Evaluation of an Interactive Tutorial for Teaching Hypothesis Testing Concepts

Christopher L. Aberson
Humboldt State University

Dale E. Berger
Michael R. Healy
Victoria L. Romero
Claremont Graduate University

In this article, we describe and evaluate a Web-based interactive tutorial used to present hypothesis testing concepts. The tutorial includes multiple-choice questions with feedback, an interactive applet that allows students to draw samples and evaluate null hypotheses, and follow-up questions suitable for grading. Students either used the interactive tutorial (n = 15) *or completed a standard laboratory assignment* (n = 10) *covering the same topics. Students who used the tutorial performed better* (p = .06) *on a quiz than students who completed the standard laboratory, supporting the effectiveness of this freely available online tutorial. A second group of students* (n = 112) *who did not participate in the assessment overwhelmingly rated the tutorial as easy to use, clear, and useful.*

Null hypothesis significance testing (NHST) procedures are a primary focus of introductory statistics courses (Friedrich, Buday, & Kerr, 2000). Recent criticisms of NHST focus on misunderstandings and misuses of hypothesis testing (e.g., Cohen, 1994; Nickerson, 2000). Nonetheless, hypothesis testing remains central to psychology students' abilities to understand research reports and conduct independent research. As such, it is essential that students gain a detailed understanding of hypothesis testing. We believe that many statistics courses emphasize rote learning of hypothesis testing, focusing on mechanical approaches to drawing statistical conclusions. Students learn to reject the null hypothesis if a computed value is larger than a comparison value. Although a mechanistic approach allows students to produce correct answers in simple situations, students who learn by rote are less likely to develop a deeper understanding of topics (Lovett & Greenhouse, 2000). Consequently, misunderstandings may occur regarding interpretation of statistical significance and their practical importance.

The Web Interface for Statistics Education (WISE) project provides instruction that addresses these shortcomings. As part of the WISE project, we created several Web-based tutorials that require only a JAVA-enabled browser. In this article, we present an interactive tutorial to assist students in learning about hypothesis testing with the *z* distribution. We use a normal distribution approach, as it is our impression that many introductory statistics courses use *z* as an introduction to hypothesis testing. The tutorial (found at http://wise.cgu.edu under "tutorials") consists of a paper-based assignment that guides students' use of an interactive applet, follow-up questions appropriate for grading or discussion, and on-screen multiple-choice questions that allow students to gauge understanding as they progress through the tutorial. This assignment assumes knowledge of *z* and normal distribution probabilities as well as some in-class introduction to hypothesis testing.

The tutorial begins with a description of a research scenario. The task is to investigate the effectiveness of three training programs. Students examine the mean and standard deviation of scores for a population of students who took a standardized test but did not take part in any training course (null population) and the means for populations of students who completed one of three training programs. One program is very effective (i.e., program mean is much larger than the mean for students with no training), the second is moderately effective, and the third is slightly effective. The effectiveness of the three programs corresponds to Cohen's large, medium, and small effect sizes when compared to the null population (Cohen, 1987).

Next, the student answers a series of multiple-choice questions involving computation of *z*, probability, and judgments as to the effectiveness of a training program. These questions review topics and informally introduce hypothesis testing concepts (i.e., judgments of likely and not likely outcomes). Incorrect answers correspond to typical errors, such as failing to consider sample size for standard error calculations or choosing the wrong area under the normal curve. Incorrect answers lead to feedback that addresses why the answer is incorrect and provides guidance for obtaining the correct answer. For example, one question asks the student to choose the probability of obtaining a certain *z* score. Feedback for incorrect answers address likely faults in reasoning.

Students then draw samples using an interactive applet that graphically represents the population distributions and sampling distributions for the training and no-training groups. Using the applet, the student can manipulate population mean, sample size, and standard deviation. The student can observe changes in these values immediately in the display of the sampling distributions. However, this exercise asks the student to modify only the means to represent the various training populations. The student begins by drawing a single sample. The applet plots the distribution of the individual scores in the sample and presents the calculated values for the mean and z. The student then compares the obtained z to a criterion ($z = \pm 1.96$) and draws a conclusion regarding the null hypothesis. A brief description of decision criteria presents alpha as a standard value for determining whether a sample result is unlikely given that the null hypothesis is true. The student draws 19 more samples, indicating a decision for each sample. For the first exercise, the student examines the highly effective training program, for which most sample means lead to rejection of the null hypothesis.

