

An Interactive Tool for Teaching the Central Limit Theorem to Engineering Students

Anas Basalamah

Department of Computer Engineering
Umm Al-Qura University
Makkah, Saudi Arabia

Abstract—The sole purpose of this paper is to guide students in learning the introductory statistical concepts, such as, probability distribution and the central limit theorem (CLT) in an intuitive approach through an interactive tool. When a used data has different probability distributions, this paper intends to clarify the notions of the CLT and the use of samples in the hypothesis testing of a population by demonstrating step-by-step procedures and hands-on simulation approach. This paper discusses the relationship between the sample size and the nature of the sampling distribution, which is a vital element of the CLT, in different population distribution using the developed interactive tool. Finally, the impact of the developed interactive tool is measured via a survey experiment that illustrated the success of the developed tool in teaching the CLT.

Keywords—Probability distribution; CLT; population; interactive tool; sampling distribution

I. INTRODUCTION

The fundamental concept of statistics field in which students more often face difficulties to understand is the central limit theorem (CLT). The CLT provides an overview of using random sampling method for making an inference about any population. Therefore, it is considered as a vital concept in inferential statistics, a critical knowledge for any statistician, and one of the foundation concepts of any statistics course. In [1], authors highlighted to focus on developing the ideas of central statistics before moving on to the set of tools and procedures. The CLT is one of the central ideas of statistics. While learning it, students get puzzled more often to interpret the theory and implication of this important concept. This paper aims to create an engaging way of teaching and learning the theory and implication of the CLT for both teachers and students through explanation, simulation, and visualization. We built an web application (interactive tool) using html, css, and javascript for generating and visualizing the uniform, normal, positively skewed, and negatively skewed distribution with corresponding sampling distribution.

Students can collaborate by selecting their desired distribution, sample size, and number of samples during the simulation process in the class for giving them a part of the simulation process and creating an engaging environment in the class. Students can select different sample size and number of samples for a particular distribution that enables them observing the changes of sampling distribution more closely with the changes of these parameters. It helps the student to understand the impact of sample size on the sampling distribution. The sampling distribution is the probability distribution obtained from a large number of samples, which are drawn from a

particular population. It is the distribution of frequencies of all possible outcomes that could occur for a statistics of that particular population.

The pedagogical approach taken in this paper for teaching CLT is different comparing to the other approaches because of the features of the developed interactive tool, which creates an engaging environment for the students and ensures students participation alongside to make the understanding of the CLT clear to the students.

The remaining part of this paper is organised as the following: Section II discusses the background and motivation of this research topic along with the literature review, Section III studies the concept of the central limit theorem and different probability distributions, Section IV describes the importance of the central limit theorem, Section V shows the empirical demonstration of the CLT with the help of visualization using the interactive web-application tool, Section VI investigates the impact of the developed tool in teaching the CLT in the classroom, and Section VII concludes the paper.

II. BACKGROUND

To make valid statistical interference about the concept of the CLT and sampling distribution, students must get the opportunity to draw multiple samples [2]. This paper is highly motivated by the recommendation given in [2], hence it incorporates multiple samples in simulation to clear the logic of the CLT to the students. The simulation provides a realistic scenario to the students for intuitively understanding these concepts. Although, simulation is not exempt of issues [3], but it is the best way of teaching the CLT to the students.

In [4], [5], the authors investigated that the concept of CLT is not only difficult for the non-math major students, but also for math major students. They added that the central limit theorem is the most vital result conveyed theorem in the introductory statistic course since it includes many of the statistical inferences that are required for the later part of the course. Although, students are lost more often in understanding the logic behind the CLT. Therefore, it becomes a necessity presenting the concept in a intuitive way through simulation such that it can be easily understandable to the students.

Interactive learning experience has always been considered as the most effective pedagogical tools [6], [7], [8], [9], [10], [11] and the use of interactive environment is considered as the effective way of learning and visualizing a complex theory. There are a large number of simulation tools and online applets for learning the CLT as well [12], [6]. Most of these simulated

environments has complex interface, requires knowledge of programming, and has complex procedures, which makes students less attractive to these simulation applets. The time spent to these applets for understanding different distributions and the CLT is comparatively higher and less engaging. The sampling distribution of any given population always follows normal distribution. To clear this concept, this paper brings all the distributions along with various sample sizes together in a single interface without showing the background codes that helps students get the idea easily in the shortest time. This paper intends to bring a simulated interactive teaching applet, which can be used both online and offline by the students and can meet all the above criteria.

