IMPACT OF A SIMULATION/RANDOMIZATION-BASED CURRICULUM ON STUDENT UNDERSTANDING OF P-VALUES AND CONFIDENCE INTERVALS

Beth Chance and Karen McGaughey
Department of Statistics, California Polytechnic State University, San Luis Obispo, CA, USA
bchance@calpoly.edu

This paper describes how changes in course sequencing and pedagogy have impacted students' understanding of p-values and confidence intervals in introductory algebra-based and calculus-based tertiary-level courses. Using assessment data (for example CAOS, common exam questions, and a transfer problem) across several institutions in various stages of implementation, this paper focuses on how the use of simulation and randomization-based inference has developed, what have been found to be the main student gains, and potential cautions with the approach.

INTRODUCTION

As discussed in Roy et al. (2014) we have been experimenting with a curriculum centered on simulation and randomization-based inference for learning statistical inference in introductory algebra-based tertiary level statistics courses. These ideas have been advocated in the statistics education community for a few years now (e.g., Cobb, 2007) and are now part of the Common Core State Standards for grades 9-12 in the United States. Preliminary evaluation results have shown very promising signs for improved student learning (e.g, Tintle et al, 2011; Tintle et al., 2012; Holcomb et al., 2010ab). Tintle et al.'s results (using the CAOS instrument, delMas et al., 2007) showed that students demonstrated gains in understanding of key inferential concepts while maintaining their understanding of other course goals such as descriptive statistics, as well as improved retention. The Holcomb et al. results illustrated some areas where students still struggled. We believed a more complete revitalization of the introductory course around these ideas, with an integrated textbook and laboratory assignments, could improve student gains. Several textbooks have recently appeared that do heavily integrate such an approach (Zieffler, 2013; Lock et al., 2013; Tintle et al., 2016). This past year, the Tintle textbook was class tested at several institutions and common assessment items were given at different times during the course. Results from the SATS attitude survey (Schau, 2003) are summarized by Swanson et al. (2014). In this paper, we will focus on a set of common assessment questions given typically as part of midterm and final exams which focus on student understanding of significance and confidence at different points in the course. In addition, we discuss results of a pre-test and post-test based on the CAOS instrument for items focusing on confidence and significance (see also Tintle et al., 2014, for further discussion of these tests). We will discuss the results of these items, how they coincide with our additional observations as instructors in the course, and how they reveal areas of strength and of improvement for the curriculum.

BACKGROUND

The logic of statistical inference can be difficult for many students. We feel this can be compounded if introduced through a focus on abstract mathematical models. We see two main advantages of the proposed curricula:

- Move the learning of inferential concepts to the very beginning of the course, allowing for an initial and then repeated exposure to the entire statistical investigation process as a whole.
- Using simulation and randomization-based inference methods not only to illustrate concepts of inference but as the primary mechanism through which students perform statistical inference.

So one of our key research questions is the level at which students are able to understand the logic of inference after the initial exposure and how their understanding develops with repeated exposure in an introductory course as outlined in Roy et al. (2014).

In the first two chapters of the Tintle text, students are introduced to inference for one proportion. This approach includes a simulated-based introduction to p-values (starting with coin tossing and moving to spinners to generate repeated observations from a random process specified by the null hypothesis to create a null distribution), followed by standardized statistics as an alternative measure of strength of evidence against a hypothesized value, followed by a theory-

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

based approach (one-sample *z*-tests). In Chapter 2, confidence intervals are introduced as the set of plausible values of the parameter that would not be rejected by a two-sided p-value at a certain significance level. Students "find" this interval using a simulation-based applet, recording the p-values for several hypothesized parameter values until they "zoom in" on the boundaries for values that are not rejected. This is then reported as the confidence interval for the parameter at the corresponding level. Students then compare this confidence interval to what they obtain using the point estimate plus/minus two standard deviations, where the standard deviation of the sample proportions is found from the null distribution simulation. This approach is then compared to the conventional "theory-based" one-sample *z*-interval, especially for other confidence levels.

In these first few chapters we are hoping to convey two overarching themes:

- Could the observed statistic have plausibly occurred by random chance alone?
- How far could the observed statistic plausibly be from the parameter value?

