

Teaching Quantitative Methods Vignettes

Using Dynamic Graphics to Teach the Sampling Distribution with Active Learning

Kathryn J. Hoisington-Shaw^a and Jolynn Pek^a

^aThe Ohio State University



Abstract ■ The sampling distribution and the Central Limit Theorem (CLT) are the basis for many statistical procedures and inferences. Despite their ubiquitous nature in statistics, these concepts are some of the most abstract and difficult for students to understand. To foster a deeper understanding of these concepts, a web-based application was created that uses dynamic graphics to illustrate the concepts and engage students with active learning. We provide an outline of three in-class activities using the web application to promote the learning of population distributions, simple random sampling, sampling variability, the idea of a statistic, the sampling distribution, the Law of Large Numbers, and the CLT. These in-class activities tie the concepts together and place emphasis on their role as building blocks of statistical inference. By linking abstract theoretical concepts together before introducing statistical inference, the web application facilitates statistical thinking that students can utilize both inside and outside the classroom.

Keywords ■ dynamic graphics, sampling distribution, active learning, web application, Central Limit Theorem.

hoisington-shaw.1@osu.edu

[10.20982/tqmp.17.2.v001](https://doi.org/10.20982/tqmp.17.2.v001)

Acting Editor ■ Sébastien Béland (Université de Montréal)

Reviewers ■ Three anonymous reviewers.

Concept to be presented

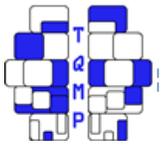
Sound statistical analyses are key to valid research. Statistical thinking, that is, thinking of data in terms of its underlying statistical properties and how these properties interact with one another, is a critical skill necessary for prudent statistical inference (Doerr & English, 2003; Wild & Pfannkuch, 1999). Teaching statistical thinking in an accessible manner is the first step in helping cultivate new researchers and providing students with the resources to become savvy statistical consumers. Furthermore, it is important to foster statistical thinking beyond simply following ritualistic procedures in a cookbook to facilitate critical thinking behind research results (Tukey, 1969).

One of the core tenants in statistical inference is the sampling distribution. The sampling distribution is essential to the Central Limit Theorem (CLT), which underlies the normal distribution's ubiquitous role in statistical inference. However, these concepts lack intuitive inter-

pretations, and their abstraction presents a challenge for many students (see Sotos, Vanhoof, Noortgate, & Onghena, 2007, for a review on the many misconceptions surrounding these topics). The purpose of this paper is to present three activities that make use of an interactive web-based utility to illustrate the sampling distribution, the Law of Large Numbers (LLN), and the CLT. The activities scaffold one another, fostering understanding of these related concepts and facilitates statistical thinking in a manner that aligns with recommendations in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (GAISE College Report ASA Revision Committee, 2016; see also Wood, Mocko, Everson, Horton, & Velleman, 2018).

The target audience for these activities is undergraduate introductory statistics students, and it is estimated that these activities will span the course of two or three weeks of class time.

First, we briefly review statistical theory behind sam-



pling distributions and CLT. Next, we give an overview of the web application and then outline the concepts each activity highlights. Finally, we describe the learning activities connected to each statistical concept and highlight their link to statistical inference.

Statistical Theory

A population is said to contain all data relevant to a phenomenon under study. This population is often represented by a probability distribution with (unknown) parameters (e.g., normal distribution with mean and standard deviation parameters μ and σ , respectively). In a research study, because it is usually impossible to observe the entire population, a sample (or subset of observations) is taken from the population. Sample statistics (e.g., the mean and standard deviation denoted as \bar{x} and s , respectively) summarize or describe the sample. These statistics are also estimates of their population parameters. Each statistic calculated from the sample can be conceived as a single point in a sampling distribution. The sampling distribution quantifies the extent and form of uncertainty across an infinite number of samples with the same number of observations (sample size, n). The variability in a sampling distribution is termed sampling variability (i.e., the variation between each statistic that summarizes data from a unique sample drawn from the population). The sampling distribution of a statistic allows for statistical inference; confidence intervals (CIs) and p -values are calculated from the sampling distribution.

