ASSESSING STUDENT LEARNING ABOUT STATISTICAL INFERENCE

John Holcomb[1], Beth Chance[2], Allan Rossman[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
bchance@calpoly.edu

Statistical significance and p-values can be a particularly challenging topic for introductory statistics students. In an effort to assess curricular changes aimed at deepening student understanding of significance, we have developed assessment strategies to diagnose students' conceptualization of p-value and their ability to communicate their understanding. We will present our approaches and discuss student performance after participating in randomization-based modules introducing the concept of significance.

INTRODUCTION

Statistics educators have developed a remarkable consensus about the content of the introductory, algebra-based (that is, non-calculus-based) statistics curriculum. In his keynote address at the first United States Conference on Teaching Statistics, Cobb (2007) argued that while this curriculum consensus has many desirable features, it has had the unintended and negative consequence of restricting curricular experimentation in statistics. For example, modern computing provides the opportunity for statisticians to carry out statistical inference through simulation of the randomness inherent in the design of the study. This capability is certainly available to introductory students as well, and we believe that this approach may lead to a deeper understanding of statistical concepts.

Toward this goal, we have endeavored to develop a dramatically different curriculum for the algebra-based, introductory statistics course. The two primary goals of this curriculum are to help students (1) understand the "big picture," connections between ideas, and overall process of statistical investigations; and (2) understand fundamental concepts of statistical significance at a deeper than superficial level.

The two most distinctive ways in which the proposed curriculum aims to achieve these goals are through (1) a "spiral" curriculum that presents key ideas of statistical inference early in the course, and then revisits them at deeper levels repeatedly throughout the course; and (2) use of randomization-based tests, rather than standard, normal-based parametric tests, as the entry point and focus for students' developing their understanding of inference concepts.

Classroom materials available for use in this curriculum may be viewed at http://statweb.calpoly.edu/csi/. As we developed these classroom materials, we also examined assessment resources to utilize, adapt, and generate new tools to determine if the new curriculum was successful in generating student understanding of statistical significance. In this paper, we describe our current slate of quantitative and qualitative assessment instruments that we believe are helpful in reaching our goals of helping students to understand fundamental concepts of statistical inference.

PREVIOUS WORK

In our investigations of published work in assessing student understanding of statistical significance, we first looked to the Comprehensive Assessment of Outcomes in Statistics (CAOS) developed by delMas, Garfield, Ooms & Chance (2007). This instrument is a collection of forced choice questions to assess student understanding of conceptual issues after students take a college-level introductory statistics course. Because the goal of this instrument is to assess issues from the entire spectrum of concepts in the first course, there are only a few items that specifically address the issue of statistical significance. These same authors, however, also developed "Topic Scales," which consist of 7–15 multiple choice questions on individual topics within the first year curriculum. One topic scale is designed to assess student understanding of significance testing. These resources are available at the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) web page at https://app.gen.umn.edu/artist/scales.html.

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]

Lane-Getaz (2007, 2008) has developed the Reasoning about P-values and Statistical Significance (RPASS) scale. This instrument explicitly assesses student understanding of statistical significance issues. The latest version of the scale consists of items developed previously for RPASS in combination with items from the ARTIST topic scale on significance tests.

In conjunction with these assessments, we sought additional items that focused on the process of statistical inference, and the use of simulation tools in particular, the ability to explain the rationale behind statistical significance and to articulate the components of a p-value, as well as student ability to transfer their understanding to new scenarios.

ASSESSMENT INSTRUMENTS UNDER DEVELOPMENT

Our main efforts have consisted of developing assessment instruments with five different themes. One focus is developing instruments that determine whether students understand the basic components of the randomization activity that they just completed. A second theme relates to our desire for students to be able to produce or read a summary of a statistical study and give a thorough (and correct) interpretation of what a reported p-value means. A third effort involves determining whether students are able to apply what they have learned in an activity to a new statistical study. Our fourth effort involves developing a small set of multiple choice items that assess student understanding of significance, with a specific effort to keep these questions as jargon free as possible. Our developed multiple choice questions are slightly different in scale and scope from the CAOS and RPASS instruments in that we wanted them to be easily embeddable into a pre-test, unit assessment, or final examination. Our fifth effort involves determining whether students can apply the ideas of the randomization approach to a scenario that requires a different randomization approach or a different statistic than any to which they have been exposed thus far.

Understanding Components of Activities

The classroom materials developed to help students understand statistical significance require students to engage in simulation activities. One of our tenets has been to have students engage in a tactile simulation (such as flipping a coin or shuffling cards), and then turn to a technology tool such as an applet to perform a larger-scale simulation. One of our first goals was to determine whether students who have worked through a simulation activity understand what they have done and how the components of the activity relate to the research study being analyzed. We also used our assessment of this facet to determine whether structural or stylistic changes needed to be made to the classroom materials.

