INTRODUCING CONCEPTS OF STATISTICAL INFERENCE VIA RANDOMIZATION TESTS

John Holcomb[1], Beth Chance[2], Allan Rossman[2], Emily Tietjen[2] and George Cobb[3]
[1]Cleveland State University, United States of America
[2]California Polytechnic State University, United States of America
[3]Mount Holyoke College, United States of America
j.p.holcomb@csuohio.edu

For over a decade now, technology tools have been advocated to assist student understanding of statistical concepts. We have designed and used applets that simulate sampling and randomization tests as a means for introducing students to concepts of statistical inference. In this talk we present the results of our investigation on the impact of using applets for this purpose with tertiary students. The foci of our investigation include appreciating the reasoning process behind statistical significance, understanding what a p-value is, and recognizing factors that affect p-values. We present the results of small classroom experiments designed to help inform our curricular materials and manner of teaching.

INTRODUCTION

Advances in technology have dramatically affected the practice of statistics and, to a lesser extent, the teaching of statistics. Technology enables instructors to ask their students to analyze data using statistical packages and also to explore statistical concepts with interactive software. One common example is using simulation software to investigate the concept of a sampling distribution and to learn what factors, such as sample size, affect the sampling distribution of a statistic.

Despite technological advances, the *content* of introductory statistics courses has remained largely unchanged, often focusing on normal-based inference procedures. Granted, technology enables students to conduct a *t*-test without resorting to a *t*-table in order to determine a *p*-value, and technology also makes it feasible to check the normality condition of a *t*-test with appropriate graphical displays. Technology even allows students to conduct simulations that investigate the robustness of *t*-procedures against departures from normality in the underlying population.

But a *t*-test is still a *t*-test, so these pedagogical uses of technology do not represent a very substantial advance. Moreover, for many students, the logic of statistical inference is obscured by studying such things as normal distributions and *t*-tests as a prelude to the concept of *p*-value.

Cobb (2007) made this point by arguing that "our curriculum is needlessly complicated because we put the normal distribution, as an approximate sampling distribution for the mean, instead of putting the core logic of inference at the center." Cobb cites the "tyranny of the computable" as the reason for this focus on the normal distribution and *t*-tests. He argues that the logic of inference is closer to the surface with randomization tests, which technology now enables us to introduce to students.

For the past few years, we have developed classroom activities that use technology to lead students to develop an understanding of the concepts of statistical significance and *p*-values as they conduct binomial simulations and randomization tests. In the process we have identified many issues that create difficulties for student understanding, and we have also raised many questions about the most pedagogically effective way to implement this approach. For some of these issues we have conducted small-scale classroom experiments, and for others we have gathered data on how students work through the classroom activities. In this paper we highlight these issues and present some preliminary findings.

EXAMPLES OF CLASSROOM ACTIVITIES

In this section we describe some of our recent materials and goals by presenting overviews of two classroom activities that we have developed. These, and other modules for classroom activities related to inference, can be found online at http://statweb.calpoly.edu/csi/.

Example 1: Naughty or Nice? This activity is based on a study reported in *Nature*, in which researchers investigated whether infants take into account an individual's actions towards

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

others in evaluating that individual as appealing or aversive, perhaps laying for the foundation for social interaction (Hamlin, Wynn & Bloom, 2007). In one component of the study, 10-month-old infants were shown a "climber" character (a piece of wood with "google" eyes glued onto it) that could not make it up a hill in two tries. Then they were shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character ("helper") and one where the climber was pushed back down the hill by another character ("hinderer"). The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood from the video (the helper and the hinderer) and asked to pick one to play with. The researchers found that 14 of the 16 infants chose the helper over the hinderer. (The videos can be viewed at http://www.yale.edu/infantlab/socialevaluation/Helper-Hinderer.html).