The LLN states that as n increases, the statistic approaches its population parameter value (i.e., $n \rightarrow \infty$, $p \rightarrow \pi$, where p is a sample probability and π is the population probability in a Bernoulli or binomial distribution). Stated differently, the LLN implies that as n increases, the sampling distribution of the statistic shrinks in variability while the statistic converges to the population parameter.¹

The CLT governs the shape of sampling distributions. Because the sampling distribution is often unobserved (and therefore taken to be hypothetical), the CLT provides a way to make theoretical claims about the sampling distribution. The CLT states that as $n \rightarrow \infty$, regardless of the shape of the population distribution, the sampling distribution of a statistic approaches the normal distribution. This theorem allows one to assume normality of the sampling distribution of the sample means even when the population (and sample data) is not normally distributed. Thus, the CLT is the basis for many statistical tests taught in early statistics courses, such as the z-test, t-test, and F-test. The sampling distribution, LLN, and CLT will be demonstrated in the web application to follow.

Web Application

The web application can be accessed online at https://seeing-statistics.shinyapps.io/Sampling_Distribution_and_CLT/.² The application makes use of Monte Carlo simulations to teach statistical concepts, which has been found to be beneficial in learning (Revelle, 2020). Specifically, random samples of data are drawn from particular population distributions to simulate the concept of collecting a simple random sample of data for a study. Students can choose to sample data from three different population distributions: normal, Bernoulli, and exponential. Each population distribution is accompanied by its own contextual example to highlight the use of different distributions to categorize different types of data. Having students engage in active learning by interacting with dynamic graphics that illustrate statistical concepts helps foster a deeper learning and understanding (Wainer & Velleman, 2001).

The application is laid out with three tabs for three distinct distributions (located in the left sidebar), with each tab's window displaying four separate boxes (see Figure 1 for a screenshot of the application layout). The top-left box describes the population distribution accompanied by a contextual example with brief instructions on using the application. The top-right box presents the population distribution that the samples are randomly drawn from. The bottom-left box has a slider to determine the size of the simple random sample (n) drawn from the population. After choosing the desired n , clicking the Draw Sample button returns a histogram of the randomly drawn sample. For the normal and exponential distributions, the sample mean is represented by a vertical dashed reference line on the histogram. For all distributions, the sample mean and standard deviation are reported below the histogram. Lastly, the bottom-right box displays the sampling distribution with a histogram that is updated with each sample drawn (the mean is represented by a vertical dashed line). The mean values of each sample, as well as the mean and standard deviation of the sampling distribution, are reported below the histogram and are also updated with each draw.