The questions in Figure 1 followed an activity in which students simulated a randomization test for data presented in a 2×2 table. The simulation activity involved shuffling cards, dealing them out to two groups, and counting the number of "successes" in group A. Students repeated this process numerous times, pooling across student results, and then using a java applet, determined the number of repetitions that produced a result at least as extreme as what the research study found.

Question	*Student response we hoped to see.*
a) What did the cards represent?	*"The subjects in the research study."*
b) What did shuffling and dealing the cards represent?	*Discussion of the random assignment of subjects to the treatment groups.*
c) What kind of people did the face cards represent?	*Specification of the "successes" in the context of the research study.*
d) What implicit assumption about the two groups did the shuffling of cards represent?	*Explanation that this shuffling represents the "null model" of no treatment effect, but most mentioned only "random assignment."*
e) What observational units were represented by the dots on the dotplot?	*Realization that these observational units are the simulated repetitions of random assignment.*
f) Why did we count the number of repetitions with 10 or more "successes"?	*Statement that the reason is to compare to the actual experiment results, but most just say "to see if it's extreme."*

Figure 1. Activity questions and desired student responses

One limitation we have found is getting the students to be specific enough in their responses for us to fully assess their understanding. These questions could work well with follow-up interviews of select students.

We have also used this assessment for small-scale classroom experiments. For further details on these experiments and results see Holcomb, Chance, Rossman, Tietjen and Cobb (2010).

Components of P-value Interpretation

In evaluating students' ability to interpret p-value as a probability, we identified the following four desired components.

- Probability of observed data - can students identify the correct "outcome" of this probability statement?
- Tail probability - do students recognize that p-values are tail probabilities? Do they correctly indicate the direction that is more extreme (Ha)?
- Based on randomness - do students correctly identify the source of the randomness in the study (e.g., random sampling or random assignment)?
- Under null hypothesis - do students understand, and recognize the importance of noting, that the p-value is calculated under the assumption that the null hypothesis/model is true?

Based on the scoring of free response questions on the AP Statistics exam, each of these components was scored as essentially correct (E), partially correct (P) or incorrect (I) rating. Generally, the difference between an (E) or a (P) was if the general idea was stated, but it had not been correctly applied to the context at hand, then we graded it a (P).

Our goal was to see whether students performed differently on some components (e.g., depending on context or study design) and whether that performance changed consistently from early in the course to the end of the course. For example, students were given the question:

> *In 1977, the U.S. government sued the City of Hazelwood, a suburb of St. Louis, on the grounds that it discriminated against African Americans in its hiring of school teachers (Finkelstein and Levin, 1990). The statistical evidence introduced noted that of the 405 teachers hired in 1972 and 1973 (the years following the passage of the Civil Rights Act), only 15 had been African American. But according to 1970 census figures, 15.4% of teachers employed in St. Louis County that year were African American. Suppose we model Hazelwood's hiring practices as a Bernoulli process, then we find the p-value is less than .0001. Provide a one-sentence interpretation of this p-value (what is it the probability of?) in this context.*

The following are some example student responses to which we have applied the rubric in an attempt to differentiate quality of response:

- This p-value is the probability that 15 African-Americans or less would be part of the 405 teachers in Hazelwood, if Hazelwood was employing equally the same proportion of African-Americans as St. Louis county. (EEPE)
- This p-value describes the probability of 15 or less A.A. teachers out of a hired group of 405 teachers (in 1972-1973 in St. Louis) strictly by chance in 400 different sampling groups. (EEEP)
- This is the probability of observing 15 hired African-Americans out of a random sample of 405 teachers if 15.4% of teachers are African-American. (EIEE)
- There is a small probability, close to 0, that by randomization we would get fewer than 15 African-American teachers hired. (EEPI)

Some examples of responses that are more difficult to categorize with this rubric include:

- Common misconception response (e.g., p-value is probability of rejecting mull): The probability that a Hazelwood teacher is African-American is less than 0.154.
- Student uses *p*-value to make a decision concerning hypotheses without interpretation: Since the p-value is so small, we reject Ho. So there is strong evidence that the City of Hazelwood discriminated against African Americans in its hiring.

One concern with this type of assessment is that many students often say something to the effect that a small p-value tells them "we wouldn't get results like this by chance" or that "we can't attribute it to random chance." We consider this to demonstrate positive gains in understanding early in the course, but we eventually want students to articulate what "like this" and "by chance" mean in the context of the study. Again, such responses may not accurately reveal the entire depth of their understanding. Do we need to modify what we teach to raise their understanding level, or does this represent an assessment issue in that we have not provided questions that prompt students to reveal their full understanding?