The remainder of the curriculum revisits these questions (as well as scope of inference) through different data settings: two proportions, one mean, two means, two variables.

As part of NSF Grants #9950476, #0321973, we collected data from class testers of the Tintle textbook during the Fall 2013 term. The institutions are a mix of public and private, including one community college and two high schools, with diverse student backgrounds. The class testers include members of the author teams, additional faculty members at the author team institutions, and instructors at other institutions who were new to the curriculum (most of whom attended a short workshop giving them an overview of the materials prior to the fall term). Class testers were asked to give students the SATS attitude survey and a concept inventory modelled after CAOS at the beginning and end of the term. Instructors were also provided with four common exam questions that could be used as part of midterm and final exams. To account for instructor to instructor variation, including class size, in summarizing student performance below, we weighted the class averages and standard deviation of the class averages by the class sizes. Results from a more complete hierarchical analysis, including incorporating the SATS results and additional instructor level variables, will be presented at the conference. We also hope to present more comparisons with students in other curriculums at that time as well.

UNIT 1 ASSESSMENT

Below we discuss results from the common multiple choice assessment item given as part of the first midterm by 14 instructors (sample sizes ranging from 4 to 69 students) at 12 different institutions (a total of 518 students). This question gave students a one proportion scenario (do adult city residents prefer NetFlix to watching movies at the theater) and asked them a series of multiple-choice questions, which included identifying the appropriate null and alternative hypotheses, possible versus plausible conclusions, and interpretation of p-value.

In questions 1 and 2, students were very successful in selecting the correct null hypothesis statement (mean = 96.8%, sd = 5.4%), and the correct alternative hypothesis (mean = 92.4%, sd = 6.6%). In the alternative, the most common error was choosing the incorrect direction. In this curriculum, students are introduced to the logic of inference by posing two competing explanations for an observed sample majority – (1) maybe there really is a majority in the larger population or process, and (2) maybe there isn't, and we observed the sample majority by random chance alone. Small p-values are then seen as evidence against the "by chance alone" explanation. Question 3 gives students a p-value of 0.012 and asks them to select the most "plausible (i.e., believable or reasonable) explanation:

- More than half of the adult results in the city prefer to watch the movie at home;
- There is *no* overall preference for movie-watching-at-home in the city, but by pure chance the sample happened to have an unusually high number of people choose to watch at home;
- (a) and (b) are equally plausible explanations.

Performance on this item is much more mixed. Response (a) was chosen by approximately 68% of students (sd = 17.5%), response (b) by approximately 6.5% (but as high as 15% in one class), and response (c) by just under 30% of students (sd = 16.9%). From this result, we would say roughly 35% of students still do not understand the distinction between *plausible* and *possible* explanations. Seeing similar results to this assessment in previous terms, we have wondered whether it is more of

a semantics question than a statistics question. However, this difference in terminology was emphasized more in the Fall course materials and some students still struggle with the distinction.

Question 4 asks students to select an appropriate interpretation of p-value.

- A sample proportion as large as or larger than this would rarely occur if the study had been conducted properly. (mean = 7.2%; sd = 5.0%)
- A sample proportion as large as or larger than this would rarely occur. (11.8%; 7.9%)
- A sample proportion as large as or larger than this would rarely occur if 50% of adults in the population prefer to watch the movie at home. (67.2%; 15.8%)
- A sample proportion as large as or larger than this would rarely occur if more than 50% of adults in the population prefer to watch the movie at home. (14.8%; 9.5%)

A majority of students chose the correct interpretation, with incorrect interpretations not assuming any basis behind the simulation and assuming the alternative hypothesis chosen similarly.

The last question asked students to comment on whether a confidence interval for the population proportion, based on the analysis presented so far (the p-value), would include the value 0.5. The intention was for students to see that 0.5 was not a plausible value for the parameter based on the p-value of 0.012 given previously. About 12.8% of students found 0.5 to be plausible, 43.2% (sd = 18.7%) found 0.5 to not be plausible, and 45.9% said they did not have enough information to decide (this reduces to 35% if one small school is removed). When given the option to indicate the missing information, most students claimed they needed to know the value of the sample proportion, often clearly describing how they would compute the interval, but failing to see the connection between the previously provided p-value and whether or not the confidence interval would include the value 0.5.