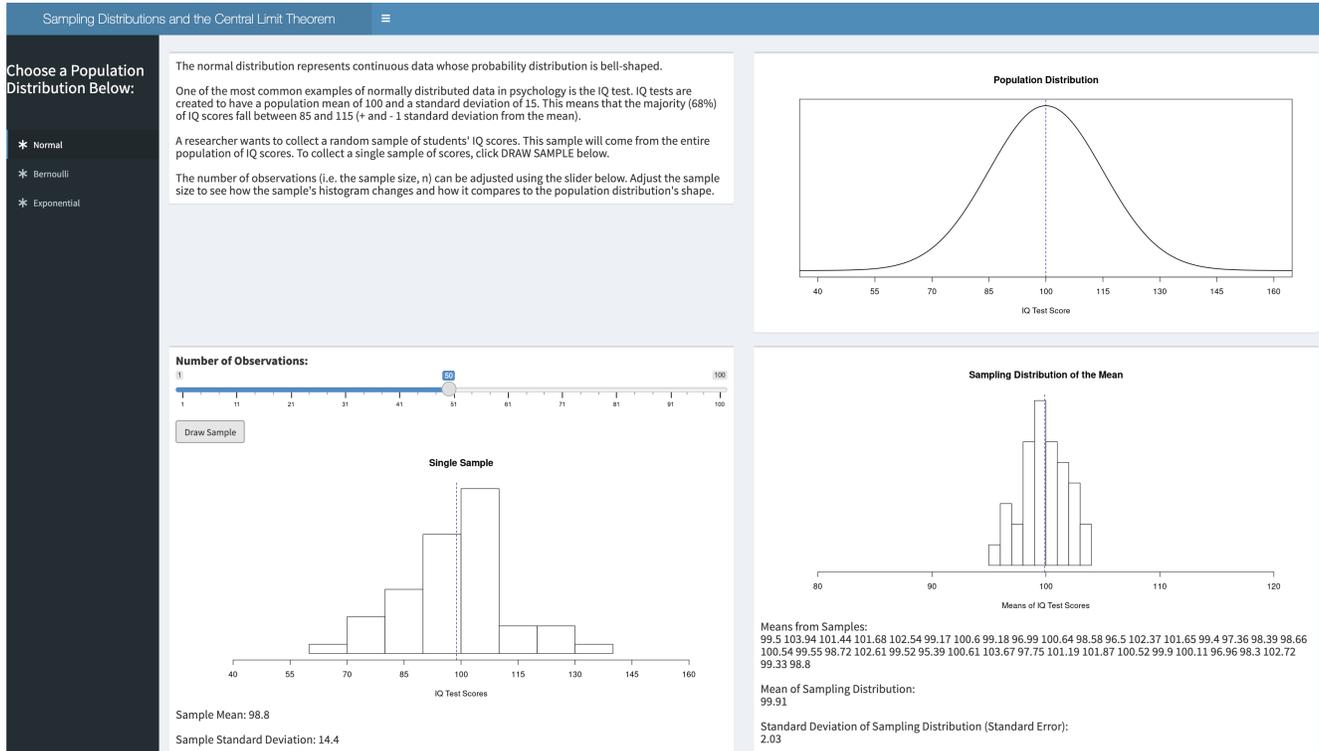
Activities Using the Web Application

The web application and its accompanied activities were created to align with the six GAISE recommendations (GAISE College Report ASA Revision Committee, 2016). All activities were designed to cultivate statistical thinking (recommendation 1) by focusing on a deeper, conceptual understanding of the theoretical topics (recommendation 2). These activities use realistic examples (recommenda-

¹The standard deviation of the sampling distribution of a statistic is called the standard error when the standard deviation σ is estimated by s .

²If modifications to the application are sought, contact the first author for the application's source code.

Figure 1 ■ Web Application Layout. Seen is a screenshot of the application with the different population distributions located in the left sidebar and the four boxes displaying the example description, population distribution, sample distribution, and sampling distribution.



tion 3) to tie statistical theory to the real world, promoting active learning (recommendation 4) through engagement with the application (recommendation 5). Activity assessments that can be used in class are provided in Appendices A-C (recommendation 6).

Activity 1: Population vs. Sample

The goals for the first activity are to distinguish between the population and a sample, illustrate simple random sampling, and highlight the role of n in sampling variability. The first activity features the normal distribution by using IQ test scores as an example of a known population. Students can randomly draw samples of IQ test scores from the population distribution ($\mu = 100, \sigma = 15$). Differences between the graph of the population distribution and the histogram are due to sampling. When several random samples are drawn, differences in the histograms (e.g., means, standard deviations) illustrate sampling variability. Students can also change the n to be between 1 and 100. Sample size influences sampling variability in that samples with larger n should look more similar to the

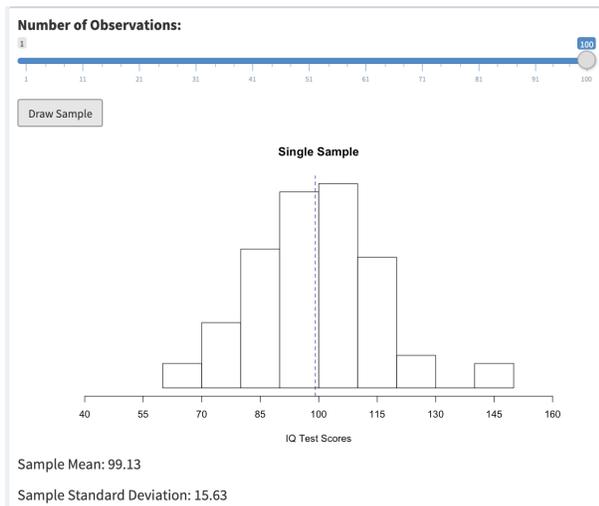
population normal distribution than samples with small n (e.g., see Figure 2). An example worksheet to walk students through these activities can be found in Appendix A.

