Journal of Statistics Education

Do Hands-On Activities Increase Student Understanding?: A Case Study

Thomas J. Pfaff
Aaron Weinberg
Ithaca College

Journal of Statistics Education Volume 17, Number 3 (2009), www.amstat.org/publications/jse/v17n3/pfaff.html

---

Key Words: Hands-On Demonstration; Active Learning; Central Limit Theorem; Confidence Interval; Hypothesis Testing.

Abstract

This article describes the design, implementation, and assessment of four hands-on activities in an introductory college statistics course. In the activities, students investigated the ideas of the central limit theorem, confidence intervals, and hypothesis testing. Five assessments were administered to the students, one at the beginning and end of the course, and three in between the activities. We found that, despite our attempts to engage our students in active reflection, their performance on the assessments generally did not improve. These results raise important issues about the design of pedagogical tools and activities as well as the need to gather data to assess their effectiveness.

1. Introduction

As statistics has become a focal point of both K-12 curricula (National Council of Teachers of Mathematics, 2000, 2006) and a required course for many undergraduate majors, there has been an increasing emphasis on helping students develop statistical reasoning. At our institution, which is a predominantly undergraduate comprehensive college, we teach introductory statistics courses to over 200 students each year.

As an essential component of statistical literacy, we want our students to move beyond simply computing confidence intervals and *p*-values to understanding what these concepts really mean and where they come from. Our goal was to design in-class, hands-on activities (which we called "modules") that would help our students develop an understanding of important statistical ideas. We decided to focus on determining the effectiveness of our activities in helping students increase their understanding of statistical concepts.

Our primary focus was our students in an entry-level Business Statistics course, of which our institution runs 6-9 sections each school year. The goal of our Business Statistics course is to teach students how to ask statistical questions, design an experiment, collect data, decide on an appropriate statistical test, and interpret the results. We were not concerned with deep theoretical explanations of statistical theorems. Instead, we decided to focus on important concepts that would help students interpret their results—such as a *p*-value or a confidence interval—in a meaningful way. We wanted our students to develop an understanding of:

- The central limit theorem
- Hypothesis testing and *p*-values
- The meaning of confidence intervals
- The role of variability

We believe that interactive data collection activities may increase student understanding of statistical concepts. Our belief is based on the idea that students do not learn passively, but instead learn by making new connections with their previous understanding to develop new knowledge structures (Von Glaserfeld 1987). In the learning process, the student's mind responds to cognitive conflict by actively arranging and rearranging mental structures. Our goal as educators is to create and structure this cognitive conflict in order to facilitate our students' active prediction and reflection that will generate new knowledge and understanding.

Our theory is borne out by other researchers. Snee (1993) makes an argument for experiential learning, and Gnanadesikan, Scheaffer, Watkins and Witmer (1997) claim: "Activity-based courses and use of small groups appear to help students overcome some misconceptions of probability and enhance student learning of statistics concepts." Some researchers have suggested the importance of experiencing data collection (e.g. Hunter 1977; Hogg 1991; Mackisack 1994) and many have recommended laboratory-based courses, in-class activities, and class projects (e.g. Hunter 1977; Dietz 1993; Fillebrown 1994; Mackisack 1994; Ledolter 1995; Bradstreet 1996; Chance 1997). Mills (2002) cites examples of students actively involved in data collection and analysis who developed a better understanding of various statistics concepts (e.g. Goodman 1986; Hubbard 1992; Mittag 1992; Gratz, Volpe, and Kind 1993; Packard, Holmes, and Fortune 1993; Sullivan 1993; Giesbrecht 1996; Marasinghe, Meeker, Cook, and Shin 1996; McBride 1996; Velleman and Moore 1996).

Gnanadesikan et al. (1997) describe this process of cognitive conflict and reflection taking place in a statistics classroom:

> When students are tested and provided feedback on their misconceptions, followed by corrective activities (where students are encouraged to explain solutions, guess answers before computing them, and look back at their answers to determine if they make sense), this "corrective-feedback" strategy appears to help students overcome their misconceptions.

Lunsford, Rowell, and Goodson-Espy (2006) argue that lectures and demonstrations were not as effective as hands-on activities for developing understanding:

> Just demonstrating graphical concepts in class via computer simulation was not sufficient for our students to develop these skills. We believed our students needed to have a directed (either through an activity or an exercise) hands-on experience (either in-class or out-of-class) with simulations that emphasize graphical representations of distributions (such as Sampling SIM or many of the simulations in the VLPS).

