THE USE OF RANDOM DATA IN ONLINE DISCUSSION BOARDS TO PROMOTE STUDENT UNDERSTANDING OF SAMPLING DISTRIBUTIONS

Mary H. Bruce
Purdue University Global, USA
mbruce@purdueglobal.edu

With the recent pandemic, education systems responded to this challenge with reliance on virtual instruction and distance learning. As a result, instructors were pushed to create learning environments that foster meaningful learning through discussion boards and video conferencing. This paper seeks to illustrate how the use of random data in online discussion boards provides a setting rich in embedded statistical theory that serves as an engaging and fruitful environment for instruction in sampling distributions and hypothesis testing in an introductory statistics course.

INTRODUCTION

It is well-documented in the literature that the concept of sampling distributions in statistics education is met with challenges in conceptual understanding (Chance et al., 2004; Ozmen & Guven, 2019). Many of the earlier studies conducted on sampling distributions took place in traditional, face-to-face classes (Chance et al., 2004). Researchers have used technological tools to improve the learning of this difficult concept (Chance et al., 2004). While results are favorable, the author has often encountered difficulties with some simulations as students struggle to bridge the connection between the simulation interface and the desired learning objectives.

With the pandemic, participation in online learning soared globally. While many students will return to face-to-face learning later this year, projections are favorable for steady growth in the online sector, and hybrid learning will most likely be a more adaptive, viable replacement for strictly face-to-face environments (Bozkurt & Sharma, 2020, Xie & Siau, 2020).

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) is the standard framework used to pave the path for course design and instruction in the teaching of statistics. Recommendations in the college report include the following:

1. Emphasize statistical literacy and develop statistical thinking.
2. Use real data.
3. Stress conceptual understanding rather than mere knowledge of procedures.
4. Foster active learning in the classroom.
5. Use technology for developing concepts and analyzing data.
6. Use assessments to improve and evaluate student learning. (ASA, 2016)

DESIGN

The author served as a subject matter expert for a graduate online introductory statistics course. In designing the weekly discussion boards for this course, GAISE principles guided the creation of these collaborative forums. The author specifically wanted to implement the use of random data sets for varied reasons – the generated data would be unique to each student preventing the redundant, repetitive responses that often plague online discussion boards, the uniqueness would encourage each student to interact with the software and add individual value to the collective discussion, and the use of random data would provide the opportunity to simulate random sampling to foster discussion on differences in center and variability from sample distributions to sampling distributions. This distributional thinking through the use of technology and the interaction with actual data reflects key ideas in the GAISE framework. The author posits the control of the student in the random sampling process allows for a deeper, conceptual understanding of the creation of data distributions and the changes that occur with measures of center and spread. The collaboration is believed to successfully mimic the idea of repeated sampling to generate distributional thinking.

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

The course is a ten-week online course. Each week, the instructor has a one-hour live seminar to actively teach the students the content for the week using a platform that allows for an upload of a pdf file and screen writing on the pdf file during the session. There is a chat log enabling students to comment and ask questions. The author uses this time to teach new concepts, but also to facilitate the linking of concepts from week to week to enable students to make important and necessary connections. Specifically, the author pulls ideas that emerge from the online discussions and positions them to stimulate new thinking that leads into subsequent units. The discussion boards run weekly and serve as the primary means of communication, collaboration, shared cognitive processing, and formative feedback. In this course, students use Microsoft Excel as their primary software supplemented with StatKey, a statistical package that accompanies the text *Statistics: Unlocking the Power of Data with WileyPlus* (Lock et al., 2021).

SAMPLE DISTRIBUTION

In the first discussion, the primary objective is a typical introduction to descriptive statistics. The author created an Excel template using the Excel function =ROUND(NORM.INV(RAND(), mean, standard),0) to randomly generate 30 observations from a hypothetical population of IQs with a mean of 100 and a standard deviation of 15. The choice of variable here is flexible – any variable with a known normal distribution, mean, and standard deviation could be used to generate interest in a simulated, but real context. Students are instructed to type their name in a particular box on the worksheet to generate a new sample of 30 IQs and then do a special paste of the values to a new worksheet to perform their descriptive analysis. The discussion directions require the students to upload their Excel worksheets and discuss the central tendency, variation, and shape of their distributions in context. They are instructed to look at their minimum and maximum values to process the degree of variability within their sample.

Most of the students do very well with the basic requirements of this discussion. Video resources guide them with the execution of Excel, so even those who are not well-versed with the software show the ability to calculate summary statistics. Interpreting the measures of center and spread pose more of a challenge, although most students show the ability to distinguish center from spread and to discuss these measures in a realistic context. At this early point in the course, it is evident that some students rely on patent definitions for their interpretations, especially with the less familiar standard deviation.

