculty.sfasu.edu/phelpsrt/bsd/. The instructions, provided in Appendix A, guide students through the process of compiling three approximations of sampling distributions of sample means. The process is composed of three phases. The students generate lists of random means, use the results to compile relative frequency distributions and graph the results.

The lists are generated by making simple alterations to the existing VBA code in the Excel work environment. Once students have entered the sample size, 40, 60, or 100, they click the “Start” button. The code then randomly generates 10,000 means based on the criteria. Students then move the list of means to another worksheet and compile the results into relative frequency tables. The bins for all of the tables are automatically generated by the work environment as a function of the population. Finally, students build a scatter plot of the relative frequency distributions of the sampling distributions and the population. Students are then prompted to provide a description of the sampling distribution of the sample mean, describe how its shape changes with alterations to the sample size, contrast the shape of the sampling distributions to that of the population, and relate how their work is relevant to the course content.

The Excel Work Environment

The Excel work environment is preloaded with a dataset. It contains an “About” tab which describes and cites the data, a “Population” tab which contains the code that generates the random samples and reports their means, three staging tabs for the construction of the sampling distributions, a staging area for the graph with a preloaded relative frequency distribution of the population, and a tab that calculates results and compares them to aspects of the population. For an example outcome see Appendix B.

While the file is ready to use with preloaded data, it is also completely customizable. One may wish to substitute a population that is more interesting to the class or an estimator other than the sample mean. Data with extreme outliers should be avoided as they may diminish the usefulness of the graph output. The only other limitation on the population is that it must be small enough to fit in one column in Excel 2003 (N < 65,536). The code produces randomly generated sample means for samples of any size less than N. For insight into the evolution of the VBA code used here see Giudici, Phelps, and Calafiore (2012). The number of repetitions is also limited by the number of rows in an Excel column; however this number is only limited by the number of rows in the version of Excel being used.

The size of the population and the number of samples generated will alter the time needed to produce the results. For example, using a population of size 3,220 and a sample size of 40 the code took twenty-four seconds to generate 10,000 random sample means. Cutting the size of the population in half reduced the runtime to sixteen seconds. Increasing the sample size from 40 to 100 did not add to the time required. Available software packages can achieve these results in less time. However, the benefits of this solution include nearly universal student access to and familiarity with Excel. Additionally, no foreign Excel add-ons are required which prevents the inevitable complications that they entail.

RESULTS

Evidence of Effectiveness

Students at a regional public university were assigned the learning experience as part of an applied statistics course. Introduction to probability and statistics is a prerequisite for the course. The assignment was given after reviewing probability distributions and after all other coverage of the sampling distribution of the sample mean and the application of its properties. To measure the effectiveness of the assignment, identical pre- and post-tests were given. For the text of the pre- and post-tests see Appendix C. The pre-test was given in class after all of the relevant material had been covered. Only after the pre-tests had been collected, were students informed that their accuracy would not affect their grade. The assignment was handed out immediately after the pre-test. The post-test was completed in class, four days after the assignment had been collected. Table 2 details pre- and post-tests results.

The results demonstrate that there was a statistically significant improvement between the pre- and post-tests. This, however, is not very strong evidence that the assignment *caused* the improvement. Students may have been motivated by their uncertainty on the pre-test to study for the post-test. Both the short period of time between the tests and the fact that their pre-tests were not returned until after the post-tests were graded, support *causality*. Also, there is a statistically significant correlation between student scores on the assignment and the results on the post-test. The correlation between student results on the pre-test and the assignment is much smaller and insignificant.

Table 2 Within-Cohort Comparison

	Assign.	Pre-test	Post-test
Sample Size	41	41	41
Average	73.63%	47.90%	68.87%
Correlation with Assignment		0.1019	0.4003
T-test of Difference		4.6408	

Table 3 details a comparison of the results of the first exam (with) versus the same exam in the prior term, when the assignment was not given, (without). The results add support to the before/after results in Table 2. The with/without results provide evidence that student understanding was increased by either the teaching tool or the pre- and post-tests. The results from the first item: “Explain, in your own words, what the sampling distribution of the sample mean is.” are reported under Item-1. There is a clear improvement in student performance between the two terms. The improvement resulted in a t-test that is significant at the 5% level. The result is negative in sign for the second item: “Use our summary approximations and the above information to approximate the standard deviation of the sampling distribution of the sample mean.” The (with) results were not significantly different. The results for the exam as a whole suggest a statistically significant improvement. The results do not control for the overall quality of students, but are at least suggestive that the teaching tool resulted in an improved understating of the sampling distribution identity. Also, the results of the (with) exam are significantly correlated to the assignment results.