In this activity we ask students to consider whether the experimental result provides convincing evidence that the infants have a genuine preference for the helper toy rather than the result occurring merely "by chance." We start by asking whether the observed result (14 of 16 choosing the helper) could *possibly* have occurred if there were really was no preference between the two toys, and then we ask *how likely* such an extreme result would be under that null model of no preference. We lead students to investigate this latter question by flipping a fair coin 16 times, representing the infants' choices for the helper or hinderer under the null model of no preference. Students combine their results and begin to develop a sense for how unusual it would be to obtain 14 or more heads in 16 independent tosses of a fair coin. Students then use an applet to simulate 16 tosses of a fair coin to visually witness the variability in the number of heads from set to set, and then generate a large number (say, 1000) of repetitions of 16 tosses each. They examine this distribution of the number of heads and then use the applet to determine the proportion of these repetitions that produced 14 or more heads. This turns out to be a very small proportion (the *p*-value is .0021), so we want students to conclude that the observed research result provides fairly strong evidence that the infants genuinely do have a preference for the helper toy, that is was not merely a coincidence that so many picked the helper toy. More importantly, we hope that this activity leads students to be able to explain the reasoning process behind this conclusion.

Example 2: Sleep Deprivation? This activity is based on an experiment that investigated whether harmful effects of sleep deprivation on visual learning linger for several days (Stickgold, James & Hobson, 2000). The 21 subjects were randomly assigned to one of two groups: one group was deprived of sleep on the night following training and pre-testing with a visual discrimination task, and the other group was permitted unrestricted sleep on that first night. Both groups were then allowed as much sleep as they wanted on the following two nights. All subjects were then re-tested on the third day. The response variable was the improvement in response time to a visual stimulus on a computer screen. The mean improvement in the unrestricted sleep group turned out to be 19.82 milliseconds, compared to 3.90 milliseconds in the sleep deprived group.

We ask students to simulate a randomization test in order to assess whether this difference between the groups is statistically significant. They explore the likeliness of obtaining such an extreme difference under the null model that there's no effect of sleep deprivation by randomly assigning the 21 improvement scores (written on 21 index cards) between the two groups and calculating the difference in the re-randomized group means. Students combine results with classmates and examine the resulting distribution of differences in group means to see how unlikely the observed difference (19.82 – 3.90 = 15.92 milliseconds) is under the null model that the sleep deprivation has no effect.

Students then turn to an applet which shows the improvement scores from the actual study moving off the original dotplots, mixing together, and then being randomly reassigned to the two groups, color coded by the initial group membership and displaying the new difference in group means. Students then repeat this re-randomization process a large number of times. (See Figure 1.) The *p*-value turns out to be quite small (≈ .007), and we expect students to explain that it would be very unusual for random assignment alone to produce a difference between the groups as large as the actual experiment found, if there were no effect of sleep deprivation. Based on this reasoning, students conclude that the experiment therefore provides strong evidence that sleep deprivation is harmful even three days later.

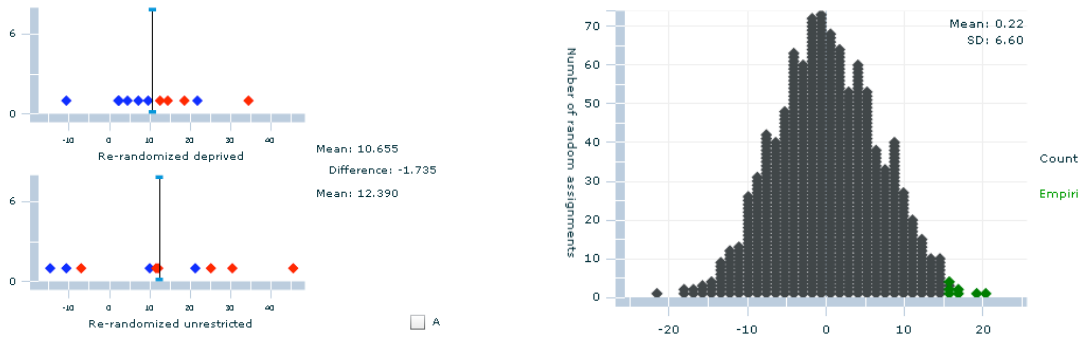

Figure 1. Screen Shot of Randomization Test for Quantitative Response Applet

FEATURES OF CLASSROOM ACTIVITIES

The two classroom activities described above share many features that we consider important for developing students' understanding of statistical significance:

- The activities are based on data from genuine research studies. These studies are of general interest, hoping to appeal to diverse groups of students and motivate the analysis.
- Both activities ask students to use the technology as a convenient tool to investigate the same core question of inference: How often would such an extreme result occur by chance alone?
- The general strategy for tackling this question is the same in the activities: Simulate the random process many times under the null model, and then see how unusual the observed result is.
- Students start with a tactile (literally "hands-on") simulation and then they proceed to use technology for simulating a large number of repetitions more efficiently.

We contend that this approach to introducing students to statistical inference has many advantages over a more conventional normal-based approach, including:

- These activities require virtually no prerequisite knowledge, so students can engage in them from the very beginning of the course. With such an important and difficult concept as statistical significance, we argue that an early experience, with frequent follow-up, allows a similar reasoning process to be reinforced over and over in new settings throughout the course. Thus, students can deepen their understanding with each activity.
- Descriptive statistic and inferential statistics no longer have to be separated in the course. In this way, even in the beginning of the course, students see the entire investigative process.
- The reasoning process of statistical inference is more transparent with randomization tests than with parametric inference procedures, so we expect that students can better understand concepts and reasoning of statistical inference with this introduction.
- Randomization tests make clear the role of randomness (either random assignment to treatment groups or random sampling from populations or processes) in the data collection process. We hope that this helps students to realize how important randomness is in collecting data, and also to recognize how the scope of conclusions that can be drawn from a study depend critically on how the data were collected.
- These randomization-based procedures take advantage of modern computing, in a much more fundamental manner than simply treating a statistics package as a glorified *t*-table with infinitely many rows and columns. Again, students can very quickly investigate additional questions, like the role of sample size and testing multiple plausible values for the population parameter/size of the treatment effect.
- With randomization tests, it is also quite straightforward to change the statistic of interest, for example to compare differences in medians rather than differences in means or the relative risk rather than the difference in proportions. It is also fairly straightforward to generalize to more complicated situations, for example to comparing a quantitative response variable across three groups rather than simply two groups.

CURRICULUM DESIGN ISSUES

Developing these activities has prompted us to ask many questions about curriculum design that we would like to see addressed by pedagogical research. These include:

