Exploring Student Understanding of Significance in Randomization-Based Courses

Holcomb, John
Cleveland State University, Department of Mathematics
2121 Euclid Avenue, RT 1515
Cleveland, OH, USA
j.p.holcomb@csuohio.edu

Rossman, Allan
Cal Poly San Luis Obispo, Department of Statistics
Building 25, Room 107D
San Luis Obispo, CA, USA
arossman@calpoly.edu

Chance, Beth
Cal Poly San Luis Obispo, Department of Statistics
Building 25, Room 107D
San Luis Obispo, CA, USA
bchance@calpoly.edu

Introduction

Cobb (2007) argued for a new curriculum for the introductory statistics course that is "centered not on the normal distribution, but on the logic of inference" (p. 4). He goes on to say, "We would introduce inference by way of the permutation test for randomized experiments." Furthermore, Cobb presents twelve arguments for why one should teach introductory statistics in this manner. Five of the most compelling reasons to us include:

- the model matches the production process, and so it is easy to emphasize the connection between data production and inference;
- the model is simple and easily grasped;
- the distribution is easy to obtain by physical simulation for simple situations;
- The entire paradigm generalizes easily to other designs (e.g., block designs), other test statistics, and other data structures (e.g., Fisher's exact test);
- we should do it because Fisher (1936) told us to.

Thus, we were inspired to develop a curriculum that would put the "core logic of inference" at its very center. We designed activities that would allow students to explore, from start to finish, the entire inference process starting on the very first day of an introductory statistics course. With the excitement of the newly enlightened, we developed a spiral curriculum that then builds on these key ideas of statistical inference and them at deeper levels repeatedly throughout the course. In this curriculum, randomization-based tests, rather than standard, normal-based parametric tests, serve as the entry point and provide the focus for students' developing their understanding of inference concepts.

In this paper, we briefly describe our curriculum and then describe our research on how well students are grasping the desired concepts. We also describe some of our research questions on how best to present this curriculum and results of some small-scale experiments that we conducted to inform how better to implement such a course. In short, although we believe such an approach is leading to deeper and longer-lasting understanding of statistical significance for most students, our optimistic hopes that this curriculum would quickly yield substantial understanding has been tempered with the bitter reality that for some students the ideas of statistical significance and inference can remain quite challenging.

Classroom Activities

We have developed a sequence of computer-classroom activities that center on randomization-based approaches to inferences for the following settings:

- a single proportion;
- a two-by-two table;
- comparing two groups with a quantitative response;
- quantitative data from a paired design;
- the slope of a regression line.

We have designed these activities to use data from genuine research studies or from classroom studies. These studies are of general interest, as we are hoping to appeal to diverse groups of students and attempting to motive the analysis as well as the complete statistical process. Our general strategy in all the activities is to simulate the random process many times under the null model, and then see how unusual the observed result is. For the simulation, we first use a tactile simulation (often involving coins or playing cards) and then move to a web-based applet. Our approach with the applet is for students to investigate often, "how often would such an extreme result occur by chance alone?"

Herein we describe 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 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 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 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 reaction 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 likelihood of obtaining such an extreme difference under the null model that there is 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 at least 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 genuinely detrimental even three days later.