Table 3: Between Cohort Comparison

	Item 1	Item 2	Exam
Without (n = 52)	59.90%	38.56%	62.06%
With (n = 59)	72.80%	37.29%	69.15%
Correlation with Assignment	0.1957	0.0424	0.3626

The results of the first quiz from each term provide further evidence that the improvement may have been causal. Table 4 details quiz results over the relevant material. This quiz was given in both terms after the material had been covered and before any activity related to the teaching tool. The results suggest that students were not different in their capacity to understand the sampling distribution of the sample mean or its properties and application.

Table 4: Cohort Content Knowledge Comparison

Quiz 1 Without (n = 49)	72.17%
Quiz 1 With (n = 57)	69.88%
T-test of Difference	-0.6023

Student Reaction

An early version of the VBA code was revised as a result of student input. A Computer Science major took an interest in the code and revised it to make it more efficient. His efforts reduced the time required by the program by more than 50%. His contribution was not motivated by course credit. Also, anecdotal evidence and student comments suggested that the assignment was generally appreciated.

Student Comments

- "I actually appreciated the assignment because it cleared some things up for me that I didn't understand prior to the assignment. And I am a visual person so it helped me understand in that way, too."
- "I got a chance to play with Excel when I got stuck at certain points. It helped me with using and seeing what Excel can do."
- "I think the homework really helped me learn more about the class and how sample size affects outcomes. I really enjoyed this homework. It is a great hands-on approach to learning!..."
- "… Running the program was interesting and fun…"
- "The assignment was beneficial in that it helped me learn more of how to work in Excel and it helped to see what actually happens to the sampling distribution as the sample size grows."

CONCLUSION

This assignment is intended to augment available resources aimed at guiding students to an understanding of the sampling distribution identity and its properties. Understanding the nature of the sampling distribution of the sample mean and its relationship to the Central Limit Theorem is as vital for statistical literacy as it is challenging. This understanding will give students a better foundation for developing statistical literacy. Success in this area will reward students throughout their efforts in the area of statistics and empower them to conduct statistical inference more appropriately and with increased confidence.

Our results suggest that giving students an active role in compiling large-population sampling distributions can improve their understanding of the sampling distribution identity. Our teaching tool is unique in that it effectively provides students with the structure needed to empower active participation in this involved process. Going through this independent learning experience is both straight-forward and meaningful. Feedback and personal experience suggest that the assignment is engaging and fun. Finally, utilizing this teaching tool requires little added instructor effort.

Exposing students to the existence of the Data Analysis ToolPak and VBA within Excel are added benefits of the assignment. Absent this experience, many students may have an atrophied view of the capabilities of Excel. This exposure may prompt some students to take courses that allow them to get more out of Excel, or even prompt a program tract that will enhance their quantitative skills and computer literacy.

This teaching tool has benefited from several years of active service and constructive comments from faculty and students. It is our hope that others will implement this learning experience and enjoy similar results. We look forward to receiving constructive comments and perhaps independent attempts to measure its effectiveness.

ACKNOWLEDGEMENTS

We would like to thank all of those students and faculty who have provided constructive feedback. Specifically, we would like to thank James Sanders for his contribution to the development of the code as an undergraduate student. Finally, we would like to thank Michelle Everson for her kind direction and helpful comments.

REFERENCES

Becker, W. E., and Greene, W. H. (2001). “Teaching Statistics and Econometrics to Undergraduates.” Journal of Economic Perspectives 15 (4): 169–182.

Chance, B., del Mas R., and Garfield, J. (2004). “Reasoning about Sampling Distributions.” In The Challenge of Developing Statistical Literacy, Reasoning and Thinking, edited by Dani Ben-Zvi and Joan Garfield, 295–323.

Claremont Graduate University. (n.d.) Sampling Distribution of the Mean Tutorial. Web Interface for Statistics Education (WISE). Accessed June 1, 2020. http://wise.cgu.edu/portfolio/samplingdistribution/.

Garfield J., delMas R., and Chance, B. (2000). Tools for Teaching and Assessing Statistical Inference (TTASI). Unavailable as of June 1, 2020. http://www.tc.umn.edu/~delma001/stat_tools/.