Activity 2: The Sampling Distribution and LLN

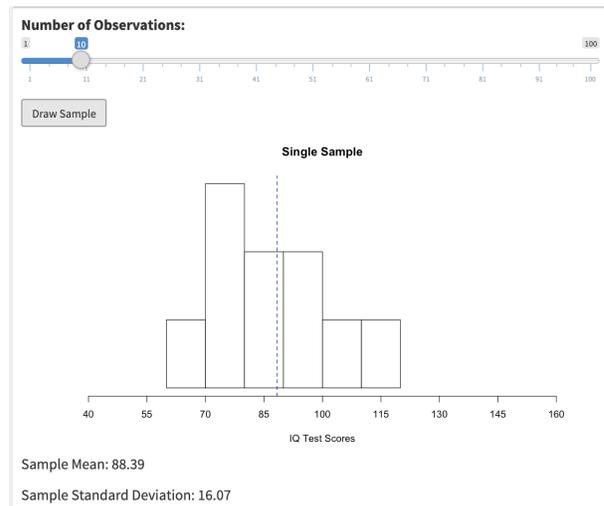
The goals of the second activity are to introduce the idea of a statistic as a value that summarizes a sample, illustrate the sampling distribution of a statistic, and show the influence of n on the statistic under the LLN. The second activity features the Bernoulli distribution that reflects the probability that a person shows up to take an IQ test as the “success” event ($\pi = .80$). Before beginning the activity, students can be introduced to the idea that different distributions will represent different types of events (e.g., normally distributed data vs. Bernoulli). Similar to the first activity, students can randomly sample from the Bernoulli distribution, reviewing what was learnt about sampling variability in relation to n with the normal population distribution. Instead of focusing on a single sample and its mean (p in this case), students should be encouraged to continue to draw samples (of the same n) and note how the sam-

Figure 2 ■ Example of Activity 1 With the Web Application. The screenshots display samples taken from a normally distributed population with parameters $\mu = 100$ and $\sigma = 15$. The left screenshot shows the sample with an $n = 100$ and a sample mean $\bar{x} = 99.13$ (which is very close to the population mean of $\mu = 100$); whereas the right screenshot shows a sample mean $\bar{x} = 88.39$ with an $n = 10$.

(a)



(b)



pling distribution’s histogram changes. Because the sampling distribution is dependent on the sample size, changing the sample size resets the sampling distribution’s histogram. By varying n , it can be shown that the LLN also affects the sampling distribution in which larger n has a tighter sampling distribution compared to smaller n (Figure 3). Finally, it can be illustrated that LLN generalizes across the form of population and the statistic (\bar{x} or p), and is at work when the statistic approaches the population parameter (μ or π). An example of this student learning activity can be found in Appendix B.

Activity 3: The Central Limit Theorem

The goal of the third activity is to introduce the CLT. This activity uses an example of the exponential distribution to represent the time that passes between a person taking an IQ test for a second time (with the population parameter $\lambda = 1/2$). The parameter λ in an exponential distribution represents the rate of change. Suppose the scale of time is 1 year. Then, $\lambda = 1/2$ quantifies the average time between the first and second IQ tests as $1/\lambda = 2$ years. All the concepts introduced in the previous activities can be reviewed with the exponential distribution. The exponential distribution represents time data that is distinct from normal or Bernoulli distributed data. With simple random sampling, the concepts of sampling variability and the LLN in relation to the size of n can be reviewed. We recommend introducing the CLT by pointing out that as n is in-

creased, and the number of samples is large (around 30 or so draws), the histogram of the sampling distribution begins to appear bell-shaped (i.e., normal; see Figure 4). Stated differently, at larger sample sizes, the sampling distribution of the mean converges faster to the normal distribution. It is important to remind students here that the CLT is theoretical, and the sampling distribution is hypothetical and often unobserved. Individually drawing samples one at a time aids in emphasizing to students what is meant to draw a simple random sample from the population. It should also be highlighted that beyond large-scale replication studies (where study n s are different) researchers are unable to take many, many samples unless conducting quality control or operations research in engineering. However, the larger the n , the more likely CLT has occurred to justify using a normal sampling distribution for inference (e.g., see Pek, Wong, & Wong, 2017). Lastly, it can be highlighted that the CLT applies to population distributions of unknown form by showing that it occurs across the three different population distributions. An example of this activity for students can be found in Appendix C.