1. Should the first activity that students encounter focus on inference for a single proportion, as in the "Naughty or Nice" example, or on a comparison of two groups, as in the "Sleep Deprivation" example? A clear advantage of starting with one proportion is that the situation is easier both to understand and to simulate. Tossing a coin 16 times to represent the toy choices of 16 infants is a simpler process to understand and implement than dealing out 21 index cards with numbers written on them , representing the random assignment of subjects with fixed response variable outcomes into treatment groups. But an advantage of comparing two groups is that such comparisons are an important and recurring theme in statistics, and such a scenario also allows for a greater variety of scientific studies to analyze. A third option, to start with comparing two groups on a categorical response, in other words with a 2×2 table, is also viable. (See the "Dolphin Therapy" activity with our materials.)
2. Should the first example that students encounter be one where the result is statistical significant, as with both examples presented above, or one where the result is not significant at all? An advantage of starting with a significant example is that students may find it easier to judge when an observed result is very unlikely to occur, as opposed to making a judgment about whether it is not unlikely to occur. A disadvantage to starting with insignificant results is this may reinforce students' natural inclinations to regard a non-small *p*-value as evidence in support of the null model, rather than a lack of evidence against the null model, or even to consider the simulation by itself as evidence in favor of the null hypothesis and not considering the observed result at all.
3. In the case of a simulation involving a single proportion, as with the "Naughty or Nice" example, how should the tactile simulation be conducted? For example, should each of 16 students flip a coin once, which has the advantage of representing each person (infant) in the study with one person (student) in the classroom? Or should each student flip a coin 16 times, which might make it harder for the students to recognize that the 16 flips represent 16 different infants? A third option is for each student to flip 16 different coins, each representing a different infant, but this plan would pose logistical problems and require a lot of coins.
4. In the case of a randomization test for a 2×2 table, what statistic should the students calculate in their simulations? Because a 2×2 table with fixed margins has just one degree of freedom, one possibility is to focus on one cell of the table, for instance the number of successes in the treatment group. An advantage of this choice is its simplicity; a disadvantage is that many students won't realize why one cell is sufficient to represent the table, and this approach also loses the idea of comparison. An alternative is to use the difference in proportions of success between the two groups. This has the advantage of emphasizing the comparison and also setting the stage for taking the difference in group means or medians with quantitative data. It also has a conceptual simplicity in that the center of the distribution is zero (under the null model of no treatment effect) and the variability decreases with sample size increases. A disadvantage that we have noticed is that some students get bogged down in calculating the two proportions and the difference between them, and some make mistakes in the calculation, thus distracting their attention from the inferential reasoning process we are focusing on.
5. We have suggested beginning each simulation with a tactile version before turning to technology, but does the tactile aspect really add value to the students' learning experience? Our thinking is that students are in a better position to understand what the technology simulation is doing if they have first performed a tactile simulation themselves. We have also tried to develop applets that mirror the hands-on simulation as closely as possible so the technology is not simply a "black box" to them. But would the activity be just as effective, and would students understand equally well, if they skipped the tactile part and began the simulation analysis with technology.
6. How much of the work should the technology do automatically? For example, to calculate an approximate *p*-value from a simulation, should the applet provide this value automatically, or should the student user need to specify the direction in which to count, or should the student user also need to indicate the value to count above or below?

7. Should the type of randomness used in the simulation always reflect the role of randomness used in the actual data collection process? For example, should we use randomization-based methods when analyzing results of randomized experiments, as with the "Sleep Deprivation" examples, but then switch to resampling-based methods (bootstrapping) when analyzing data that arise from independent random samples from two populations? Or should we create finite or artificial populations to sample from? An advantage of using the different kinds of randomness would be to emphasize that random sampling and random assignment are very different processes with different goals that support different analyses and scopes of conclusions. A disadvantage is that this adds an additional (and ultimately unnecessary) layer of complexity for students to think through when conducting simulation analyses.

CLASSROOM EXPERIMENTS

We have conducted some small-scale classroom experiments to address some of these questions. The question for which we have the most data is in regard to #2 above: whether the first example should have a significant or non-significant result. Our experiment involved four sections of an introductory course at Cal Poly. We told approximately half the students that 9 of the 16 infants chose the helper toy (referred to herewith as the "non-significant result group"), and the other students were told the true experimental result that 14 of 16 chose the helper ("significant result group"). Students were given the activity and told to work in pairs. Two instructors were involved, with one instructor randomizing across sections and the other randomizing by individuals. After completion of the activity, students in the non-significant result group were given the following question (with correct answer (d)):

> *When I conducted the simulation using 1,000,000 repetitions, I obtained a proportion in part (l) of .402. Based on this result, which assumes the null model of genuine preference, the actual obtained by the researchers (9 of 16 choosing the helper) is*
> *a) impossible* *b) very surprising*
> *c) somewhat surprising* *d) not at all surprising*

The analogous question for the significant result group reported the empirical *p*-value as .002, were told "14 of 16 choosing the helper," and the correct answer was "very surprising." The results were that 60.6% (n=71) students in the non-significant result group (60.6%) answered correctly, while 77.5% (n=71) in the significant result answered correctly (two-sided p-value ≈ .030). Our interpretation of this result is that students find it easier to spot a surprising outcome than a non-surprising one.