2. Design and Implementation Considerations

Several researchers (e.g. Hodgson (1996), Schwartz, Goldman, Vye and Barron (1997), and delMas, Garfield and Chance (1999)) found that introducing computer simulation activities into their classes increased their students' understanding but that this increase—while statistically significant—was not dramatic. Consequently, we sought recommendations for designing and implementing successful activities. Based on their own experience, delMas, Garfield and Chance (1999) recommended that activities:

1. Provide guidance to "facilitate exploration and discovery."
2. "Use simulations to draw students' attention to aspects of a situation or problem that can easily be dismissed or not observed under normal conditions."
3. "Provide a supportive environment that is rich in resources, aids exploration, creates an atmosphere in which ideas can be expressed freely, and provides encouragement when students make an effort to understand."
4. Provide representations for interrelated concepts and build connections "between different representations of the same phenomena."
5. "Help students evaluate the difference between their own beliefs about chance events and actual empirical results."

To incorporate these activities into our modules, we made the following design guidelines:

- The activity should be motivated by a "natural" question and students should explore the situation by gathering and interpreting data.
- The teacher should ask scaffolding questions, solicit conjectures, and ask students to explain their reasoning.
- Multiple representations (numerical descriptions, tables, and graphs) and modes (kinesthetic, verbal, and written) should be integrated into the activity and the teacher should facilitate making connections between the representations.
- The context should build on students' intuitions and experiences by using situations with which students are familiar and asking questions that students see as non-abstract. This should enable students to make predictions before engaging in the activity.
- Students should be asked to reflect on their predictions after they have collected data and again during a guided class discussion to articulate the difference between their predictions and the results.

3. Concrete Activities and Computer Simulations

A common strategy for implementing hands-on activities is to use a computer simulation method (CSM). Wood (2005) notes that computer simulations "tend to be more general [than conventional approaches], require far less technical background knowledge, and, because the methods are essentially sequences of physical actions, it is likely to be easier to understand their interpretation and limitation." Mills (2002) notes that while many researchers have recommended the use of CSMs, "a review of the literature reveals very little empirical research to support the recommendations."

When having our students perform simulations on their calculators or computers, we noticed that they seemed to not believe that the simulation reflected what would "actually" happen. For example, we had students simulate putting a group of people in an elevator by randomly selecting sample weights from a distribution of U.S. adult weights. After simulating several elevators, students still claimed that "in reality" the filled elevators would be much heavier than their simulation predicted.

Instead of using computer-based simulations, we decided to incorporate physical objects into our activities. We hypothesized that by using concrete objects, the activity would provide more opportunities to create and structure cognitive conflict and to facilitate our students' active prediction and reflection. In an appropriate concrete activity, students cannot be passive and simply observe the data collection and interpretation. Instead, they are personally involved with the underlying population, making their sampling more realistic. Furthermore, the results of a concrete activity wouldn't merely reflect "reality," but would be unarguably real.

For example, a computer simulation approach to investigate the meaning of a 90% confidence interval might have a student rapidly simulate the generation of 100 (or more) confidence intervals from some fixed population and observe that about 90 of them trap the true parameter. While this does not take much class time, students do not interact with the underlying population, do not experience the sequence of samples that gradually suggests the value of the underlying parameter, and may not readily experience the confusion that arises when some of the resulting intervals fail to trap this parameter.

In a concrete version of the same activity, we might give each student a population (e.g. a bag of blue and purple bingo chips) and have each student try to estimate the proportion of chips that are blue. In drawing a sample from the bag to create a confidence interval, students develop an intuitive understanding of the underlying population and, when comparing confidence intervals with their classmates, extend this intuition to the

relationship between the sample and the population. Through this process, students are constantly predicting the results and reflecting on these predictions in the class discussion.

With this reasoning, we added two additional guidelines for designing our activities:

- The population should be tangible and its parameters, while unknown to students, should be simple to compute.
- Data collection should be "slow" enough for students to understand the composition of their sample and compare it to the underlying population.

This reasoning suggests expanding the study to compare the effects of our concrete approach with a CSM. For this pilot project, however, we decided to focus on evaluating whether or not our activities were effective in developing students' understanding, leaving the comparison with CSMs for a future study.

4. Methods

We designed four modules to engage the students in actively making sense of the big ideas of the course. We used each module at the time in the semester when we would normally be discussing the corresponding topic.

Over the course of the semester, we also administered five written assessments to the entire class. The goal of these assessments was to evaluate our students' understanding of the "big ideas" before using the modules, soon after using the corresponding module, and again near the end of the semester. We primarily drew our questions from the "Tools for Teaching and Assessing Statistical Inference" web site (http://www.tc.umn.edu/~delma001/stat_tools/). Many of the questions were repeated on multiple assessments so we could determine if our students' performance changed over the semester. The questions used for assessment were not used on any other exam or homework problem in the course. The assessments were not administered by the course instructor, and the instructor did not see them until after final grades were submitted.