In sum, the above stated context motivates us working on the same problem and providing a more interactive method of teaching CLT to the students. The interactive tool designed for the students has an interactive interface that allows the students to take different size and different number of samples iteratively for plotting their respective sampling distribution. As a result, students get an overview on how the estimated means of different samples with a particular sample size get centralized and form a bell-shaped curve, which is the principle of the CLT. The following section investigates different probability distribution and the concept of the CLT along with its properties.

III. POPULATION DISTRIBUTIONS, SAMPLING DISTRIBUTION, AND THE CLT

A. Population Distribution

The central limit theorem states that the sampling distribution of a given population forms a bell-shaped curve or follows normal distribution regardless the variable's distribution in the population. The distribution of a variable can follow different probability distributions, i.e. uniform distribution, normal distribution, positively skewed distribution, and negatively skewed distributions (see in Fig. 1).

The distribution of a variable is the distribution of the random sample that is drawn from the population. The CLT acts on all the probability distribution that have a finite variance. Therefore, the CLT cannot be applied on Cauchy distribution because of having infinite variance. In addition, the CLT acts on independent and identically distributed variables, where the value of a particular variable does not depend on the values of other variables. But the distribution of all these variables must remain constant throughout the measurement process.

B. Sampling Distribution

The sampling distribution is the mean of the randomly drawn samples from the population. For example- lets assume that someone draws a sample with a fixed sample size from a given population, then calculates its mean and plots it on a histogram. If this process is repeated many times, the produced histogram displays the distribution of the sample mean. This mean distribution is considered as the sampling distribution in statistics.

The shape of the sampling distribution depends on the sample size and number of samples. For different sample

size the shape of the sampling distribution differs. Similar observation found with the number of samples. The following sections discusses these phenomena in detail.

C. The Central Limit Theorem

The central limit theorem states that the sampling distribution of sufficiently large size of samples drawn with replacement from a given population that has the mean μ and standard deviation σ , will form an approximated normal distribution curve. This statement is true regardless the distribution of the population. Usually, a sufficiently large sample size is the samples having size of 30 or more ($n \geq 30$). If the population distribution is normal, the CLT is true for smaller size of samples as well. If the population distribution is strongly skewed, it may require larger size and number of samples. The relationship between the sample size and the shape of the sampling distribution is clarified in section V.

Also, the CLT statement holds correct for the binomial population distribution, provided that $\min(np, n(1-p)) > 5$, where p is the probability of success and n is the sample size. As a result, normal probability distribution can be used to quantify uncertainty for making inferences about a population mean.

1) *Properties of CLT*: There are two attribute of any distribution, namely- the mean (μ) and the standard deviation (σ). The sampling distribution converges the normal distribution, when its mean is equal to the population mean and the standard deviation is σ/\sqrt{n} . The standard deviation σ decreases by \sqrt{n} with the increment of the sample size n .

In sum, the sampling distribution approximates the normal distribution with the increment of sample size and the spread of the distribution suppressed. These properties have significant implications, which will be discussed in the later sections of this paper.

IV. IMPORTANCE OF THE CLT

The central limit theorem is important because of the following to reasons, namely- normality assumption and precision estimates.

A. CLT and the Normality Assumption

The normality assumption is very essential in statistics for parametric hypothesis testing of the mean, i.e. t-test. The CLT supports the assumption as it states that the sampling distribution of any population can approximate a normal distribution. This critical implication of the CLT allows the hypothesis testing even if the data is non-normally distributed. However, the testing is allowed if and only if the sample size is large enough. Because, a non-normally distributed data also behaves like a normal distribution for larger sample size.

Moreover, parametric tests of the mean are robust from the normality assumption when the sample size is sufficiently large. This is also a contribution of the central limit theorem.