Connecting the Activities to Statistical Inference

The statistical concepts of the sampling distribution, LLN, and CLT are key to understanding statistical inference. In our teaching experience, students have struggled to understand them in the absence of data. Although the web application makes use of simulated data, it is important for

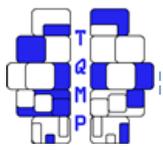
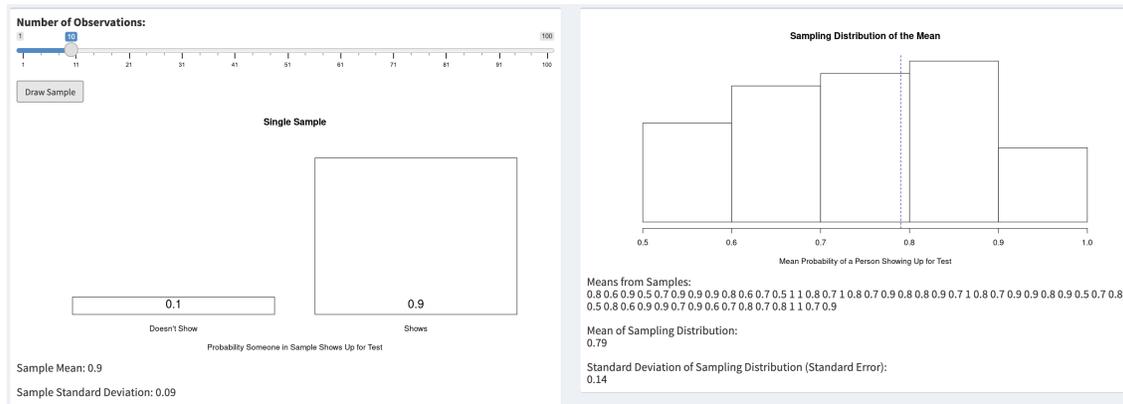
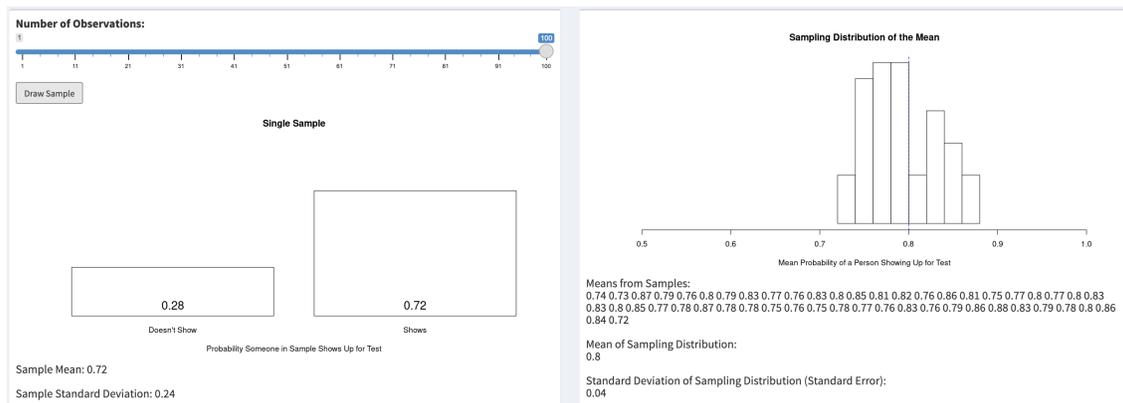


Figure 3 ■ Example of Activity 2 With the Web Application. The top screenshot displays the sampling distribution of the mean of samples of size $n = 10$, whereas the bottom screenshot displays the sampling distribution of the mean of samples of size $n = 100$ (both taken from a Bernoulli population with parameter $\pi = .80$). Each sampling distribution is made up of 50 samples.

(a)



(b)



students to realize that only a single sample is drawn for a single study in practice.