4.1 Modules

Detailed descriptions of the modules we used in class along with the associated worksheets are included in Appendix A.

4.1.1 Using Cards to Illustrate the Central Limit Theorem

The goal of this module is to introduce the central limit theorem and observe its effect on distributions. The activity suggested here is similar to one proposed by Gnanadesikan et al. (1997) who used the dates of a random collection of pennies for their initial sample. Our module offers an important improvement on the penny activity by allowing students to know the probability distribution for the population prior to sampling. While the distribution of pennies is often skewed and unimodal, our module begins with a distribution that is bimodal, ensuring that it will not look normal; it can also be easily modified to have an initial distribution that is skewed, although this will increase the sample size needed for the sampling distribution to look normal.

In this module, each student gets a suit of 13 cards. Each card is assigned a number equal to its face value with a Jack equal to 10, and the Queen and King each equal to 0. This creates a population with a U-shaped distribution. Students are asked to predict the shape of the sampling distribution (for their individual suit of cards) if they draw 30 cards with replacement, then compute μ and s. Each student then draws 30 cards with replacement and records the sample mean of the first 1, 10, 20, and 30 cards. As a class, the students record their results in a spreadsheet and predict what each distribution should look like, describing their shape, spread, and center. After generating a histogram of their results, the class computes the means and standard deviations for each distribution, describes what they see and discusses the results.

4.1.2 *T*-test with Dice

In this module, students construct confidence intervals and find *p*-values using a *t*-distribution. Although Dambolena (1986) and Gordon and Gordon (1989) encouraged readers to use computer simulations and graphics to enhance students' understanding of the t-distribution, it is not readily apparent that their methods offer a more effective instructional method than a hands-on approach.

Students are given three dice (a six-sided die, an eight-sided die, and a twelve-sided die) and investigate whether the mean of the sum of the dice is identical to the sum of their means by taking a simple random sample of n = 30 rolls of the three dice. Of course, students could calculate both of these means, but they recognize that calculating the mean of the sum involves substantial effort and so the statistical approach is helpful.

Before beginning the experiment, each student describes the population, the parameter of interest, the statistic that will be computed, states and writes a sentence describing the meaning of the null and alternative hypotheses, and predicts whether or not they think the null hypothesis will be false. Doing this should help students understand what hypothesis they are testing before they begin and forces them to use formal notation and language to describe the situation.

Each student then rolls the dice 30 times, computes $\bar{x}$ and *s*, and computes a 90% confidence interval and *p*-value. They enter their proportion, *p*-value, and confidence interval into a class spreadsheet. The class then examines a table of results, a graph of the distribution of samples and a graph of the confidence intervals and collectively decides whether the mean of the sums is equal to the sum of the means. After revealing to the class that these quantities should be equal, the students investigate the connection between these quantities and their confidence intervals along with the *p*-values and the cases in which they rejected H_0.

4.1.3 Bags and Chips: Proportion Tests and Confidence Intervals

The goal of this module is to explore the ideas of a hypothesis test and a confidence interval by having students try to determine if a bag contains equal proportions of two different colors of bingo chips. As with the previous module, students describe the population, parameter, statistic, H_0, and make a prediction about H_0 before collecting data. Each student then gets a small canvas bag containing 70 blue bingo chips and 30 purple chips; while the activity could be done with 10 chips in each bag, we believe it is helpful for the population to be larger so that a student can't easily describe it by glancing into the bag.

Students sample 45 chips with replacement and use this to compute a *p*-value and an 80% confidence interval (chosen so that we have a good chance that a few students would not trap the parameter). Each student enters their proportion, *p*-value, and confidence interval into a class spreadsheet. The class then examines a table of results, a graph of the distribution of samples and a graph of the confidence intervals and collectively decides whether their bags had equal proportions of the different chip colors; if they decide that the bags were not equally split, then they try to estimate what the actual split was. After revealing to the class that there was a 70/30 split, the students investigate the connection between this split and the confidence intervals along with the *p*-values and the cases in which they rejected H_0.

When comparing their confidence intervals, students should quickly notice that their centers vary widely (but the widths only vary a little); consequently, their *p*-values will also vary and not all of the intervals will capture the true proportion. The 70/30 split gives a power of 78.5%, which means roughly one fifth of the students will fail to reject the (false) null hypothesis.