Giudici, E., Phelps, R., and Calafiore, P. (2012). “An Excel Tool for Teaching the Central Limit Theorem to Undergraduate Business Students,” Journal of Information Systems Technology and Planning, 5(13).

Hakeem, S. A. (2001). “Effect of Experiential Learning in Business Statistics.” Journal of Education for Business 77 (2): 95–98.

Kennedy, P. E. (2001). “Bootstrapping Student Understanding of What Is Going on in Econometrics.” The Journal of Economic Education 32 (2): 110–123.

Lane, D. M. (1997). Sampling Distribution Simulation. Accessed June 1, 2020. http://onlinestatbook.com/stat_sim/sampling_dist/index.html in Lane, D. M. (Ed.) Online Statistics Education: A Multimedia Course of Study, http://onlinestatbook.com/.

Lane, D. M., and Tang, Z. (2000). “Effectiveness of Simulation Training on Transfer of Statistical Concepts.” Journal of Educational Computing Research 22 (4): 383–396.

Larwin, K. H., and Larwin, D. A. (2011). “Evaluating the Use of Random Distribution Theory to Introduce Statistical Inference Concepts to Business Students.” Journal of Education for Business 86 (1): 1–9.

Martin, L. R. (2011). “Teaching Business Statistics in a Computer Lab: Benefit or Distraction?” Journal of Education for Business 86 (6): 326–331.

McDaniel, S. N., and Green, L. (2012). “Independent Interactive Inquiry-Based Learning Modules Using Audio-Visual Instruction in Statistics.” Technology Innovations in Statistics Education 6 (1).

Mills, J. D. (2002). “Using Computer Simulation Methods to Teach Statistics: A Review of the Literature.” Journal of Statistics Education 10 (1): 1–20.

Mills, J. D. (2004). “Learning Abstract Statistics Concepts Using Simulation.” Educational Research Quarterly 28 (4): 18–33.

Rossman, A., and Chance, B. (n.d.) Rossman/Chance Applet Collection (RCAC). Accessed June 1, 2020. http://www.rossmanchance.com/applets/index.html.

APPENDIX A: ASSIGNMENT INSTRUCTIONS

Building Sampling Distributions of the Sample Mean

Please read completely prior to beginning. Visit the Webpage: http://faculty.sfasu.edu/phelpsrt/bsd/ **Download the Excel Work Environment**. Preview the "About" tab in order to get an understanding of the nature of the population that you will be working with and the appropriate scale of the sample means to be generated. It is assumed that you are familiar with the concept of a relative frequency distribution. The **Building and Graphing Probability Distributions in Excel Video** demonstrates how to compile and plot three relative frequency distributions for comparison.

Overview:

You will be using the Visual Basic tool within Excel to generate and graph approximations of sampling distributions of the sample mean. These approximations will be composed of 10,000 sample means obtained from 10,000 samples drawn randomly, with replacement, from the population. The **Excel Work Environment** together with the instructions below will make this process simple.

When finished you will print and turn in one graph including the relative frequency distribution of the population and three probability distributions of sample means, each labeled with its corresponding sample size. Your graph should include four probability distributions (Each should look distinctly different)

1) "Population"
2) "Sampling Distribution of X-bar (n = 40)"
3) "Sampling Distribution of X-bar (n = 60)"
4) "Sampling Distribution of X-bar (n = 100)"

Instructions:

After **SAVING** the **Excel Work Environment** make sure to "enable editing" and to "enable content". Once the file is opened, you will need to access the code. To do this, simply right-click on the red "Population" tab label and choose "View Code". This will reveal the VBA code that will produce the lists of sample means. The default sample size is n = 40. Also as a default, the code will produce the sample averages of 10,000 samples.

Simply return to the population tab of the Excel file and click the "Start" button to watch Excel work in conjunction with the Visual Basic code.

The code carries out the following actions:

- Draws a random sample (sample size is specified in the code) from the population in **Column A**
- Adds one to "Use Count" in **Column B** for each member of the population included in the sample
- Calculates the average of the randomly generated sample
- Writes the average in the first available cell in "X-bar" in **Column E**
- Repeats the above 9,999 times

Do not attempt any work in Excel while the program is running. Let the program run in the background. Your computer may appear to be frozen while the code runs. Simply wait until the program signals that it is done. Once the program has stopped, you will have your first list of sample means generated from the population using a sample size of 40.