Next several multiple-choice questions focus on formal aspects of hypothesis testing. One problem provides the mean and z for a sample drawn from one of the training programs. The student must correctly determine the null hypothesis and statistical conclusion. These questions combine with the first sampling exercise to provide the student with a formal understanding of the terminology (e.g., alpha), logic (e.g., low probability suggests null hypothesis is unlikely), and mechanics of hypothesis testing procedures (e.g., reject null hypothesis when z exceeds criterion).

Following these questions are two additional exercises. One exercise examines results for samples taken from a population that differs moderately from the null population (medium effect size). The final exercise uses a population that differs only slightly from the null population (small effect size). Again, the student draws 20 samples, records means, and makes decisions regarding the null hypothesis for each sample.

After completing the sampling exercises, students answer follow-up questions that do not involve computer-based feedback. These questions are appropriate for grading or discussion. One question asks students to examine differences between the frequency of hypothesis rejection for the small and large effect size examples. The student comments on differences, indicating that the example with the larger effect size yielded more rejections of the null hypothesis. The problem then asks the student to suggest reasons for these differences. This type of question requires the student to think about the relations between the two distributions and consider plausible reasons for different rejection rates. Another question asks the student to evaluate a situation in which a group of program graduates requests a refund from one of the training programs. The graduates claim that the average of their scores was so low on the standardized test that the program's claims of a certain mean score for graduates could not be true. The student must apply hypothesis testing principles to address the probability that a sample of students would obtain a mean value that deviates from the population mean by a specific amount and discuss the implications of this result for the training program.

Method

Participants

One section of introductory statistics ($n = 25$; 23 women, 2 men), enrolled at a medium-sized rural state university, participated in an assessment of tutorial effectiveness. Students in this course registered for one of two laboratory sections. Each laboratory section consisted of a 50-min period held twice a week, wherein students received weekly assignments. We selected this group for assessment because the two laboratory sections allowed for assignment of separate exercises. These students completed the tutorial ($n = 15$) or a standard laboratory assignment ($n = 10$).

A second group of 112 students, 33 men and 75 women (4 no response on sex) enrolled in introductory statistics ($n = 67$; three sections, primarily sophomore/junior) or intermediate statistics ($n = 45$; two sections, primarily senior/graduate), completed the tutorial assignment and rated the tutorial but did not participate in the effectiveness assessment. Participants were students from a small, private liberal arts college ($n = 67$), a large urban state university ($n = 31$), and a medium-sized rural state university ($n = 14$). Most students were traditional college age.

The first author taught all lecture and laboratory sections and used the tutorial as one of the course's required laboratory assignments. Laboratory assignments prior to using the tutorial included computer-assisted data analysis, hand calculation problems, and other interactive tutorials.

Effectiveness Assessment: Comparison With a Standard Laboratory

The tutorial group received a packet with an overview of the tutorial, an exercise using the tutorial, and follow-up questions. The standard laboratory group received a paper-and-pencil assignment including four z-test problems from a statistics text (Howell, 1999) and the same follow-up questions as the tutorial group. One laboratory section used the tutorial and the other completed the standard laboratory. Practical issues prevented the use of random assignment (no separate facilities to conduct two conditions simultaneously). The instructor spent the first 5 min of laboratory reviewing the exercise and addressing questions about the assignment.

Students worked as they would regularly on their weekly laboratory assignments. Each student had access to a computer and could consult with the professor, teaching assistant, and other students for help. The initial laboratory for both groups followed a lecture introducing hypothesis testing with z. There was no lecture period preceding either group's second laboratory session. During the second laboratory session, students completed and turned in their assignments, after which we administered the comprehension test. Students in both groups also rated ease of use and interest in using similar assignments in the future.

Comprehension. A 10-item quiz covered the statement of hypotheses, calculating z, drawing statistical conclusions, and normal distribution probabilities. Items came from the test bank of the textbook that provided the standard laboratory questions. The use of two equivalent versions of the quiz minimized the potential for dishonesty. Students in both groups received either one of the two quizzes randomly. The two quizzes yielded comparable scores, $F(1, 23) < 1$.

Ratings-Only Group

Students in the ratings-only group ($n = 112$) responded to questions regarding how easy the tutorial was to understand, how clear they found explanations of statistical concepts, how useful they viewed the tutorial, and their desire to use similar assignments in the future.