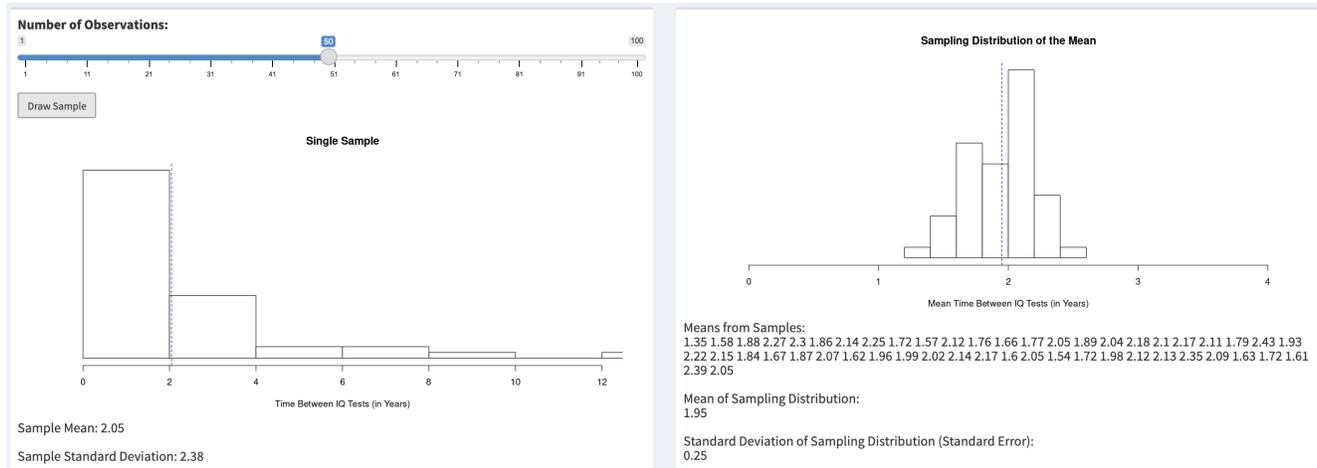
Once students have a strong grasp on the concepts underlying inference as outlined above, these activities can be used to link the sampling distribution to more concrete examples later in the course (e.g., t-test). The sampling distribution, LLN, and CLT come into play when a researcher tests whether a population mean is different from 0. A randomly drawn sample that is large enough to elicit the use of the LLN is more likely to result in a mean that is similar to the population mean. Large sample sizes have more precise estimates (tighter CIs) because the variance of the sampling distribution is smaller than the sampling distribution of smaller sample sizes. An estimate of the mean that is both precise and close to the population value is desirable for ensuring that the value used in the statistical test is rep-

resentative of what the researcher intended to evaluate. The CLT then allows the researcher to assume normality of the test statistic when testing the mean against 0, even if the data itself is not normally distributed. Therefore, the CLT allows a researcher to assume normality of the statistic, in which its sampling distribution represents sampling error, conduct a test based on this normality, and have an estimate with precision based on the sample size used.

Conclusion

Students often struggle to grasp many of the abstract and theoretical concepts present in introductory statistics courses (Ramsey, 1999). The three activities outlined in this paper make use of a dynamic web application to visually illustrate and show students first-hand the effects of n , sampling variability, the sampling distribution, the LLN, and

Figure 4 ■ Example of Activity 3 With the Web Application. This screenshot shows the sampling distribution of the mean of 50 samples (with $n = 50$) taken from an exponential population distribution with parameter $\lambda = 1/2$. Although the data is exponentially distributed, the sampling distribution of the mean is normally distributed.



the CLT. These concepts lay the groundwork for statistical inference and can be tied to inferential devices discussed later in the course (e.g., p -values and CIs). Providing a link between statistical theory and inference is key in promoting statistical thinking because it teaches students to think critically of how different situations (e.g., violations of statistical assumptions) affect the interpretations and conclusions that can be made based on data.

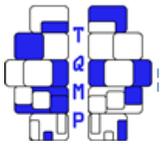
The use of dynamic and interactive graphics has become more popular in both data analysis and reporting (Grant, 2013). Dynamic graphics can show real-time changes within data, present data with multiple dimensions, and are much better than a traditional static graph in conveying abstract and complex information (Wainer & Velleman, 2001). As dynamic graphics become a more common and easily accessible format for the general public to consume (e.g., Allen et al., 2021), it is imperative to utilize these tools for teaching statistical concepts (e.g., Forbes, Chapman, Harraway, Stirling, & Wild, 2014). Beyond visualizing changes in data, graphics that engages the user to interact, or “converse” with data (Wainer & Velleman, 2001) are known to facilitate active learning, which is a key objective outlined in the GAISE College Report (GAISE College Report ASA Revision Committee, 2016).