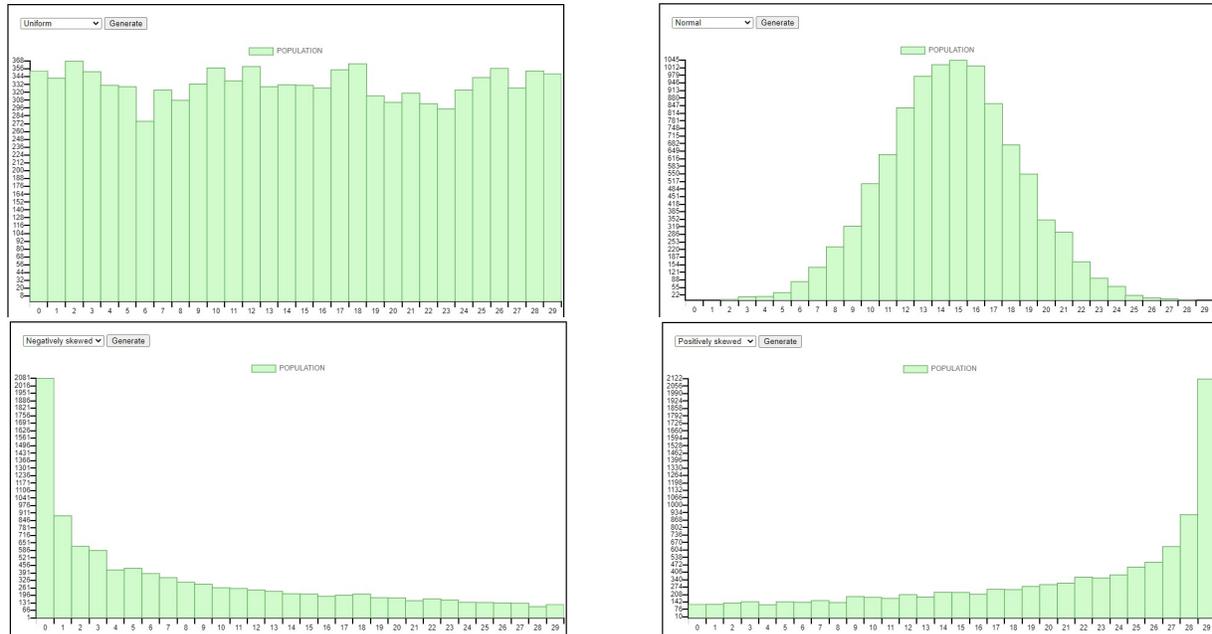


Fig. 1. Different Population Distributions, Namely- uniform, Normal, Negatively Skewed, and Positively Skewed Distribution.

B. Precision of Estimates

In all the figures, the sampling distribution of the mean clusters around the population and becomes denser as the sample size increases (see in Fig. 2, 3, and 4). This property of the Central Limit Theorem is relevant when using samples to estimate the mean for the entire population. The larger the sample size, it is more likely that the sample mean will be closer to the actual population mean. In other words, the estimate is more accurate.

On the other hand, the sampling distribution of the mean is much wider for smaller sample sizes. When the sample size is small, it is not uncommon for the sample mean to be away from the actual population mean. In this case, the estimate will be less accurate. Finally, understanding the Central Limit Theorem is important when relying on the validity of the results and assessing the accuracy of the estimation. We should use a large sample size to meet the normality assumptions and get more accurate estimates even if the data is non-normally distributed.

V. EMPIRICAL DEMONSTRATION OF CLT

This section demonstrates the central limit theorem with respect three different different population distribution, namely- uniform distribution, normal distribution, and severely skewed distribution. For illustrating the impact of CLT on severely skewed distribution, we consider the negatively skewed distribution. All the population distributions taken for the demonstration are shown in Fig. 1.

A. The CLT with Uniform Distribution

At first, we consider the uniform distribution shown in Fig. 1. We take the sampling frequency distribution for the sample size less than 30 and greater than 30 in Fig. 2, respectively. We observe that the obtained sampling distribution for

the sample size of 10 is ranging from 6.9 to 22.7, whereas the sampling distribution for the sample size of 30 ranges from 6.9 to 22.7. Besides, the shape of the sampling distribution is more normal using the sample size 50 comparing to the sample size 10.

Therefore, we can conclude that if we increase the sample size, the sampling distribution gets more tighten and centralized in uniform distribution.

B. The CLT with Normal Distribution

Similarly, we check the behaviour of sampling distribution using a normal distribution that is depicted in Fig. 1. We consider sample size of 10 and 50, respectively, and observed the similar conclusion as uniform distribution. The demonstration of CLT using normal distribution is depicted in Fig. 3. We observe that the obtained sampling distribution for the sample size of 10 is ranging from 11.1 to 17.9, whereas the sampling distribution for the sample size of 30 ranges from 12.68 to 16.48.

Hence, it is clear that the sampling distribution for larger sample size is more centralized comparing to the smaller sample size. We further observe that sampling distribution for the normal population using sample size of 10 is more tighten and centralized comparing to the sampling distribution of uniform population using sample size of 50 (see in Fig. 2 and Fig. 3).