Results

Effectiveness Assessment: Comparison With a Standard Laboratory Assignment

An ANCOVA, controlling for student grade percentage (arsine transformed) prior to laboratory completion, found that the tutorial group (adjusted $M = 7.50$, $n = 15$) performed better than the standard laboratory group (adjusted $M = 6.14$, $n = 10$), $F(1, 22) = 3.96$, $p = .06$ (two-tailed), $\eta^2 = .15$. This analysis allowed us to control for some systematic differences between students enrolled in different laboratory sections, as the first author observed an unequal distribution of the courses' top students between conditions. A test of the homogeneity of regression assumption revealed similar relations between current grade and performance for the tutorial and standard laboratory groups, indicating that ANCOVA was appropriate for these data. Although this result did not reach traditional levels of statistical significance, the effect size was encouraging and the two-tailed approach was conservative. Student reports of time spent completing the assignment indicated no substantial difference between the tutorial group ($M = 117$ min) and the standard laboratory ($M = 107$), $F(1, 22) < 1$.

All students using the tutorial judged it as easy to use, compared to only 30% of the students who judged the standard laboratory assignments, Fisher's exact test, $p < .001$. Students who used the hypothesis testing tutorial also indicated greater interest in using similar assignments in the future (71% *very interested,* 29% *somewhat interested*) than students who completed the standard laboratory (30% *very interested,* 60% *somewhat interested,* 10% *not interested*), $\gamma = -0.72$, $p = .02$.

Ratings-Only Group

The second group of students rated the tutorial as somewhat or very easy to use (83%), the explanation of statistical concepts as somewhat or very clear (86%), judged the tutorial as somewhat or very useful for teaching statistics (95%), and were somewhat or very interested in using similar assignments to learn about other statistical topics (94%). Intermediate statistics students rated the tutorial as more useful, $\gamma = 0.72$, $p < .001$, and indicated greater interest in using tutorials to learn additional topics, $\gamma = -0.71$, $p < .001$, than did introductory statistics students.

Discussion

A test of comprehension suggested that our Web-based tutorial may be more effective than a standard laboratory assignment in teaching basic concepts of NHST. Students rated the tutorial as easier to use and expressed more interest in using similar assignments than students who completed the standard laboratory. Additionally, most students who used the tutorial viewed the explanation of statistical concepts as clear and useful. This combination of increased learning, student interest, and ease of use supports the effectiveness of the tutorial.

The hypothesis testing tutorial incorporates techniques for enhancing learning that may account for improved performances. Bjork (1994) reported that learners often believe that they understand concepts better than they really do. Multiple-choice questions with options that correspond to common mistakes can promote confrontation of misconceptions. When students make mistakes, the tutorial provides instruction that immediately addresses specific misunderstandings. Another potential problem in instruction is a failure to engage students in elaborative processing (i.e., thinking about topics; Hofer, Yu, & Pintrich, 1998). Lack of elaborative processing may be problematic in statistics courses if instruction focuses exclusively on process. Our follow-up questions promote elaborative processing by asking students to explain and apply the concepts they have learned. For example, one question asked students to address the complaints of people who have completed a training course but performed poorly on a standardized test. To address this complaint the student must apply hypothesis testing concepts. Our tutorial may also enhance learning through multimedia presentation. Information presented in multiple formats improves memory (Paivio, 1971). Our interactive applet presents information using text by giving the mean and z and graphically by plotting the mean of each sample in relation to the null and true distributions.

Several limitations temper these conclusions. Our assessment does not establish whether use of the computer, opportunity for feedback, or interactive content led to improved performance. Additionally, the sample for the assessment is small. Finally, it is unclear whether students only learn more initially or if our tutorial leads to long-term learning.

However, our hypothesis testing tutorial demonstrates that Web-based materials can incorporate important principles of good instruction. Given the encouraging results of this assessment, we suggest that interactive computer-based tutorials may effectively supplement traditional assignments.

References

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalf & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Cohen, J. (1987). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49,* 997–1003.

Friedrich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology, 27,* 248–257.

Hofer, B. K., Yu, S. L., & Pintrich, P. R. (1998). Teaching college students to be self-regulated learners. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulated learning: From teaching to self-reflective practice* (pp. 57–85). New York: Guilford.

Howell, D. C. (1999). *Fundamental statistics for the behavioral sciences* (4th ed.). Pacific Grove, CA: Duxbury.

Lovett, M . C., & Greenhouse, J. B. (2000). Applying cognitive learning theory to statistics instruction. *The American Statistician, 54,* 196–206.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

Notes

1. Grants from the Mellon Foundation, Claremont Graduate University, and Humboldt State University support the Web Interface for Statistics Education.
2. Send correspondence to Christopher L. Aberson, Department of Psychology, Humboldt State University, Arcata, CA 95521; e-mail: CLA18@humboldt.edu.