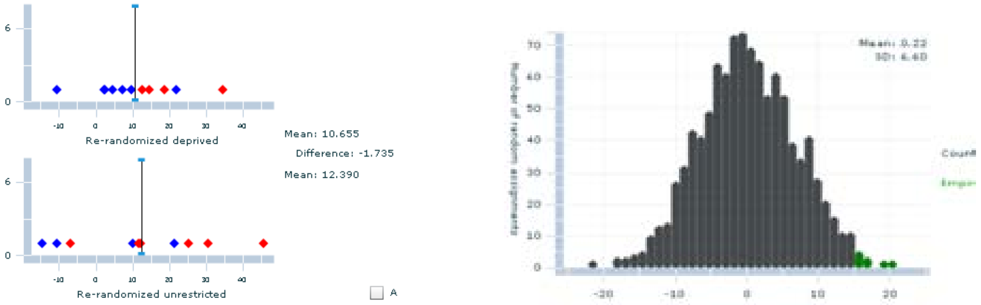


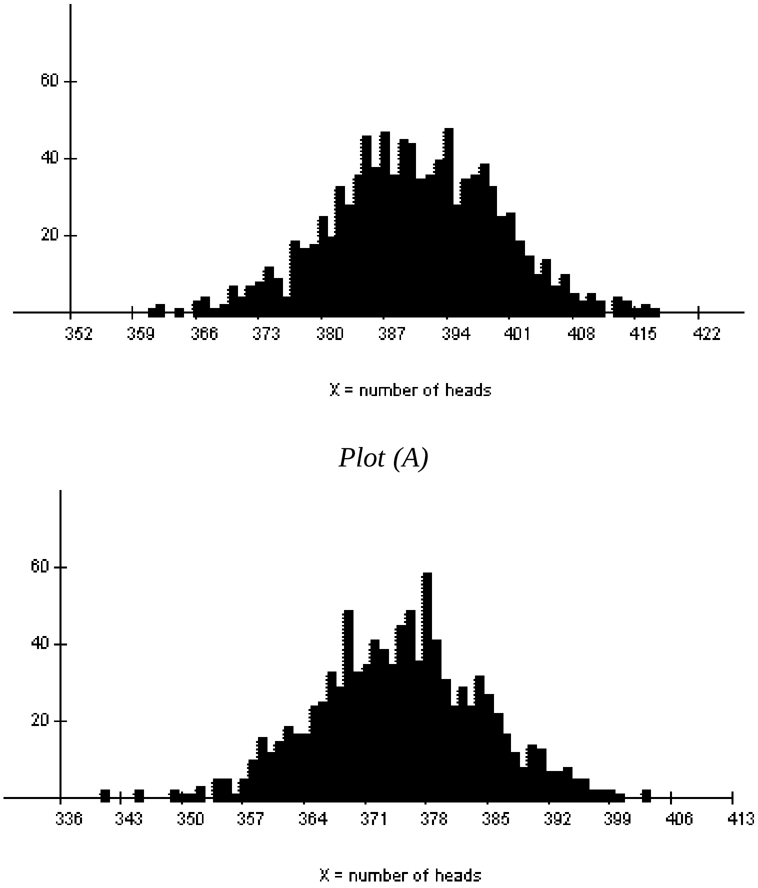
Figure 1: Screen Shot of Simulation Results of Randomization Test

Evaluating Understanding

We had originally believed that such an approach would lead almost every student to divine understanding. Our excitement over the ability to teach the entire inference process on the first day made us overly optimistic on what students would actually grasp after just the first example. For example, an exam question after Example 1 described above for a single proportion yielded disappointing results. First, the question:

> In a recent Gallup survey of 500 randomly selected US adult Republicans, 390 said they believe their congressional representative should vote to repeal the Healthcare Law. Suppose we wish to determine if significantly more than three-quarters (75%) of US adult Republicans favor repeal. The

coin tossing simulation applet was used to generate the following two dotplots (A) and (B). Which, if either, of the two plots (A) and (B) was created using the correct procedure? Explain how you know.



Plot (A)

Plot (B)

Figure 2: Simulated Distribution for $\hat{p}$ *centered at 390 and 375.*

We found that 35% answered Plot (B) correctly with 29% choosing plot (A) which is centered on the sample result, 23% choosing neither (many indicated they wanted the center to be .5*500=250), and 13% gave other answers.

Perhaps the spiraling nature of revisiting the ideas of significance throughout the course does have a positive impact on instruction. We have placed some questions from the Comprehensive Assessment of Outcomes in Statistics (CAOS) assessment tool developed by delMas et al. (2007) on our final examination. We compare our results (Cal Poly students, denoted CP) to national normative data collected by the CAOS team and results from a team of researchers at Hope College utilizing a similar curriculum based on randomization methods (Tintel et al. (2011)). Below we give results for a traditional introductory curriculum compared to a randomization-based curriculum for these three groups.