C. The CLT with Skewed Distribution

Finally, we take a highly skewed population distribution, i.e. negatively skewed distribution depicted in Fig. 1, for testing the sampling distribution behaviour and investigating the CLT. The same process is followed for the skewed population analysis as before. Two sample size of 10 and 50 are considered. The obtained sampling distribution figures for both the sizes

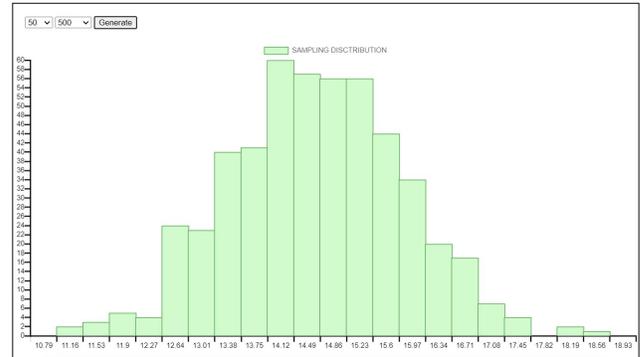
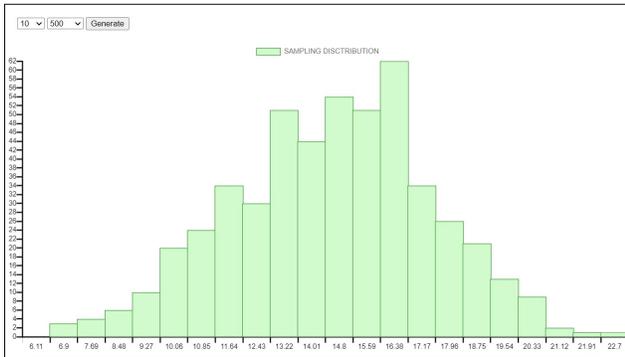


Fig. 2. The Central Limit Theorem Explained with Uniform Distribution.

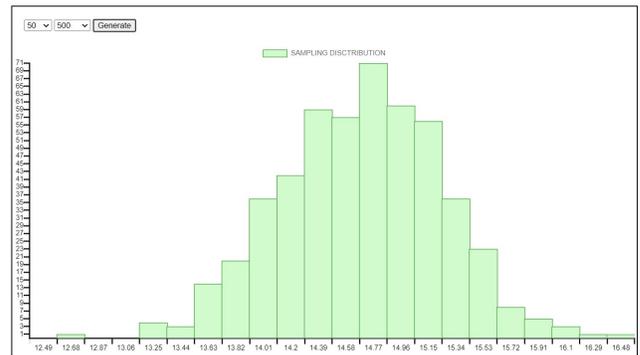
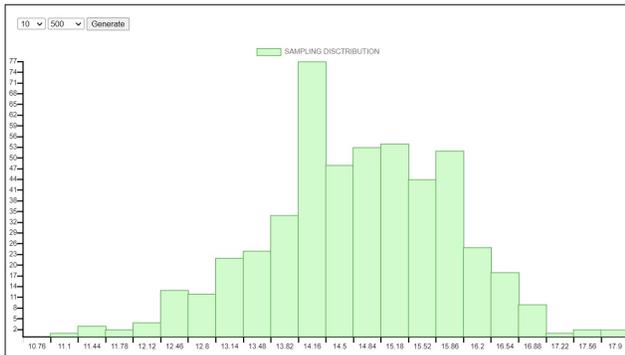


Fig. 3. The Central Limit Theorem Explained with Normal Distribution.

are depicted in Fig. 4. Similar observation is found, i.e. if the sample size increases the obtained distribution gets tightened and the shape approximates a normal bell-shape.

The range of sample mean obtained for the size 10 is from 1.6 to 16.6 and for the size 50 is from 4.92 to 11.92, which supports the previous statement and justifies the central limit theorem.

VI. IMPACT

We have developed an interactive teaching applet using HTML5, CSS, and Java-script. The interactive tool is a dynamic, responsive, and device independent web-application that can be opened by double clicking on the index.html file. Students only need to have a web-browser installed in their device. The teaching applet has an attractive and simple user interface, where user can generate different population distributions, can select random samples of different sizes, and can plot their sample mean. Student can add a single sample to the sampling distribution iteratively or can add large number of samples at a time, which enables them to visualize the changes occurred in the sampling distribution after adding each sample to the sampling distribution. Students can easily absorb the motto of the CLT by watching how the sampling distribution becomes tightened after the addition of new sample means to the distribution.