In the spirit of accessibility and affordable learning, we hope that our freely available web application will be widely used in the classroom to promote statistical thinking via active learning in traditional, hybrid, and fully online courses. Although we focused on teaching abstract concepts that lead up to inference, dynamic graphics can

be expanded to teach other abstract statistical concepts (e.g., see McClelland, 1999). Freely accessible web applications in the form of dynamic graphics, built on Monte Carlo methodology, promise to enhance teaching and learning in statistics, especially in a future that is trending towards online learning and hybrid classrooms (Anshari, Alas, Yunus, Sabtu, & Hamid, 2016).

References

- Allen, J., Almukhtar, S., Aufrichtig, A., Barnard, A., Bloch, M., Cahalan, S., ... Marcus, I. (2021). Coronavirus world map: Tracking the global outbreak. *The New York Times*, 2021. Retrieved from <https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html>
- Anshari, M., Alas, Y., Yunus, N., Sabtu, N. I., & Hamid, M. H. (2016). Online learning: Trends, issues and challenges in the big data era. *Journal of e-Learning and Knowledge Society*, 12(1), 121–134.
- Doerr, H. M., & English, L. D. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for Research in Mathematics Education*, 34, 110–136. Retrieved from <http://doi.org/10.2307/30034902>
- Forbes, S., Chapman, J., Harraway, J., Stirling, D., & Wild, C. (2014). Use of data visualization in the teaching of statistics: A new zealand perspective. *Statistics Education Research Journal*, 13(2), 187–201.
- GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statis-*



- tics education college report. Washington: American Statistical Association. Retrieved from <http://www.amstat.org/education/gaise>
- Grant, R. (2013). Nathan yau. *Significance*, 10(1), 32–35. Retrieved from <http://doi.org/10.1111/j.1740-9713.2013.00637.x>
- McClelland, G. (1999). *Seeing statistics*. Belmont, CA: Duxbury Press. Retrieved from <http://www.seeingstatistics.com/>
- Pek, J., Wong, O., & Wong, C. (2017). Data transformations for inference with linear regression: Clarifications and recommendations. *Practical Assessment, Research & Evaluation*, 22(9). Retrieved from <https://doi.org/10.7275/p86s-zc41>
- Ramsey, J. B. (1999). *Why do students find statistics so difficult?* Finland: Paper presented at the International Association for Statistical Education Conference in Helsinki.
- Revelle, W. (2020). Teaching research methods using simulation. In J. L. Rodgers (Ed.), *Teaching statistics and quantitative methods in the 21st century* (pp. 217–237). Abingdon-on-Thames: Routledge.
- Sotos, A., Vanhoof, S., Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, 2(2), 98–113.
- Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, 24(2), 83–91. Retrieved from <https://doi.org/10.1037/h0027108>
- Wainer, H., & Velleman, P. F. (2001). Statistical graphics: Mapping the pathways of science? *Annual Review of Psychology*, 52(1), 305–335. Retrieved from <http://doi.org/10.1146/annurev.psych.52.1.305>
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67, 223–248. Retrieved from <http://doi.org/10.1111/j.1751-5823.1999.tb00442.x>
- Wood, B., Mocko, M., Everson, M., Horton, N., & Velleman, P. (2018). Updated guidelines, updated curriculum: The gaise college report and introductory statistics for the modern student. *Chance*, 31(2), 1–99. Retrieved from <http://doi.org/10.1080/09332480.2018.1467642>

Appendix A: (Activity 1) Population vs. Sample

Learning Outcomes:

1. Distinction between population and sample
2. Role of sample size
3. Sampling variability
4. Simple random sampling

Background:

- A population consists of all data relevant to the event being studied and is represented by a probability distribution (e.g., normal, Bernoulli, exponential) that reflects the probability a specific observation (or event) occurring.
- Most of the time we cannot collect data from all possible observations, so we take a sample (i.e., a small number of observations from the population).
- Consider the IQ test, whose population follows a normal distribution (bell-shaped curve) with a mean of 100 and a standard deviation of 15.

