proceedings

MDPI

Proceedings

Understanding the Central Limit Theorem the Easy Way: A Simulation Experiment †

Monica E. Brussolo

Operations Management at the Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080, USA; monica.brussolo@utdallas.edu; Tel.: +1-972-883-4411

† Presented at the 2nd Innovative and Creative Education and Teaching International Conference (ICETIC2018), Badajoz, Spain, 20–22 June 2018.

Published: 5 November 2018

Abstract: Using a simulation approach, and with collaboration among peers, this paper is intended to improve the understanding of sampling distributions (SD) and the Central Limit Theorem (CLT) as the main concepts behind inferential statistics. By demonstrating with a hands-on approach how a simulated sampling distribution performs when the data used has different probability distributions, we expect to clarify the notion of the Central Limit Theorem, and the use of samples in the hypothesis testing process for populations. This paper will discuss an initial stage to create random samples from a given population (using Excel) with collaboration of the students, which has been tested in the classroom. Then, based on that experience, a second stage in which we created an online simulation, controlled by the professor, and in which the students will participate during class time using an electronic device connected to internet. Students will create simple random samples from a variety of probability distributions simulated online in a collaborative way. Once the samples are generated, the instructor will combine and summarize the resulting sample statistics using histograms and the results will be discussed with the students. The objective is to teach some of the central topics of introductory statistics, the Central Limit Theorem and sampling distributions with an interactive and engaging approach.

Keywords: Central Limit Theorem; simulation; sampling distribution; random samples; approximation to normality

1. Introduction

At the core of the statistics field, there is a fundamental concept which students find difficult to understand. It is the Central Limit Theorem and how it provides the foundation for using a random sample to make inferences about any population. This idea is vital for inferential statistics, one of the main objectives of any introductory statistics course, and critical knowledge for any statistics practitioner. As Cobb and McClain mentioned, statistics education first should "focus on developing central statistics ideas rather than on presenting set of tools and procedures" [1]. One of these central ideas is the Central Limit Theorem. As this is not a trivial concept, students are generally puzzled in how to interpret this theorem and they learn its theoretical framework without a clear understanding of its implications.

The aim of this paper is to create an engaging way to teach the Central Limit Theorem with an experiment based on simulations of probability distributions and their corresponding sampling distributions. The students will collaborate by creating their samples as part of a single sampling distribution, making them an active part of the simulation and creating a more engaging environment during class. The main difference of this approach and other existing simulations is that the group

will work together to create a sampling distribution, for which the statistics will be analysed as a group effort.

This paper includes the following sections: the first section discusses how other sampling distribution simulators have approached the topic and which areas of opportunity were identified. The second section presents a different and innovative approach by emphasizing collaboration among peers while performing simulations. The paper then explains how this simulation works and what it has been our experience with collaborative effort in simulations using Excel in the classroom. Then, we discussed future paths for research describing an experimental design, which is in planning stage. The paper concludes referring to the results obtained by the simulation and its application to the Central Limit Theorem and sampling distributions.

2. Related Work and Areas of Opportunity

To explain the Central Limit Theorem and sampling distribution in introductory statistics courses, instructors have resourced to the use of simulations allowing them to replicate realistic scenarios for the students to understand these topics in an intuitive way. However, the use of simulations is not exempt of issues. As it is explained by Hodgson and Burke, some students have difficulty understanding the logic of the simulation and assume "one must draw multiple samples in order to make valid statistical inferences" [2].

In order to use sampling distributions for statistical inference, students must learn three fundamental properties of the sampling distribution of the mean (SDM) with the assumption of selecting random samples of size n. These properties are:

1. "mean $\mu_{\bar{x}_n}$, equal to the mean of the population: $\mu_{\bar{x}_n} = \mu$
2. standard deviation, $\sigma_{\bar{x}_n}$, equal to the standard deviation of the population divided by the square root of the sample size: $\sigma_{\bar{x}_n} = {}^{\sigma}/_{\sqrt{n}}$
3. (Central Limit Theorem) a shape that is normal if the population is normal; for other populations with finite mean and variance, the shape becomes more normal as n increases" [3].

The current approach to simulations of the SD has some challenges. We identified three areas of improvement while comparing this project with the available resources used to run simulations.

The first problem arises when simulations are performed by demonstration only. In this case, students play a role of spectators while the instructor demonstrates the simulation. Even though the simulation is explained in detail, students do not develop a clear understanding of the SDM and its properties.

The second problem occurs when each student creates his own simulated sampling distribution. Therefore, multiple simulated sampling distributions are created, making it difficult for the students to understand the reason behind that, and it enforces the incorrect idea of having to create multiple samples in order to make inferences.

The third problem results from not having an accurate measure of the effect, if any, of using simulations to teach the CLT and SDM, so there is no evidence of the benefits that students are obtaining through this approach. A deficient job has been done to document the effect of using simulations and how different ways to do it, can report different results. We were unable to identify a robust experimental design focused on the effect of simulations to explain the CLT. Thus, as the following steps of this project, we will address that gap by using an experimental design to measure the effect of simulation of sampling distributions in the learning process.

3. Methods

By creating a simulation of different probability distributions (Normal, Uniform, Exponential) and with students' collaboration who will generate n random samples from each of these distributions, we can develop simulated sampling distributions of the mean from each of these probability distributions. Then, by using data visualization, we will provide evidence to the students

about the original probability distribution, and how the simulated sampling distribution is shaped once we have grouped all the sample statistics created by the students.