As part of the discussion requirements, students are also expected to post a minimum of two substantive peer replies to their classmates. As the facilitator, this is where the author pushes the students' learning farther along the trajectory to prepare them for upcoming lessons. For instructors, this requires informal assessments of students' understanding to maximize their potential for new learning. In the author's opinion, online discussions sharpen the instructor's ability to assess individual students as glimpses into their thinking can be gleaned from their online responses. Once students display a solid understanding of descriptive analysis of a sample distribution, students are encouraged to compare their sample distribution with those from other classmates. Students are asked to consider whether all these samples came from the same population or from different populations. Another common prompt used by the instructor in this first discussion involves asking the students to take their sample mean, extend a distance of two standard deviation units in both directions, and see if a sorted list shows that most of their observations fall within that range.

SAMPLING DISTRIBUTION

The primary objective of the second unit of the course involves visual displays. Students learn appropriate tables and graphs for both qualitative and quantitative variables, although the emphasis is on quantitative variables. During the live seminar, the instructor starts with a review of the Unit 1 Discussion. Students are reminded that each one of them generated a random sample. A typical class size is around 40 students, so the instructor explains that we have 40 samples. The idea of whether these samples all came from the same population is brought up again as student illustrations are shown on the screen (Figure 1). As students are guided to compare means, medians, standard deviations, minima and maxima, the students typically agree that these samples did come from the same population as "the values all seem pretty close."

IQ	
Mean	96.9
Standard Error	3.034041
Median	96.5
Mode	76
Standard Deviation	16.61812
Sample Variance	276.1621
Kurtosis	-0.86112
Skewness	-0.00474
Range	62
Minimum	64
Maximum	126
Sum	2907
Count	30

IQ	
Mean	99.83333
Standard Error	2.836577
Median	99.5
Mode	104
Standard Deviation	15.53657
Sample Variance	241.3851
Kurtosis	1.192677
Skewness	0.392099
Range	72
Minimum	72
Maximum	144
Sum	2995
Count	30

Figure 1. Student Illustrations from First Discussion

Specifically, the proximity of the sample means is noted, so the instructor reminds the students of some of the more extreme means from the discussion, and asks "What about the sample with a mean of 106?" The responses shift to a more hesitant "Well, maybe not that one." This leads into a rich discussion regarding the expected variation in individual observations in a sample compared to expected variation in sample means. The instructor shows a graph of an individual student's sample distribution along with a collective distribution of all students' sample means to enhance their inquiry (Figure 2).

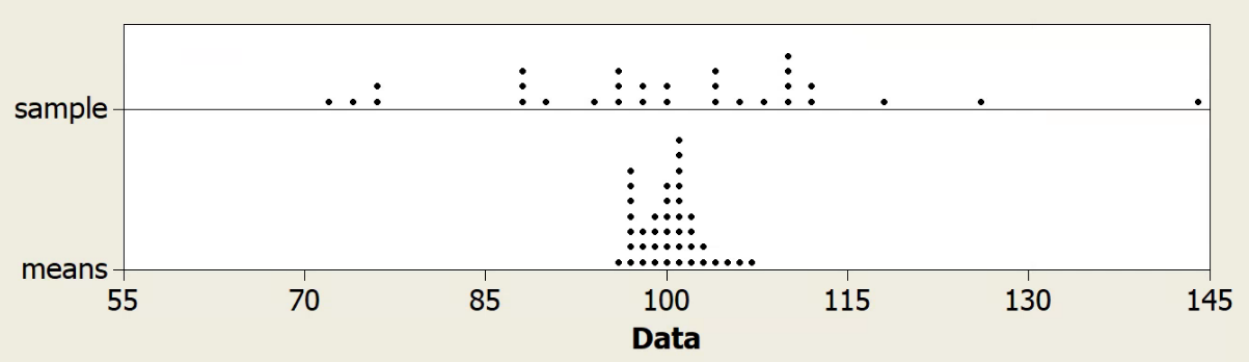

Figure 2. Dot Plots Showing a Sample Distribution and Sampling Distribution from First Discussion

With the visual aid, students understand that a mean of 106 could come from the same population, but the occurrence is unlikely. The instructor outlines a new, alternative population centered a little higher on the means graph and explains the increased likelihood of getting a sample mean of 106 with a different population. Students are then asked to guess what the mean and standard deviation of the population might be. While 100 is a common guess for the population mean due to the centering of the two dot plots, students struggle more with the standard deviation and are reminded of the probe in the discussion where students were asked to extend a distance of two standard deviation units from the mean. Once students have given this some thought, the descriptive statistics from both distributions are shared (Figure 3).