We observed that all the students got the gist of the CLT easily after taking the class with the help of the developed interactive teaching applet. Therefore, we can claim that the

interactive tool developed by us can successfully be used during the teaching/learning process of the CLT.

VII. CONCLUSION

This paper intends to provide a simulation based overview of the central limit theorem, which intuitively helps the students for cracking the concept. The relationship of the sample size and the sampling distribution is well illustrated throughout the empirical analysis section, which is an essential concept for the students as they go deep into their statistics course. The last section shows the impact of this study and the developed interactive tool about how well the students were able to absorb the concept after using the developed tool.

REFERENCES

- [1] P. Cobb and K. McClain, "Principles of instructional design for supporting the development of students' statistical reasoning," in *The challenge of developing statistical literacy, reasoning and thinking*. Springer, 2004, pp. 375–395.
- [2] T. Hodgson and M. Burke, "On simulation and the teaching of statistics," *Teaching Statistics*, vol. 22, no. 3, pp. 91–96, 2000.
- [3] M. E. Brussolo, "Understanding the central limit theorem the easy way: A simulation experiment," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 2, no. 21, 2018, p. 1322.
- [4] M. L. Lunsford, G. H. Rowell, and T. Goodson-Espy, "Classroom research: Assessment of student understanding of sampling distributions of means and the central limit theorem in post-calculus probability and statistics classes," *Journal of Statistics Education*, vol. 14, no. 3, 2006.
- [5] A. E. Watkins, A. Bargagliotti, and C. Franklin, "Simulation of the sampling distribution of the mean can mislead," *Journal of Statistics Education*, vol. 22, no. 3, 2014.

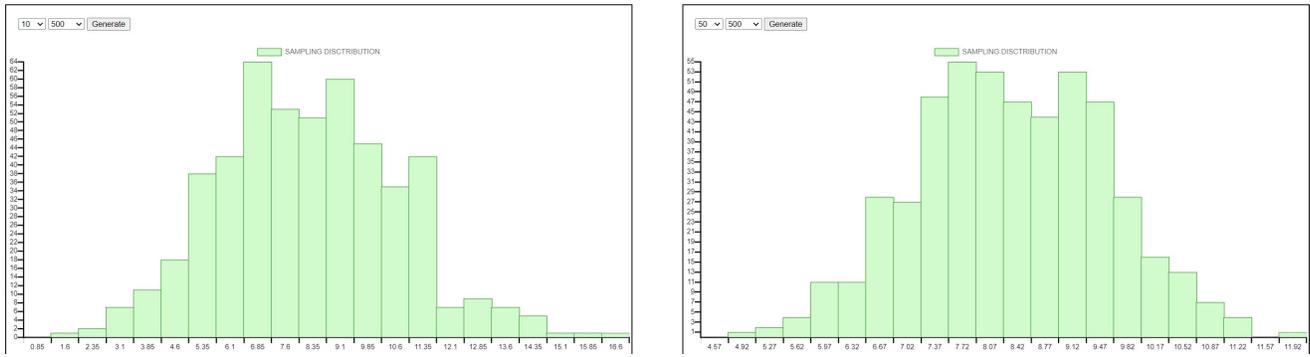


Fig. 4. The Central Limit Theorem Explained with Negatively Skewed Distribution.

- [6] S. P. Gordon and F. S. Gordon, "Visualizing and understanding probability and statistics: graphical simulations using excel," *PRIMUS*, vol. 19, no. 4, pp. 346–369, 2009.
- [7] J. Garfield, "Teaching statistics using small-group cooperative learning," *Journal of Statistics Education*, vol. 1, no. 1, 1993.
- [8] K. A. Carlson and J. R. Winqvist, "Evaluating an active learning approach to teaching introductory statistics: A classroom workbook approach," *Journal of Statistics Education*, vol. 19, no. 1, 2011.
- [9] T. J. Pfaff and A. Weinberg, "Do hands-on activities increase student understanding?: A case study," *Journal of Statistics Education*, vol. 17, no. 3, 2009.
- [10] R. K. Steinhorst and C. M. Keeler, "Developing material for introductory statistics courses from a conceptual, active learning viewpoint," *Journal of Statistics Education*, vol. 3, no. 3, 1995.
- [11] D. M. Jamie, "Using computer simulation methods to teach statistics: A review of the literature," *Journal of Statistics Education*, vol. 10, no. 1, 2002.
- [12] E. Ruggieri, "Visualizing the central limit theorem through simulation," *PRIMUS*, vol. 26, no. 3, pp. 229–240, 2016.