At this early point in the course, students appear to be developing some comfort with the overall inferential process (stating hypotheses, determining p-values, drawing conclusions, interpreting intervals) but are not yet seeing the connections between these components or the inductive nature of the reasoning process. Not surprisingly, some students are also struggling with notation and terminology (e.g., $\hat{p}$ vs. probability vs. p-value; plausible vs. possible).

UNIT 2 ASSESSMENTS

Next we summarize results from a common multiple choice question given by 12 of the class testers after the second unit (comparing two groups). Instructors varied on whether they gave the questions immediately after the second unit or at the end of the course, but most reported embedding the questions in a major exam (rather than as a quiz or review problem). Students are given a scenario comparing two proportions and are asked to interpret various conclusions which get at the issues of power and sample size (see Table 2). Question 1 tells students the difference between the two groups is found *not* to be statistically significant. In the past, we felt that students would be more likely to infer this gave "strong evidence that there was no difference," but only about 18.5% (sd = 13.5%) chose this option, with almost 70% (sd = 13.2%) selecting instead a correct response about lack of evidence of a difference.

Next students were asked how to interpret a small p-value. About 51% (sd = 12.6%) correctly identified that the "observed sample results" would be surprising if there really had been no difference (in context). Just over 25% (27.5%, sd = 10.2%) indicated that the small p-value corresponded to a small chance of no difference. The rest selected strong evidence of no difference (27.5%; 10.2%) and small probability of a difference (8.5%; 6.8%).

The third question told students that previous research indicated the difference should be significant, but this study found a large p-value, asking whether that implies that something went wrong. Over 90% (sd = 9.6%) of students correctly indicated that the sample size might have been too small to detect a difference.

When given two studies, one with a larger difference in the two sample proportions, about 86% (sd = 12.1%) of the students correctly indicated that that study would provide stronger evidence of a genuine difference, though over 11% selected "same strength of evidence." When given two studies with the same difference in the sample proportions, but different sample sizes, 79% (sd = 14.6%) correctly indicated that the study with larger sample sizes would yield stronger evidence, with almost 16% selecting "same strength of evidence."

Lastly students were given a similar "plausible vs. possible" question as the Unit 1 Assessment, for a statistically significant difference between two groups. Overall about 62% (sd = 12.4%) of students chose the alternative hypothesis as the more plausible explanation, with 24% finding them equally plausible, and 16% finding "random chance alone" more plausible.

Another question that we have been experimenting with describes a simulation where values of the response variable are randomly reassigned to males and females, as done repeatedly in the course to find p-values. Students are then asked about the purpose of this use of randomness. Some students focus on the scope of conclusion allowing causation (which is why random assignment was used in the original study) and some on generalizability. For the two most experienced instructors, 54% and 61% of students correctly identified wanting to simulate values under the null hypothesis as the purpose for this use of randomness.

Our interpretation of these results is that with the change from one proportion to two proportions, students successfully carry over some of the big ideas, and struggle to carry over others. They do fairly well in judging factors that affect strength of evidence, even with the lack of emphasis on algebraic formulas in the course. Some improvement from Unit 1 to Unit 2 is seen in understanding of possible versus plausible explanations of a study outcome. However, students still experience difficulty distinguishing more subtle interpretations in p-value. The language issue may be exaggerated a bit in the Unit 2 Assessment, as questions changed back and forth between assuming a large or small p-value for the same study.

A second post-Unit 2 item was given by 11 of the instructors to a total of 323 students. This question asked similar questions about the purpose and conclusions from a randomization test but this time comparing two sample means. Over 90% (sd = 34.1%) were able to recognize the purpose of a randomization test to assess the unusualness of a result if the null hypothesis was true (in context). Yet only 60% (sd = 24.6%) stated the null hypothesis as the assumption behind the simulation (vs. the alternative hypothesis). Then 75% (11.4%) were able to pick the correct p-value interpretation, including the assumption of no effect.