Copy and paste the list of 10,000 sample averages under the heading "X-bar" in **Column A** of the tab titled "n= 40". To copy click on the top value and highlight the data using [Ctrl] [Shift] and [down-arrow] keys. Once highlighted [Ctrl] with [C] will copy and [Ctrl] with [V] will paste.

Next, you must use the histogram tool, in the Data Analysis ToolPak to build a relative frequency distribution of the sample means (see the frequency distribution video if you need a refresher). This is an approximation of the sampling distribution of the sample mean. Paste the results of the histogram (both the bin and frequency results) under the appropriate sample size in the "Graph" tab and calculate the relative frequencies for the graph.

Next, return to the population tab in Excel and, if necessary, right click on the tab label "Population" to access the code. Once there, you will need to change the sample size in the Visual Basic code to n = 60 and redo all of the relevant steps above. To change the sample size in the Visual Basic window simply scroll to the top of the code and change "glngSampleSize = 40" to "glngSampleSize = 60".

Once you have done this you can again return to the "Population" tab of the Excel file and hit the "Start" button. The program will take the same amount of time to generate a second list of random sample means.

After you have

- Pasted the new results under the heading "X-bar" in the "n= 60" tab of the Excel file
- Generated a new histogram using these values
- Used the results to update the "Graph" tab

You need to repeat the above one more time using n = 100.

Next, generate a graph **(scatter type)** with four series. Three of the four series will contain the randomly generated means and their relative frequencies the fourth will contain the population.

!!! - NOTE: DO NOT INCLUDE THE WORD "MORE" OR ANY LABELS IN YOUR INPUT FOR THE GRAPH - !!!

In the end you should have one graph including a relative frequency distribution of the population and three (different looking) sampling distributions of sample means labeled with their corresponding sample sizes.

Trouble Shooting Graph: Check the boxes below.

□ The range of the x-axis variable is from a little above zero to around twenty for the population.

If the x-axis is not indicating the correct range of values or if the plots in your graph are very similar in appearance, then check the input used for the graph's x- and y-axes to be sure that no text input was included.

□ The height of the tallest plot is around .35.

If the height is larger than one, then go back and be sure to select the relative frequency.

Turn in the Following

Submit Excel Output:

Insert the following items into a Word document.

To do this, copy and paste special, "Picture (Enhanced Metafile)".

1) Customize and **insert** the graph containing four probability distributions
   - o **Label your axes!**
   - o Your graph should include four distributions (Each should look distinctly different)
     - □ "Population"
     - □ "Sampling Distribution of X-bar (n = 40)"
     - □ "Sampling Distribution of X-bar (n = 60)"
     - □ "Sampling Distribution of X-bar (n = 100)"

2) **Insert** the sheet labeled "Print Out 2"

Submit Analysis of Results:

In this section, you will look over and interpret the results that you have generated. To complete this section type up a paragraph that addresses each point below.

3) Detail your process and the product of that process
   - o Summarize what you did. In a few sentences, describe the entire process and outcome
   - o Carefully detail what the sampling distribution of the sample mean is
4) Describe what happens to the shape of the sampling distribution of the sample mean as the sample size increases
   - o This should be apparent from your graph
   - o Include the position of the population average in your discussion

   - Your response must line up with what you know to be true about the sampling distribution of the sample mean

5) Discuss where this change in shape can be seen in the results in “Print Out 2”
   - Your response must include numbers from “Print out 2”
   - Your response must line up with what you know to be true about the sampling distribution of the sample mean
6) Compare and contrast the shape of sampling distributions of the sample mean to the shape of the population
   - This should be apparent from your graph
   - Include the position of the population average in your discussion
   - Your response must line up with what you know to be true about the sampling distribution of the sample mean
7) Discuss what you learned through this exercise
   - Your response must line up with what you know to be true about the sampling distribution of the sample mean
8) Discuss why what you have learned is relevant to class

Note: If your graph does not line up with what you know to be true about the sampling distribution of the sample mean it may be that your graph contains errors (see the linked video for the graphing process). Also, your results may differ slightly from theory because we did not take every possible sample, but rather built approximations using only 10,000 samples. As a result of the approximation, the average sample mean will not exactly equal the population mean.