Descriptive Statistics: sample, means

Variable	Mean	SE Mean	StDev	Minimum	Median	Maximum	Range
sample	99.83	2.84	15.54	72.00	99.50	144.00	72.00
means	100.25	0.407	2.58	95.57	100.15	106.88	11.31

Figure 3. Descriptive Statistics from a Sample Distribution and Sampling Distribution from First Discussion

This becomes a powerful discussion as students are instructed to take their sample standard deviation and divide it by the square root of the sample size of 30 to see the concordance with the SE Mean (standard error) and to notice the proximity of this value to the standard deviation of the distribution of sample means. Students are notified that all of their samples were generated from the same population in the first discussion, one with a mean of 100 and a standard deviation of 15 which reflects accepted parameters for human IQs. A primary point of emphasis is the smaller variation in sample means, specifically the mathematical relationship between the population standard deviation and the sample size.

For the second discussion, the author created an Excel template using the Excel function =ROUND(NORM.INV(RAND(), mean, standard),0) to randomly generate three samples, each with 50 observations, from three populations. The first two populations are the same, a hypothetical population of IQs with a mean of 100 and a standard deviation of 15. The third population is different, a hypothetical population of IQs centered higher with a mean of 110 and a smaller standard deviation of 12. Students are now asked to compare their three samples in a discussion using descriptive statistics and graphical displays. Once again, as the instructor pushes the idea of statistical significance, the students are asked to discuss their thoughts on whether the samples all came from the same population. Based on the seminar discussion of the smaller spread of sampling distributions of means, many students recognized that the third group most likely came from a different population. Some students continued to struggle with this and reverted to thinking of the spread of an individual sample. Previous studies on sampling distributions have noted this same difficulty (Ozmen & Guven, 2019).

ERRORS IN HYPOTHESIS TESTING

The third and fourth units of the course cover probabilities with normal distributions and an introduction to hypothesis testing. By the end of the first half of the course, students have learned the steps to hypothesis testing, errors in hypothesis testing, and have conducted the three main types of t-tests for means (one-sample, independent samples, and paired samples). A unit on one-way ANOVA follows the unit on t-tests. For this discussion, the author created an Excel template using the Excel function =ROUND(NORM.INV(RAND(), mean, standard),0) to randomly generate three samples of data, each with 30 observations, two from the same population with mean of 156 and standard deviation of 10, and the third from a different population with a mean of 164 and a standard deviation of 10. The context is comparing productivity in three work settings to analyze the effect of music on productivity. The first group has no music, the second group has set music playing, and the third group has choice in the music playing. Each student is required to run a one-way ANOVA test with their three samples. The random generator typically produces a few student results that show no significant effect which provides a perfect opportunity to discuss probabilities of Type II errors. The three random datasets could also be manipulated to come from the same population providing the opportunity to discuss the probability of Type I errors and the relationship to a significance level, if appropriate.

CONCLUSION

Teachers have now overcome the hurdle of learning how to provide online instruction to their students. They can now focus their attention on the quality of their online delivery and instruction. Even if teachers are returning to the actual walls of their classroom, they have still taken away valuable lessons from their experience, and more than likely, hybrid learning will now hold a higher position than prior to the pandemic (Bozkurt & Sharma, 2020). The author believes that online discussion boards can be designed

to enhance students' conceptual understanding of some of the more difficult statistical ideas. Hopefully, instructors will be confident to use online tools to create dynamic and interactive lessons to complement their teaching.

REFERENCES

Allen, I., Seaman, J., Poulin, R., & Straut, T. (2016). Online report card: Tracking online education in the United States. Retrieved from http://onlinelearningsurvey.com/reports/onlinereportcard.pdf

American Statistical Association. (2016). *Guidelines for Assessment and Instruction in Statistics Education*. Retrieved from http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.

Bozkurt, A., & Sharma, R. C. (2020). Education in normal, new normal, and next normal: Observations from the past, insights from the present and projections for the future. *Asian Journal of Distance Education, 15*(2), i-x.
Retrieved from http://www.asianjde.com/ojs/index.php/AsianJDE/article/view/512

Chance, B., delMas, R. C., & Garfield, J. (2004). Reasoning about sampling distributions.
In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Katz, Y. J. & Yablon, Y. B. (2002). Who is afraid of university internet courses? *Educational Media International, 39*, 1, 69–73.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2021). Statistics: *Unlocking the power of data*. Hoboken, NJ: Wiley.

Ozmen, Z. M., & Guven, B. (2019). Evaluating students' conceptual and procedural understanding of sampling distributions. *International Journal of Mathematical Education in Science and Technology, 50*(1), 25-45. https://doi.org/10.1080/0020739X.2018.1467507

Xie, Xin and Siau, Keng, "Online Education During and After COVID-19 Pandemic" (2020). *AMCIS 2020 TREOs*. 93. Retrieved from https://aisel.aisnet.org/treos_amcis2020/93.