METHODS AND TECHNIQUES

A Hands-On Exercise Improves Understanding of the Standard Error of the Mean

Robert S. Ryan
Kutztown University

One of the most difficult concepts for statistics students is the standard error of the mean. To improve understanding of this concept, 1 group of students used a hands-on procedure to sample from small populations representing either a true or false null hypothesis. The distribution of 120 sample means (n = *3) from each population had standard errors that closely approximated those of the theoretical sampling distributions, thereby illustrating how the Central Limit Theorem provides a standard error to use for hypothesis testing. Performance on an exam about the standard error of the mean was significantly better for the students who had completed this exercise than for students in a control group.*

Psychology students generally consider inferential statistics to be one of the most challenging subjects in their curriculum. The logic of statistical inference depends on understanding the standard error and the related concepts of hypothetical sampling distributions under either a true or a false null hypothesis. Several studies have described hands-on exercises designed to facilitate learning about sampling distributions (e.g., Dyck & Gee, 1998; Gourgey, 2000; Johnson, 1986; Rossman & Chance, 2000). However, these studies did not address whether understanding the standard error, in particular, can be improved specifically by hands-on experience with constructing both a null-true and a null-false sampling distribution.

The Dyck and Gee (1998) and Johnson (1986) articles presented exercises similar to the one in this study, but with some important differences. In both studies, the participants themselves did not randomly draw members from a population to form samples. In Dyck and Gee's study, each participant determined the value for one member of a population by counting the number of blue M&Ms® in a package, and the instructor selected members from that population to form the samples. In Johnson's study, the instructor presented each student with three or four equal-sized samples that he told them had been randomly selected from a population. In addition, in both studies participants did not draw samples from a population that could have had either a mean specified by a null hypothesis that was true or some other mean. In Johnson's study, the participants knew the population mean in advance. In Dyck and Gee's study, the participants knew that all of the samples came from a population with the same mean, which they calculated after the sampling process.

These procedures are appropriate for exercises that focus on improving understanding of the shape of the sampling distribution and the relation between its mean and the population mean. However, a critical idea for understanding the standard error is the distinction between the variability of members of a sample versus the variability of the sample means. An exercise in which participants draw the members of the samples themselves, as well as place the means of the samples in a distribution, would provide a way for them to experience the critical distinction directly. Furthermore, drawing such samples without knowing from which of two populations they came would enable the participants to notice that it is specifically the variability of sample means around the mean of all the sample means that is most directly relevant to deciding whether to accept or reject the null hypothesis. Therefore, this article presents a hands-on exercise in which students directly experienced the hands-on construction of both a null-true and a null-false sampling distribution.

The reason for the usefulness of such an exercise stems from a common problem that often interferes with students' understanding of the standard error, namely that they do not clearly distinguish between distributions of scores and a distribution of sample means. Most texts on statistics for the behavioral sciences describe a sampling distribution of the mean as consisting of the means of all possible samples of some size taken from the same population (e.g., Christensen & Stoup, 1991; Gravetter & Wallnau, 2000; Grimm, 1993; Hurlburt, 2003; Witte & Witte, 2004). Therefore, understanding the idea of a sampling distribution requires that students understand that they must now consider one of the measures that they formerly calculated for a set of scores (i.e., the mean) to be one of many single scores in yet another distribution (i.e., the sampling distribution). At the same time students are trying to learn this distinction, however, they are already trying to deal with a great deal of abstractness. For example, the characteristics of a sampling distribution, such as its central tendency and

variability, are abstract ideas that students measure using mathematical tools, which are also abstract. Furthermore, the measurements apply not to something concrete (e.g., measuring the area of a room), but to a distribution of scores, which is itself an abstract concept. Therefore, students may focus on the idea of "every possible sample" as the most salient concept in the definition of a sampling distribution. As a result, they may believe that the sampling distribution is simply an amalgamation of a large number of samples, which would lead to the misconception that the standard error measures the variability of individual scores rather than the variability of sample means.