APPENDIX B: EXAMPLE OUTCOME

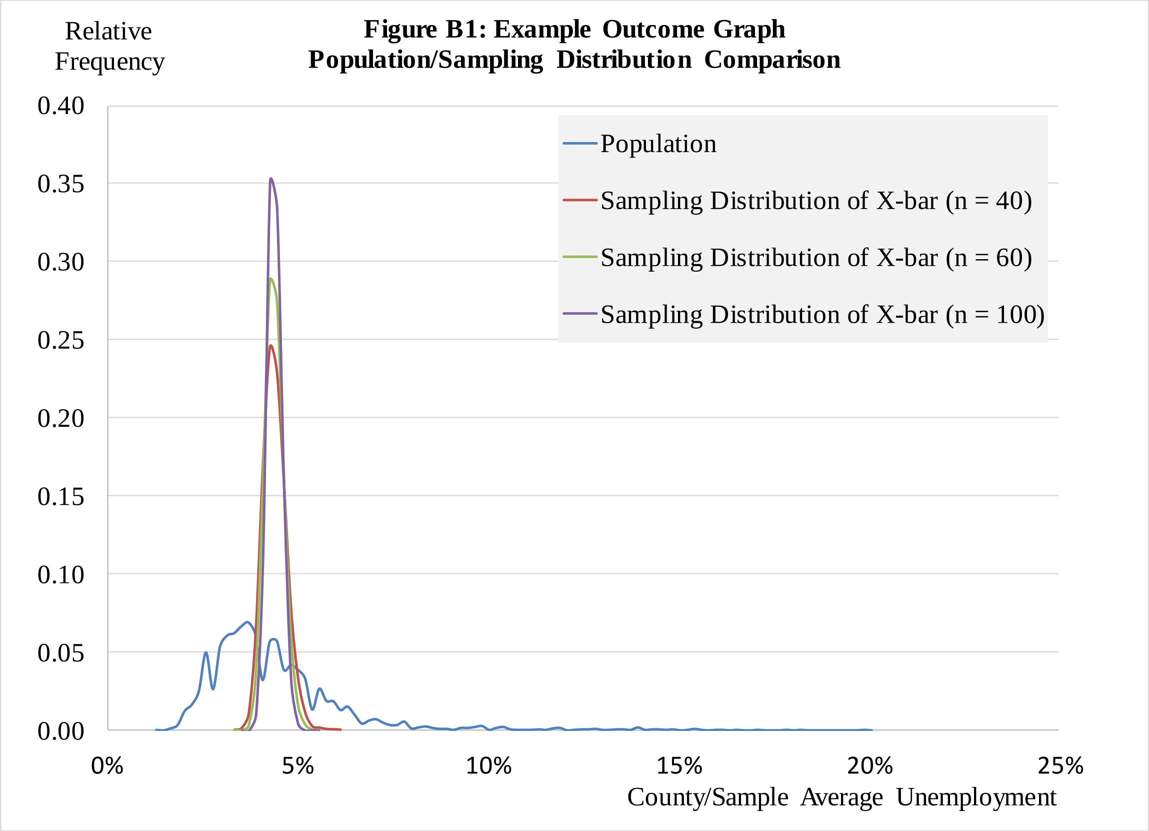
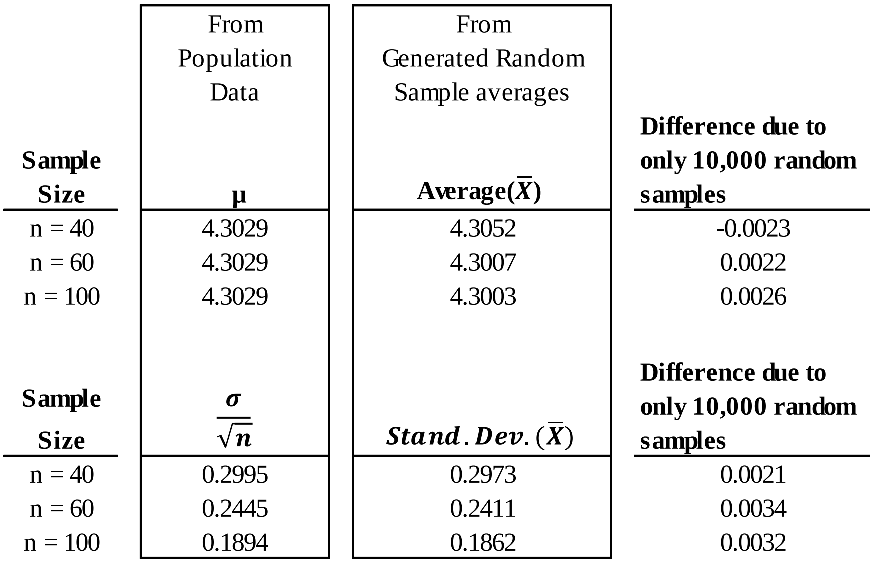