We find some inconsistencies in these results. Student appear to fairly quickly understand the notion of a test of significance helping us decide whether the observed result could have happened "by chance alone" (Unit 1) but it has also been our observation that students continue to struggle with understanding the role of "assuming the null hypothesis to be true" in estimating and interpreting the p-value. Students may also be beginning to detect patterns in the phrasing of these questions and correct responses.

A transfer question (not presented here) also has illustrated remaining student difficulty in designing their own simulation to assess a brand new type of inference question.

PRE/POST TEST

We also asked the class testers to administer a multiple choice exam to their students at the beginning of the term and at the end of the term. Across 25 instructors, we have almost 1000 responses on the pre-test and roughly 500 responses on the posttest. Many of these questions are from the CAOS instrument developed by the University of Minnesota (delMas, 2007), which allows comparison to normative results from 2006. Tintle et al. (2014) report more fully on these results. Below we highlight a few interesting trends and comparisons regarding confidence intervals and p-values. We again report weighted averages and standard deviations for correct responses (Table 1).

- Confidence interval interpretations: Students are given a set of interpretations of confidence intervals and asked whether they are valid or invalid. With no real improvement from pre to post, a slight majority correctly indicate a statement about "95% of the population is in the interval" is invalid, but fewer indicate that "95% confidence an observation is in the interval" is incorrect. Students also struggle with whether sample statistics should be captured in the interval, but with moderate improvement by posttest. Most students can recognize a correct interpretation. These results are in line with the CAOS normative data.
- Students show strong improvement in recognizing that "statistical significance" corresponds to small (rather than large) p-values (many sections at nearly 100% correct). However, they still struggle with more subtle wording of p-values. With some moderate gains from pre to post, a majority of students recognized a correct interpretation. However, about half find a probability

statement about the alternative valid, with no real improvement from the beginning to the end of the course. About one-third find a probability statement about the null valid. These are comparable to the CAOS normative data.

- Students perform poorly on a variation of the "hospital problem" with less than 40% recognizing more variation with larger samples leading to a higher probability of an extreme result, preferring the "equal-probable because both random samples" option (over 40% post).
- Students perform poorly on recognizing an appropriate simulation method to estimate a p-value (22% pre, 32% post) though slightly improved compared to CAOS 22.4%, with 50% choosing an option that all suggestions were accurate, including repeating the actual study.
- On this post-test, most found stronger evidence in a larger sample size (64%), a slight increase from the pretest (55%).

We also included a question to assess student intuition on the necessary sample size to achieve a margin-of-error of 3 percentage points for all 310 million U.S. residents. On the pretest, 10,000,000 is the most common response, followed by 1,000,000 and then 10,000, with only 9% selecting 1,000. On the post-test, responses were much more uniform 22% selecting 1,000, and 29% still selecting 10,000,000.

Table 1. Pre and post results for CAOS like items

	PRE Mean (sd) (%)	POST Mean (sd) (%)	CAOS Norm
Confidence Interval Interpretations			
95% of population is in interval (invalid)	57.0 (11.6)	60.0 (12.7)	65.4
95% confident an observation is in interval (invalid)	30.1 (8.5)	35.2 (13.7)	49.4
95% of sample averages are contained within the interval (invalid)	48.6 (9.1)	61.3 (12.0)	47.5
Correct interpretation of 95% confidence interval	70.1 (7.7)	80.0 (8.8)	75.9
Statistical Significance and p-values			
Small p-value required for statistical significance	43.5 (18.8)	83.8 (16.0)	68
Correct interpretation of p-value	50.0 (8.2)	63.0 (8.9)	57
Probability Ha statement true (invalid)	41.4 (10.4)	51.0 (11.3)	54.4
Probability Ho statement true (invalid)	54.3 (8.4)	65.9 (9.4)	60.1
More variation in smaller samples leads to a higher probability for an extreme result	31.8 (15.8)	38.7 (11.6)	33
Recognizing an appropriate simulation method to estimate a p-value	22.0 (8.8)	31.9 (10.1)	22.4
A larger sample size provides stronger evidence	54.8 (14.0)	64.6 (11.8)	
Necessary sample size to achieve 3% margin of error	8.8 (6.2)	31.9 (10.1)	

Though student performance is on par with CAOS results for most items, students still struggle with several concepts. As was evident in the Unit 1 and Unit 2 Assessments, the pre/post-test results indicate students have difficulty with the subtleties of confidence interval and p-value interpretations. However, a previously common student question of "do I want the p-value to be large or small" has virtually disappeared. Students are showing gains in understanding of the effects of sample size and variation on statistical significance, but after a full quarter/semester of instruction, they still struggle with understanding the purpose of simulation (generating hypothetical results *assuming the null hypothesis* is true in order to evaluate the observed statistic).