Misconceptions are often difficult to overcome specifically because people disregard correct information about the misconception (Eaton, Anderson, & Smith, 1984). However, information that students generate for themselves is harder to disregard (Slamecka & Graf, 1978). In fact, Chi, deLeeuw, Chiu, and LaVancher (1994) found that self-generated information was specifically beneficial for the especially difficult task of overcoming misconceptions.

Overcoming students' common misconception about the standard error rests at least in part on distinguishing between a distribution of scores and a sampling distribution. Therefore, this exercise maximizes the likelihood that students would notice that the distribution (of which the standard error measures the variability) consists of sample means, not individual scores.

A new aspect of the exercise that was not present in previous hands-on constructions of sampling distributions highlighted this distinction. Specifically, students constructed sampling distributions by drawing samples both from a null-hypothesis-true and a null-hypothesis-false population. This innovation focused students' attention on how a researcher is able to draw an inference about a population mean from a sample mean, even without knowing from what population the sample actually came, specifically because the sample mean is one of many sample means in a hypothetical distribution.

Method

Participants

The participants in the control group were 17 graduate students in an inferential statistics course in a counseling psychology master's program at Kutztown University in the fall of 2001. The participants in the experimental group were 29 graduate students in the same course in the fall of 2002.

Materials

Populations for the exercise. I used two sets of bags to hold slips of papers showing the values in the populations representing the true null hypothesis and the alternative hypothesis. Figure 1 shows the values that I used and how many of each value were in the bags. Both populations consisted of the same values but were skewed to produce different means. As a result, the participants could select samples from populations with different means, but could not tell from which population they were selecting by looking at the values.

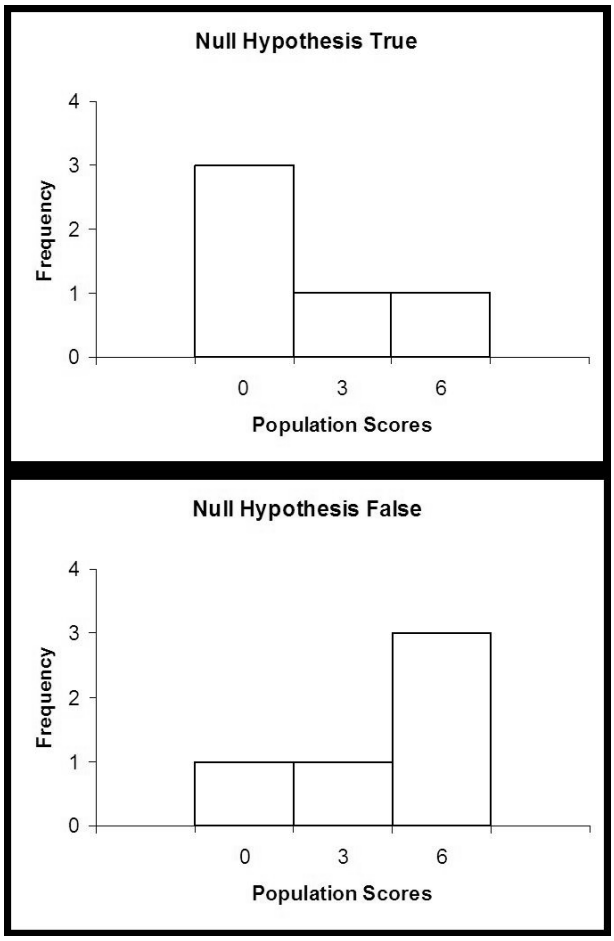

Figure 1. The populations.

Theoretical sampling distributions. As shown in Table 1, the difference between the population means, and therefore between the theoretical sampling distribution means, was 2.4. Both sampling distributions consisted of the means of all 125 possible samples ($n = 3$) using sampling with replacement.

Exam. Both groups took an exam on sampling distributions. A subset of 15 items that were identical for the two groups assessed the effectiveness of the exercise. Where there were multiple items on the same type of question, intercorrelations assessed reliability.

Evaluation survey. The students evaluated the hands-on exercise by taking a short survey. It consisted of one general open-ended question and five more specific questions.