1. Statistically significant results correspond to small p-values
   - Traditional (National/Hope/CP): 69/86/41%
   - Randomization (Hope/CP): 95%/95%

2. Recognize valid p-value interpretation
   - Traditional (National/Hope/CP): 57/41/74%
   - Randomization (Hope/CP): 60/72%
3. p-value as probability of Ho - Invalid
   - Traditional (National/Hope/CP): 59/69/68%
   - Randomization (Hope/CP): 80%/89%
4. p-value as probability of Ha – Invalid
   - Traditional (National/Hope/CP): 54/48/72%
   - Randomization (Hope/CP): 45/67%
5. Recognize a simulation approach to evaluate significance (simulate with no preference vs. repeating the experiment)
   - Traditional (National/Hope/CP): 20/20/30%
   - Randomization (Hope/CP): 32%/40%

These results show some promise that students are developing a better understanding of statistical significance and p-value in a randomization-based curriculum than with a standard curriculum.

Multiple Choice Questions

We developed the following seven multiple choice questions to serve either as an assessment immediately following the teaching of a module on inference or as a summative assessment that could be integrated easily into a final examination. Here we also provide some summary statistics of student performance in an introductory statistics class of students at Cal Poly; the correct answer is indicated in bold. We believe that these results indicate that we are on the right track in developing discriminating questions.

Questions 1-7 concern the following scenario:

You want to investigate a claim that women are more likely than men to dream in color. You take a random sample of men and a random sample of women (in your community) and ask whether they dream in color.

Note: A "statistically significant" difference provides convincing evidence (e.g., small p-value) of a difference between men and women – *This note is optional to include.*

1) If the difference in the proportions (who dream in color) between the two groups turns out <u>not</u> to be statistically significant, which of the following is the best conclusion to draw?

26% a) You have found strong evidence that there is no difference between the groups.

62% b) You have not found enough evidence to conclude that there is a difference between the groups.

12% c) Because the result is not significant, the study does not support any conclusion.

2) If the difference in the proportions (who dream in color) between the two groups <u>does</u> turn out to be statistically significant, which of the following is a valid conclusion?

12% a) It would <u>not</u> be surprising to obtain the observed sample results if there <u>is really no</u> difference between men and women.

82% b) It would be very surprising to obtain the observed sample results if there <u>is really no</u>

difference between men and women.

6% c) It would be very surprising to obtain the observed sample results if there is really a difference between men and women.

3) Suppose that the difference between the sample groups turns out not to be significant, even though your review of the research suggested that there really is a difference between men and women. Which conclusion is most reasonable?

6% a) Something went wrong with the analysis.

6% b) There must not be a difference after all.

88% c) The sample size might have been too small.

4) If the difference in the proportions (who dream in color) between the two groups does turn out to be statistically significant, which of the following is a possible explanation for this result?

8% a) Men and women do not differ on this issue but there is a small chance that random sampling alone led to the difference we observed between the two groups.

30% b) Men and women differ on this issue.

62% c) Either (a) or (b) are possible explanations for this result.

5) Reconsider the previous question. Now think about not possible explanations but *plausible* explanations. Which is the more plausible explanation for the result?

28% a) Men and women do not differ on this issue but there is a small chance that random sampling alone led to the difference we observed between the two groups.

36% b) Men and women differ on this issue.

36% c) They are equally plausible explanations.

6) Suppose that two different studies are conducted on this issue. Study A finds that 40 of 100 women sampled dream in color, compared to 20 of 100 men. Study B finds that 35 of 100 women dream in color, compared to 25 of 100 men. Which study provides stronger evidence that there is a difference between men and women on this issue?

78% a) Study A

2% b) Study B

20% c) The strength of evidence would be similar for these two studies

7) Suppose that two more studies are conducted on this issue. Both studies find that 30% of women sampled dream in color, compared to 20% of men. But Study C consists of 100 people of each sex, while Study D consists of 40 people of each gender. Which study provides stronger evidence that there is a difference between men and women on this issue?