Applying Concepts to Other Studies

Our intention here is to determine how well students can apply their understanding of the randomization process and statistical significance to a new context. For example, after completing a module that involved 2×2 tables, students would be presented with the context of a different research study for which the results could also be summarized in a 2×2 table. Students are then asked the questions in Figure 2.

(a) Describe in detail how to conduct a tactile (hands-on) simulation to investigate the research question.

b) What is the "null model" for this simulation analysis?

c) Use a dotplot of simulation results to calculate the approximate p-value for this study. Also circle the dots that you are referring to for this calculation.

d) Based on this simulation analysis, does the difference between the two groups appear to be statistically significant?

e) What conclusion would you draw from this simulation analysis, concerning the research question of whether children praised for their intelligence are significantly more likely to lie than those praised for their effort? Also explain the reasoning process by which you are reaching this conclusion.

f) Provide a complete, detailed interpretation (in one or two sentences) of what this p-value measures in this context (i.e., what is it the probability of?).

Figure 2. Questions students were asked after completing a module involving 2x2 tables

Multiple Choice Questions

As stated earlier, one of our primary goals is to use assessment instruments that will help us quickly assess students' understanding of statistical significance The following ten multiple choice questions were designed 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. One of these questions is taken from the CAOS instrument. Here we also provide some summary statistics of student performance in an introductory statistics class of students at Cal Poly. We provide the performance values for the first seven questions to indicate that we believe 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 not 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 does turn out to be statistically significant, which of the following is a valid conclusion?

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

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

8) You plan to use a random sample of students at your school to investigate a claim that the average amount spent on the most recent haircut by the population is more than $15. Why would a large (random) sample be better than a small one?
a) Because you're more likely to get extreme haircut prices with a larger sample.
b) Because larger samples produce less variability in haircut prices within sample results.
c) Because larger samples produce less variability in average haircut prices from sample to sample.

(CAOS) 9) Researchers surveyed 1,000 randomly selected adults in the U.S. A statistically significant, strong positive relationship was found between income level and the number of containers of recycling they typically collect in a week. Please select the best interpretation of this result.
a) We cannot conclude that earning more money causes more recycling among U.S. adults because this type of design does not allow us to infer causation.

b) This sample is too small to draw any conclusions about the relationship between income level and amount of recycling for adults in the U.S.
c) This result indicates that earning more money influences people to recycle more than people who earn less money.

10) Alicia wants to know if she receives higher tips, on average, when she gives customers her name compared to when she does not. She decides to track her tips for one week keeping track of the amounts and whether she gives the customers her name or not. She finds that giving her name led to a statistically significant higher average tip amount. Of the options below, what additional information is necessary for Alicia to be able to conclude that giving her name causes a higher tip, on average?
a) The number of tables where she does and does not give her name
b) The method Alicia used to decide which tables she would tell or not tell her name
c) No additional information is necessary because the result was statistically significant.

We are beginning to examine the psychometric properties of content validity and reliability of these items (though we plan to keep the total number of items small). An advisory board of statistical educators has reviewed the materials and offered feedback.

Proposing Randomization Methods For Novel Scenarios

Another goal of ours is that students understand randomization tests well enough to propose a reasonable method for such a test in a scenario that they have not yet studied. For example, having studied randomization tests for comparing two groups, can students describe a reasonable randomization procedure for comparing three groups (perhaps given a test statistic to use)? Another example would be to see whether students can propose a reasonable randomization test procedure for a matched-pairs design, without having studied such a scenario. We have asked final exam questions of this type. Some students do quite well, others are not yet able to make the full transition away from more traditional methods once they are covered in the course.

CONCLUSION

We believe that these assessment instruments provide a “good start” in attempting to investigate the depth of understanding that students develop regarding the important concept of statistical significance. Not surprisingly, we have found that assessing students’ conceptual understanding is much more challenging than determining whether students can perform the calculations of a *t*-test or make a correct decision based on a p-value and significance level. We also believe that the types of assessments that we have described here are helpful for revising classroom activities and for improving student learning.

REFERENCES

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

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.

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

Lane-Getaz, S. J. (2007). Toward the Development and Validation of the Reasoning about P-values and Statistical Significance Scale. *Presentation at the International Statistical Institute / International Association of Statistical Education Satellite Conference: Assessing Student Learning in Statistics,* Guimarães, Portugal, August 19-22, 2007.

Lane-Getaz, S. J. (2008). Introductory and intermediate students’ understanding and misunderstanding of *P*-values and statistical significance. *Proceedings of the 11th International Congress on Mathematical Education* (ICMe-11). Online: http://tsg.icme11.org/tsg/show/15.