Table 1. Characteristics of the Two Types of Sampling Distributions of the Mean Under Different Hypotheses

	Sampling Distribution	
Null Hypothesis	Theoretical	Empirical
True		
M	1.80	1.98
Standard error	1.39	1.35
Type I error rate	0.13	0.18
False		
M	4.20	4.13
Standard error	1.39	1.72
Power	0.72	0.70

Procedure

I introduced the exercise with an example to remind the students that researchers take a random sample to make an inference about a population mean. I told the students that there were six bags that had a population in which the null hypothesis that $\mu \leq 1.8$ was true and another six that had populations in which the alternate hypothesis that $\mu > 1.8$ was true. I instructed the students to randomly select samples of $n = 3$ using sampling with replacement.

After the students selected each sample, I told them to calculate its mean. If the sample mean was 4 or greater, the students rejected the null hypothesis. Otherwise, they failed to reject. Although this criterion resulted in the unusual significance level of .128, it facilitated calculating the empirical Type I error rate because every sample mean would fall clearly above or below the criterion. Finally, they recorded the value for each member of the sample, the sample mean, and the decision. Twelve small groups of students selected 10 samples, guessed which population their samples came from (in all cases the guesses were correct), and then traded bags with a group that had the opposite population and selected 10 more. I used the groups' recorded results to form empirical sampling distributions that the students compared to the theoretical sampling distributions in the next class. The entire exercise took about 30 min.

In the next meeting of this class, I illustrated the Central Limit Theorem by presenting the empirical sampling distributions formed from the 120 randomly selected samples from each population. Table 1 shows that the means, standard errors, Type I error rate, and power of the empirical sampling distributions were very close to those for the theoretical sampling distributions. I taught the concepts of sampling, sampling distributions, the Central Limit Theorem, errors of inference, and power to both groups in the same way except that I referred the hands-on group to the sampling distributions they had constructed, whereas I referred the control group to an example that I constructed at the blackboard and to textbook material. Due to the greater length of time required for explaining and conducting the exercise, the hands-on group spent two classes accomplishing the same lessons that the control group accomplished in one class. Both groups took their exam on sampling distributions at the beginning of the class that immediately followed the class (or classes) on sampling distributions.

Results

Exam Results

As predicted, the hands-on group achieved higher exam scores ($M = 94.7$, $SD = 6.5$, $N = 29$) than the control group ($M = 90.2$, $SD = 10.2$, $N = 17$). The scores of the hands-on group ranged from 80% to 100% correct, whereas the scores of the control group ranged from 63.3% to 100% correct. Because the sample sizes for the exam scores were different and the variances were unequal, $F(16, 28) = 2.47$, $p = .018$, I first analyzed the data with an unequal variance t test. Although this test did not reveal a significant difference, $t(23.7) = 1.64$, $p = .058$ (one tailed), I also conducted a more sensitive test and one that used equal sample sizes. This test used the percentage of students who were correct on each exam item as the unit of analysis. In this analysis, the test items formed matched pairs of percentage correct for the two groups. Because this paired t test took into account the differences in performance due to the different test items, it showed that the hands-on group performed significantly better than the control group, $t(14) = 3.14$, $p = .004$ (one tailed). In contrast to the exam on sampling distributions, even when all the items were identical, the hands-on group did not perform significantly better than the control group on any other exam (all ps $> .05$).

The reliability intercorrelations of four items that tested the application of knowledge of the sampling distribution of the mean were all significant at the .0001 alpha level. They ranged from $r(46) = .56$ to $r(46) = .99$ with an average of $r(46) = .81$.

Survey Results

Seventeen students (81%) reported that they believed they understood sampling distributions better as a result of doing the exercise and that the exercise was either enjoyable, fun, or both. Nine students reported other specific ideas that they understood better as a result of the exercise, including six concepts that students also have difficulty understanding (e.g., the implication of repeated sampling, the normal curve, the null hypothesis). The students rated the educational value of the exercise on a 7-point scale, ranging from 1 (*not valuable at all*) to 7 (*very valuable*). The mean rating was 5.29 ($SD = 1.52$). However, because there was a positive correlation between judgment of educational value and the extent

to which the exercise was fun, $r(21) = .75$, $p < .005$, these opinions may have simply reflected how much the students enjoyed the exercise.