The second question asked of these students was:

> *Fill in the blanks in the following sentence to interpret this proportion from part (l).*
> *This proportion says that in about __(1)__% of _____(2)_____, the researchers would get ____(3)____ who choose the helper toy, assuming that _____(4)_____.*

For (1) above, the "non-significant result group" did significantly better than the "significant result" group. The correct answer for the non-significant result group was 40.2% while the correct answer for the significant result group was 0.2%. The difference may largely be in misunderstanding how to convert .002 to a percent. There was very little difference in the correct response rate for (2) (we were looking for the number of repetitions, 1 million) with just over 50% for both groups answering correctly. For (3), where we were looking for answer of 9 or more for the non-significant result and 14 or more for the significant result group, the significant result group did better (54.2% vs. 38.9%, two-sided $p-\text{value} = .066$). In regard to (4), the difference between groups was not statistically significant with approximately 76% answering correctly that there is no preference. Although students seem to equally understand the null model, they did differ slightly in realizing what the simulation told them.

Interestingly, there was not a significant difference between the groups on a third question that asked for an overall interpretation:

> *Based on your answer to (1) and (2), which of the following would you consider the most appropriate conclusion from this study? (choose one)*

(a) These 16 infants have no genuine preference and therefore there's no reason to doubt that the researchers' result is different from .5 just by random chance. (b) The researchers' results would be very surprising if there was no genuine preference for the helper and therefore I believe there is a preference. (c) There is a large chance that there is a genuine preference for the helper.

Approximately 77% answered the correct answer: (a) for the "non-significant result" group and (b) for the "significant result" group, not demonstrating the common misconception that the p-value corresponds to the probability of the null model being true.

A second classroom experiment was conducted to investigate the value of using classroom time to engage students with a tactile simulation (#5 above). Here we randomly assigned 43 students to two treatment groups, where the class topic was investigating the sampling distribution of a single proportion. In the tactile group, the instructor and 20 students worked through materials that included giving each student a sample of 25 actual Reese's Pieces candies to determine the sample proportion of orange candies. The students created a dotplot of their sample proportions, and then they turned to an applet to simulate many samples of size 25. The second group of 23 students did not perform the tactile simulation but instead immediately moved to simulating the colors of 25 candies using the applet, with a teaching assistant available for answering questions.

After completion of this activity, students in both groups were given a quiz that consisted of five questions with a new situation that involved a single sample proportion (the questions are available at http://statweb.calpoly.edu/bchance/csi/advisors.html). An independent and blinded statistics instructor scored the quizzes and did not find a statistically significant difference in student performance on this quiz. An interesting aspect of this study was that the students in both the tactile group and the other group appeared to finish the activity in about the same amount of time, suggesting that the tactile aspect does not take more time and does not hinder learning. We are now engaged in further analysis of these data and are considering repeating a similar experiment.

In a follow-up questionnaire to an activity for a case of analyzing data from data in a 2×2 table, we asked students: "Do you think that the hands-on simulation with the cards added to your understanding of the randomization process, in addition to the computer applet?" We characterized 50% (of 46 respondents) as saying they found the cards helpful. The responses fell into the following categories: the cards helped them understand what the computer was doing, involved them in the process, are better for visual learners, or the student said they learn better by doing.

CONCLUSION

We have been collecting and analyzing additional data to help inform the development of classroom materials for a randomization-centered curriculum. Of course such research requires developing assessments to capture student understanding of desired concepts. In tandem with curriculum development, our developing assessment instruments are described in Holcomb et al. (2010). With the development of additional curriculum modules and the refinement of others, we hope to gain more insight into the most common stumbling blocks displayed by students as they develop an understanding of statistical significance. The designs of the simulations appear effective, but students still struggle using the simulation results to draw appropriate conclusions.

REFERENCES

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

Holcomb, J., Chance, B., Rossman, A., & Cobb, G. (2010). Assessing student learning about statistical inference, *Proceedings of the 8th International Conference on Teaching Statistics.*

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*, 557-559

Stickgold, R., James, L., & Hobson, J. A. (2000). Visual discrimination learning requires post-training sleep. *Nature Neuroscience*, *2*, 1237-1238.