This process has been tested in the classroom in its first stage. Initially, the simulation was conducted using Excel, by asking students to create random samples from a uniformly distributed population using the function RAND in Excel. Then, the students calculated descriptive statistics of their samples and they reported them back to the professor. Once all the samples were analysed, the professor created a simulated sampling distribution with the sample statistics from the students.

Through several iterations and by changing the sample size, it was evident that the collaboration of the students was a positive addition to this process. The students were able to identify their own samples among all the samples, giving them a better understanding of this process. After performing this exercise, it was also evident that this procedure required improvements. Thus, a simulation was created with an online application that allows the instructor to control the simulation of population data, and it allows the collaboration of the students simultaneously while they create their samples. The benefit of this approach is the collaboration of the students, the expedite feedback they receive, and how easy is to use for the professor. This innovation will provide a teaching tool for statistics professors to improve the learning process of one of the most significant areas of knowledge in the statistics field.

The procedure for the online simulation is as follow. First, we start this process with a pre-test about sampling distributions and the CLT, then we select the type of probability distribution that we want to simulate. After that, we select the sample statistics that we want to evaluate (mean, median, proportion, standard deviation), and the sample size. It is recommended to start with a small sample size such as $n = 5$ and in subsequent iterations, move to sample sizes of 20, 30, 50 and 100. Based on the class size, we will determine how many samples are needed from each student. For groups of 50 or more participants, the recommendation is to create up to three samples per student.

At this point, once the instructor provides students with the link and the code to access the simulation, we need to verify that they are connected online. Then, we instruct them to create their samples and their sample statistics, so we can group all together and prepare a histogram with the simulated sampling distribution. After that process is completed, we will show these results to the students and discuss the shape of the histogram. We will repeat the process by creating different combinations between probability distributions and sample size. This process is concluded by summarizing the different iterations of the simulation. We will then discuss how collaboration among peers makes this process more effective and how our findings relate to the theory discussed in class. At the end of this process, a post-test for knowledge check will be given to the students.

4. Results and Discussion

Once we test the online simulation, and based on our experience using Excel, we anticipate that this approach to demonstrate the theoretical principles of sampling distribution and the CLT is effective for any probability distribution as long as the samples are randomly selected and their size is large enough. The general recommendation is to select samples of size $n \geq 30$. This will provide a practical example of the Central Limit Theorem, how to understand it and apply it.

This project's contribution to the topic is twofold. First, students will participate in the simulation, as they collaborate with each other, so the entire group will create an aggregated simulated distribution. That will increase their interest in the process. The benefit is that if students can identify their individual samples and where they are located within the sampling distribution, they can measure as well the quality of their samples through the calculation of standardized scores.

Secondly, when students identify their own samples, and how close or far they are from the parameter of interest, this will allowed them to understand how the confidence level plays a part in the significance of the samples created. This will provide the intuition behind confidence intervals, hypothesis testing and significance of results.

The next stage of this project is to perform an experiment. To set up the experiment, we will randomly assigned different treatments to 4 groups with a pre-test and post-test to measure effects

on learning outcomes. We will randomly assigned the groups to each intervention and then control for systematic differences among the elements within these groups through control testing.

One of the groups will be randomly assigned as control group, for which we will teach this topic by demonstration without simulation. Then, treatment 1 will have a simulation by demonstration of the professor; treatment 2 will have simulation generated by each student at the individual level; and treatment 3 will have a simulation by collaboration for which students will create a simulation as a group effort. The objective of this experiment is to isolate and identify if there are differences on these approaches to teach the CLT and the SD principles. The hypothesis is that simulation by collaboration is the most effective approach to help students understand these statistics concepts, which will be assessed with a knowledge test. The procedure is to give all the groups the same assessments and use the same teaching methodology, to control for other differences and to validate if any significant effect is identified on students' outcomes after using the simulation with collaboration (treatment 3).

5. Conclusions

Through simulation, collaboration among peers, experimentation, and a hands-on-approach, students will understand and they will trust the use of a random sample as a source for inference for the population, under the framework of the sampling distributions and the CLT, making it applicable for any population. The first stage of this project, which has already being tested in the classroom, has demonstrated that the simulation with collaboration provides a good setting to learn these core statistics concepts. The second stage of this project, which implied the creation of an online simulation to improve the effectiveness of this exercise, will be tested using a robust experimental design. Its results will allow us to provide evidence either to validate or to reject the idea that a simulation with collaboration is an effective learning tool to teach core statistics principles. This is a future research path aiming to contribute further to the use of simulations in the statistics field and in higher education.

Author Contributions: The idea of the simulation and the application of the experiment was done by M.E.B. The development of the online application was contracted to Sophie Labs, but the author maintains intellectual property of the simulation with collaboration.

Acknowledgments: Funds to develop this application were provided through the Instructional Improvement Awards of the Center for Teaching and Learning of the University of Texas at Dallas.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Cobb, P.; McClain, K. Principles of instructional design for supporting the development of students' statistical reasoning. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*; Ben-Zvi, D., Garfield, J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2004; pp. 375–396.
2. Hodgson, T.; Burke, M. On simulation and the teaching of statistics. *Teach. Stat.* **2000**, *22*, 91–96.
3. Watkins, A.E.; Bargagliotti, A.; Franklin, C. Simulation of the sampling distribution of the mean can mislead. *J. Stat. Educ.* **2014**, *22*, 1–21.



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).