Discussion

The students who participated in the hands-on sampling exercise performed better on an exam about the sampling distribution of the mean than the students who had not participated, even though they performed no better on other exams. The hands-on students' evaluations of the exercise suggested that they enjoyed it and believed that it was educationally valuable. However, given that the exercise for the hands-on group was spread over two classes, the possibility exists that their superior performance was due either to having more time to learn the material or to having spaced learning sessions.

The results for individual exam questions provided some insights, as well as some questions, about the exact nature of the benefits of the exercise. One of the exam questions asked what the standard error measures. The correct alternative stated, "How much the sample means differ from the population mean." Selecting that alternative suggests understanding that the sampling distribution consists of sample means and that their grand mean equals the population mean. However, students often selected the incorrect alternative, "The variability of the scores in the sample around the sample mean." Selecting that alternative shows confusion between the standard error and the standard deviation of the sample. In this study, the single greatest improvement was on that question, with only 53% of the control participants but 69% of the hands-on participants answering it correctly. This difference suggests that selecting samples and calculating their means called attention to the fact that not only is there variability among the scores in the sample, but also that the means of each of the samples differ from one another.

The other exam questions on which the hands-on group outperformed the control group, although not significantly, were those that asked how population variability and sample size affect the size of the standard error. Because it is so important for students to understand these relations, it might be worth speculating that the failure of this result to reach significance may have been due to Type II error, even though it could also have been due to chance alone. Perhaps the reason students typically fail to understand how population variability and sample size affect the size of the standard error when they hear about these effects in a lecture is because they lack the prerequisite knowledge. If so, then perhaps the hands-on exercise provided at least some of that knowledge, although not enough so that its effect was detected at a significant level in the data reported here. Therefore, future research might profitably examine ways of modifying the exercise to strengthen its ability to improve this particular aspect of students' understanding.

The overall exam improvement was only 4.5%, but it was an improvement over the already respectable 90.2% obtained by the control group. Furthermore, the hands-on group's score of 94.7% was the highest score for either group on any of the exams. Because graduate students generally have more background knowledge, more motivation, or both, than undergraduate students, undergraduates may benefit even more from this exercise. Future research should address this issue.

References

Chi, M. T. H., deLeeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477.

Christensen, L. B., & Stoup, C. M. (1991). *Introduction to statistics for the social and behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Dyck, J. L., & Gee, N. R. (1998). A sweet way to teach students about the sampling distribution of the mean. *Teaching of Psychology, 25,* 192–195.

Eaton, J. F., Anderson, C. W., & Smith, E. L. (1984). Students' misconceptions interfere with science learning: Case studies of fifth-grade students. *Elementary School Journal, 84,* 365–379.

Gourgey, A. F. (2000). A classroom simulation based on political polling to help students understand sampling distributions. *Journal of Statistics Education, 8.* Retrieved November 16, 2003, from http://www.amstat.org/publications/jse/secure/v8n3/gourgey.cfm

Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the behavioral sciences* (5th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Grimm, L. G. (1993). *Statistical applications for the behavioral sciences.* New York: Wiley.

Hurlburt, R. T. (2003). *Comprehending behavioral statistics* (3rd ed.). Belmont, CA: Wadsworth/Thomson Learning.

Johnson, D. E. (1986). Demonstrating the Central Limit Theorem. *Teaching of Psychology, 13,* 155–156.

Rossman, A. J., & Chance, B. L. (2000). Teaching the reasoning of statistical inference: A "top ten" list. *The College Mathematics Journal, 30.* Retrieved November 15, 2003, from http://www.rossmanchance.com/papers/topten.html

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592–604.

Witte, R. S., & Witte, J. S. (2004). *Statistics* (7th ed.). Hoboken, NJ: Wiley.

Notes

1. More detailed information about any of the materials, such as the evaluation survey and all of the exams, is available by contacting the author.

2. I thank Kathleen Kleissler, Adrienne Lee, Anita Meehan, Bob Voytas, Carole Wells, and two anonymous reviewers for helpful comments on earlier versions of the article.

3. Send correspondence to Robert S. Ryan, Department of Psychology, Box 730, Kutztown University, Kutztown, PA 19530; e-mail: rryan@kutztown.edu.