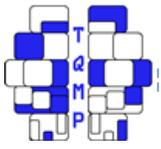
Active Learning Questions:

1. Choose the sample size to be 10 observations and click “draw sample.” What does the sample’s histogram look like? What is the mean?
2. Draw samples of size 25, 40, 75, and 100 and take note of the means and the shape of the histograms. How do these samples compare to the population distribution?
3. What role do you think sample size plays when drawing a random sample?
4. Draw three samples of size 30 – what are the means? Are they all the same? Why or why not?
5. If you are designing a study, why might you want the sample’s distribution to look similar to the population’s distribution?
6. Draws from the population are simple random samples. What would happen to the sample mean and histogram when the draws were not simple random samples?

Appendix B: (Activity 2) The Sampling Distribution

Learning Outcomes:

1. Statistic as a summary of a sample
2. Link between population, sample, and sampling distribution



3. Role of sample size
4. The Law of Large Numbers

Background:

- When you draw a sample, you can calculate a statistic to characterize/ describe the sample, such as the mean.
- Each time you draw a sample from the same population with the same sample size, and calculate the mean, these means make up the sampling distribution.
- The sampling distribution show the probability distribution (i.e., shape) of all possible values of a statistic (in this case, the mean).
- Consider the probability of someone showing up for their IQ test: either they do show up, or they do not. This distribution is called a Bernoulli distribution.

Active Learning Questions:

1. Choose the sample size to be 20 and click “draw sample.” What does the histogram of the sampling distribution look like?
2. Click “draw sample,” with the same sample size (20) 9 more times. You now have 10 means. What does the histogram of the sampling distribution look like now?
3. Change the sample size to be 50 and click “draw sample.” What happened to the histogram and why do you think this happens?
4. Now draw 20 more samples of size 50 – compare this sampling distribution to the sampling distribution in Question #2. What are the similarities and/ or differences between the two distributions?
5. Many studies are only conducted one time, meaning analyses are only performed on a single sample. What are your thoughts on this fact in relation to the sampling distribution?

Bonus: Repeat this exercise with both the normal and exponential distributions. What do you find?

Appendix C: (Activity 3) Central Limit Theorem

Learning Outcomes:

1. Link between samples, the sampling distribution, and the CLT
2. The CLT and different population distributions
3. The relationship between the CLT and statistical inference

Background:

- Each time a sample is drawn and the statistic of choice (e.g., mean) is added to the sampling distribution, the sampling distribution takes shape.
- Regardless of the probability distribution of the population and of the sample taken from said population, the sampling distribution follows the Central Limit Theorem (CLT).
- The CLT allows us to make inferences about the population using statistics (e.g., mean).
- Consider the time in between someone taking an IQ test and the next time they will take the test (assuming people take IQ tests on average every two years). This timing follows an exponential distribution.

Active Learning Questions:

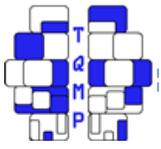
1. Choose a sample size of 50 and click “draw sample” 5 times. What does the histogram of the sampling distribution look like?
2. Click “draw sample” at least 25 more times with a sample size of 50. What shape does the histogram look like now?
3. Calculate the mean of the means in the sampling distribution – how does this compare to the population mean?
4. Repeat the above exercise with the Bernoulli distribution. What do you find?
5. Why might the sampling distribution and CLT be useful?

Citation

Hoisington-Shaw, K. J., & Pek, J. (2021). Using dynamic graphics to teach the sampling distribution with active learning. *The Quantitative Methods for Psychology, 17*(2), v1–v9. doi:[10.20982/tqmp.17.2.v001](https://doi.org/10.20982/tqmp.17.2.v001)

Copyright © 2021, *Hoisington-Shaw and Pek*. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 19/03/2021 ~ Accepted: 27/05/2021



Extended activity metadata

<i>Concept illustrated</i>	Sampling distribution	<i>Type of activity</i>	Online activity
<i>Prerequisite</i>	Basic understanding of probability, measures of central tendency, variability, statistical graphics	<i>Types of data</i>	Simulated data
<i>Co-requisite</i>	Population distributions, samples, simple random sampling, the Law of Large Numbers, and the Central Limit Theorem	<i>Computation by</i>	Web application created via R Shiny
<i>Suitable class size</i>	Any size	<i>Duration</i>	2-3 weeks of class time