CONCLUSION

In this curriculum, even for instructors new to the approach, students appear to leave with an understanding that study results are evaluated based on the notion of how likely they are to occur by random chance alone. They also develop a strong understanding that small p-values indicate statistically significant outcomes, and a majority of students appear to gain some understanding of the roles of sample size and variability in statistical significance. It is important to keep some cautions in mind:

- Anecdotal evidence suggests that students appear to have more difficulty in identifying and interpreting the roles of parameters than they might in a traditional course.
- Evidence provided in the assessments detailed in this paper, suggests that students continue to develop/maintain misconceptions about the interpretation of p-values and confidence intervals. They also fail to realize the importance of the null hypothesis in the simulation process, or perhaps even the purpose of the simulation itself.

To address these concerns, we feel it is important to

- Begin the course with some discussion of models and the role of simulation in determining probabilities, and the use of probabilities in making decisions.
- Similarly, don't yet (but may change in a few years) presume students have good understanding of descriptive statistics, particularly variability.
- Don't underestimate the difficulty students may have in the transition from one 50/50 proportion to other scenarios, including the distinction between sampling and assignment.
- Emphasize the design of the simulation at each stage (what does each repetition represent, what are the assumptions behind the simulation, what evidence is provided by an observation that is different from the simulation results, how the variation in the statistics is the key).

We continue to refine these assessment questions and to ask instructors of other courses to use them as well, as a gauge of mastery with the concepts. However, use of these assessments for class discussion may be their most effective use.

REFERENCES

Cobb, G. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education, 1*(1), 1-15. http://escholarship.org/uc/item/6hb3k0nz.

delMas, R., Garfield, J., Ooms, A., and Chance, B., (2007). Assessing Students' Conceptual Understanding after a First Course in Statistics. *Statistics Education Research Journal, 6*(2), 28-58.

Holcomb, J., Chance, B. Rossman, A., & Cobb, G. (2010a). Assessing Student Learning About Statistical Inference. In A. Rossman & B. Chance (Eds.), *Proceedings of the 8th International Conference on Teaching Statistics*. Voorburg, The Netherlands: ISI.

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010b). Introducing Concepts of Statistical Inference via Randomization Tests. In A. Rossman & B. Chance (Eds.), *Proceedings of the 8th International Conference on Teaching Statistics*. Voorburg, The Netherlands: ISI.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the Power of Data*. Hoboken, NJ: John Wiley and Sons.

Roy, S., Rossman, A., & Chance, B. (2014). *Using Simulation/Randomization to Introduce P-Value in Week 1*. Paper to be presented at ICOTS-9.

Schau, C. (2003). *Survey of Attitudes Toward Statistics (SATS-36)*. http://evaluationandstatistics.com

Swanson, T., VanderStoep, J., & Tintle, N. (2014). *Student Attitudes Toward Statistics from a Randomization-Based Curriculum*. Paper to be presented at ICOTS-9.

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J (2016). *Introduction to Statistical Investigations*. Hoboken, NJ: John Wiley and Sons.

Tintle, N., VanderStoep, J., Holmes, V-L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1). http://www.amstat.org/publications/jse/v19n1/tintle.pdf

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V-L., & Swanson, T. (2012). Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum. *Statistics Education Research Journal, 11*(1). https://www.stat.auckland.ac.nz/~iase/serj/SERJ11%281%29_Tintle.pdf

Zieffler, A. (2013). *Statistical Thinking: A Simulation Approach to Modelling Uncertainty* (2nd edition). Minneapolis, MN: Catalyst Press.