82% a) Study C

8% b) Study D

10% c) The strength of evidence would be similar for these two studies

For further details regarding our approach to assessment see Holcomb, Chance, Rossman and Cobb (2010).

Classroom Experiments

We also made a conscious decision to inform our curriculum development with data that we gathered on student understanding. For example, we hypothesized that it is easier for students to grasp the meaning of *p*-value if the result from the initial case study that they analyze is statistically significant. Thus we performed a classroom experiment that involved four sections of an introductory course at Cal Poly. With regard to Example 1 described above, we told approximately half the students that 9 of the 16 infants in the study chose the helper toy (referred to herewith as the "non-significant result group"), and the other students were told the actual 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 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 identify 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, as opposed to proceeding directly to a computer-based simulation. 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 drawing many random samples of size 25. The second group of 23 students did not perform the tactile simulation but instead immediately moved to simulating random samples 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 conducting a similar experiment.

Also with regard to the issue of the usefulness of tactile simulations, 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.

For further discussion of other research questions for classroom experiments, see Holcomb, Chance, Rossman, Tietjen, and Cobb (2010).

Discussion

We agree with Cobb that introductory statistics students can benefit from studying concepts if statistical inference is introduced with randomization-based methods rather than traditional, normal-based, parameteric techniques. With the development of additional assessment items 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 some students still struggle using the simulation results to draw appropriate conclusions. As instructors, we have noticed that some students seem to grasp the idea that "this result probably did not happen by chance alone," but moving beyond that to a more technical terminology that extends to multiple scenarios is more difficult than expected. Anecdotal evidence in more recent teachings of this curriculum suggest that having the instructor fully model how to explain the reasoning and draw a conclusion from the may have a positive impact. For example, highlighting the phrase "the observed study result" may help reinforce the idea that we look at the simulation plot to determine where the sample result lies and base inference on the tail of the null distribution beyond that one study point, rather than using the simulation results in isolation to make a conclusion (it's centered at .5 so we believe the parameter equals .5)

We also have observed the following:

- students do struggle awhile with the null result vs. the observed result. This may be helped by using classroom data where they are actively involved in observing the result and perhaps even more invested in evaluating the result.
- they struggle with our distinction between possible and "plausible" so the instructor may want to actively define that for them.
- isolated activities added to an existing curriculum, e.g., through separate labs, appears not to be sufficient, and it's much better for the reasoning/simulations to be integrated throughout the course.

We want to investigate the impact of the above changes, and our curriculum materials are constantly being refined and updated. We realize that this curriculum does not lead to "automatic" understanding of the statistical inference process, but we believe this curriculum does make strides toward the vision laid out in Cobb (2007).

REFERENCES

Cobb, George W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, 1(1),www.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.

Fisher, R.A. (1936), The coefficient of racial likeness and the future of craniometry, *Journal of the Royal Anthropological Institute of Great Britain and Ireland*, vol. 66, pp. 57-63.

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

Holcomb, J., Chance, B, Rossman, A., Cobb, G. (2010). Assessing student learning about statistical inference, 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 [© 2010 ISI/IASE].

Holcomb, J., Chance, B, Rossman, A., Tietjen, E., Cobb, G. (2010). Introducing Concepts of Statistical Inference via Randomization Tests, 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 [© 2010 ISI/IASE].

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

Tintle, N., VanderSteop, J., Holmes, V., Quisenberry, B., Swanson, T., (2011) Development and assessment of a preliminary randomization-based introductory statistics curriculum, *Journal of Statistics Education,* 19(1) [on-line].

ABSTRACT

Statistical significance and p-values can be a particularly challenging topic for introductory statistics students. Here we present the results of our research regarding implementing a randomization-based curriculum that uses Applets as the main randomization tool. We present the results of small classroom experiments designed to help inform our curricular materials and manner of teaching in a variety of classroom environments. We also present research results that indicate where students have the greatest difficulties in understanding significance.

Keyword 1: p-value **Keyword 2:** significance **Keyword 3:** randomization **Keyword 4:** statistical education