Figure B1: Example Outcome Graph
Population/Sampling Distribution Comparison

Figure B2: Example Outcome "Printout 2"



Sample Size	From Population Data μ	From Generated Random Sample averages Average($\bar{X}$)	Difference due to only 10,000 random samples
n = 40	4.3029	4.3052	-0.0023
n = 60	4.3029	4.3007	0.0022
n = 100	4.3029	4.3003	0.0026

Sample Size	$\frac{\sigma}{\sqrt{n}}$	$Stand.Dev.(\bar{X})$	Difference due to only 10,000 random samples
n = 40	0.2995	0.2973	0.0021
n = 60	0.2445	0.2411	0.0034
n = 100	0.1894	0.1862	0.0032

Note: This table works well as a fast plagiarism check. Since the random sample generator is seeded to the time, the only way to get two identical results here is to print the results twice.

Example Student Analysis

In this analysis, a population of 3,220 annual average U.S. county unemployment rates was used to generate three approximate sampling distributions and a frequency distribution of the population. Each of the sampling distributions used a different sample size (n=40, 60, and 100). The three sampling distributions were created by using an Excel macro to draw three sets of 10,000 random samples (each of the appropriate size) from the population and find the mean ($\bar{x}$) of each sample. After obtaining the results for each sample size, a graph was created in order to compare each sampling distribution to the population frequency distribution. The graph (shown above) displays an approximate sampling distribution of $\bar{x}$ for samples of size 40, size 60, and size 100, as well as the population distribution of x. The mean and standard deviation for each set of 10,000 samples can be seen in "Print out 2."

(These distributions are approximate because in order to generate a true sampling distribution, every possible sample of size "n" would have to be included. However, for a sample of size 40, there are 1.1335E+112 possible samples. Examining all of these possible samples would require far more time and computing effort than is reasonable for this analysis. For samples of size 60 and 100, the number of possible samples is much larger.)

The distribution of $\bar{x}$ can be considered normal under Central Limit Theorem 2, since all of the sample sizes described above exceed 30. This distribution has a theoretical mean that is equal to the population mean of 7.9276, though each of the above distributions varies slightly due to the limited number of samples examined. Likewise, the standard deviation of each distribution should be equal to the population standard deviation divided by the square root of the sample size, but minor variations can be seen here.

As the sample size increases, the standard deviation of the sampling distribution shrinks. This reduces the variability of the distribution, causing the results to gather more closely around the population mean of 7.9276. This can be seen on the above graph as an increase in the distribution's height and a decrease in width as n increases. In addition, larger sample sizes lead to slightly better estimates of the population mean. The results from "Print out 2" confirm this. As n increases from 40 to 100, the standard deviation of the sampling distributions shrinks from 0.4905 to 0.3090…

While each of the sampling distributions are normal and clustered tightly around the population mean, the population distribution is more irregular and has a greater spread. However, all four distributions have approximately the same mean of 7.9276. The sampling distributions are taller and narrower because taking the average of a sample reduces its variability. The population distribution must account for extreme results, while such results tend to be balanced out by less extreme or opposite results in a sample average.

This exercise demonstrates the tendency of $\bar{x}$ to gather around the population mean. It also shows the distribution shape of $\bar{x}$, and the effect that different sample sizes will have on that distribution. From the graph above, we can see that $\bar{x}$ is an unbiased estimator of Mu, and that its efficiency as an estimator increases as n increases. This is relevant to the class material because it provides a visual representation of many of the concepts covered in chapter 7. It also displays the effect of the standard deviation on the shape of a distribution and demonstrates the process of collecting samples, as covered in previous chapters.

APPENDIX C: PRE- AND POST-TEST QUESTIONS

1) What is the sampling distribution of the sample mean?
2) List some of the important properties of the sampling distribution of the sample mean.
3) What happens to the sampling distribution of the sample mean when the sample size is increased?
4) What happens to the variance of the sample mean when the sample size increases?