4.1.4 Bags and Chips: Two Proportion Tests and Confidence Intervals

This module extends the one-proportion test by using two populations instead of one. A similar activity has been implemented with a CSM by Wood (2005) using the "two bucket story" to derive bootstrap confidence intervals and simulate probability distributions.

While students used a single bag of chips in the previous module, here they use two bags of chips to determine if the bags have identical proportions of blue chips. As before, each student describes the populations, parameters, statistics, H_0 and Ha, and makes a prediction about whether they will reject H_0 prior to starting the experiment.

Each student (or pair of students) then gets two small canvas bags. The first contains 70 blue chips and 30 purple chips; the second contains 60 blue and 40 purple chips. Students sample 45 chips with replacement from each bag and use this to compute a *p*-value and an 80% confidence interval. Each student then enters their results into a class spreadsheet.

After revealing the actual proportions to the class, students investigate the connection between this split and the confidence intervals and discuss the *p*-values and the cases in which they rejected H_0. As in the previous module, students should notice that the confidence intervals and *p*-values vary even though they all drew random samples from identical populations.

4.2 Written Assessments

While all of the items in our assessments were either multiple choice or true/false, every question was followed by a prompt for students to explain their reasoning. All assessments were administered during regular class periods by a colleague who was not teaching the course. The assessments can be found in Appendix B.

- Assessment 1 was given to students near the beginning of the semester. This assessment included questions about the relationship between statistics and parameters, the central limit theorem, the meaning of confidence intervals, hypothesis tests and *p*-values. While we didn't expect students to understand some of these technical concepts, we included them here so that we could compare their performance on this preliminary exam with later assessments.
- Assessment 2 was administered after the first module. It included questions designed to measure students' understanding of the central limit theorem.
- Assessment 3 was administered after the second module. It included questions designed to measure students' understanding of confidence intervals and hypothesis tests (specifically, the meaning of the null hypothesis) using population means as the parameter of interest.
- Assessment 4 was administered after the third module. It was nearly identical to Assessment 3 except for a re-ordering of some multiple-choice answers and a focus on population proportion instead of the mean as the parameter of interest.
- Assessment 5 was administered near the end of the semester. It was designed to measure the students' "retention" of the concepts they had worked with in the modules and included a subset of the questions that had appeared on the previous assessments. We decided to not include the entire set of questions so that our students could finish the assessment in one class period (50 minutes).

4.3 Analysis Techniques

After the end of the semester, we recorded each student's multiple choice and true/false answers as well as their written explanations in a spreadsheet. For each item, we identified the correct answer, assigning it a value of 1, and in a few instances we also gave partial credit to an answer that, while technically incorrect, still reflected an understanding of the "big idea" behind the question.[1]

For each item that appeared on multiple assessments, we conducted a Mc-Nemar test using SPSS to compare students' performance between each pair of assessments. For this analysis, partial credit was converted to full credit (a value of 1) due to the requirements of the test. Since we expected that our modules would improve our students' understanding of the concepts—and that their understanding would translate to increased performance on the assessments—we used a one-sided test.

For each collection of problems that appeared on multiple assessments, we computed an "exam score" for each student on each assessment by

finding the sum of their scores for those problems. Here, we used the partial credit as noted above. We then used a paired *t*-test in SPSS to compare students' scores on each pair of assessments.

Several assessments included multiple items that addressed the meaning of confidence intervals. Students' responses for these questions were cross-tabulated for each pair of questions in each assessment and analyzed for significant associations using a χ^2 test in SPSS.

In addition, we used the methods developed by delMas, Garfield, and Chance (1999) to analyze students' reasoning on the "sampling distributions" questions, which asked students to describe how the shape of a distribution of sampling means changes as you increase the sample size.[2] Students' answers were characterized as "correct reasoning," "good reasoning," "larger to smaller reasoning," and "incorrect reasoning."[3]

Although we included a prompt for students to explain their reasoning on each question, most students did not provide explanations. Furthermore, many of their explanations were little more than a restatement of their answer and did not enable us to draw much insight into their reasoning. Because of this, we decided to not use students' explanations in our analysis.

5. Results and Discussion[4]

5.1 The Big Picture

Overall, students' understanding of the statistical concepts did not seem to improve. While students showed some significant improvement on individual items that appeared on multiple assessments, their performance actually significantly decreased on others and showed no change for most items.

We will begin by describing the results for the "exam scores" and describe results for individual items below. For each collection of questions that appeared on multiple assessments, we found the students' average percentage score on the pair of assessments; the graph below (Figure 1) shows students' performance and results of the paired T test:

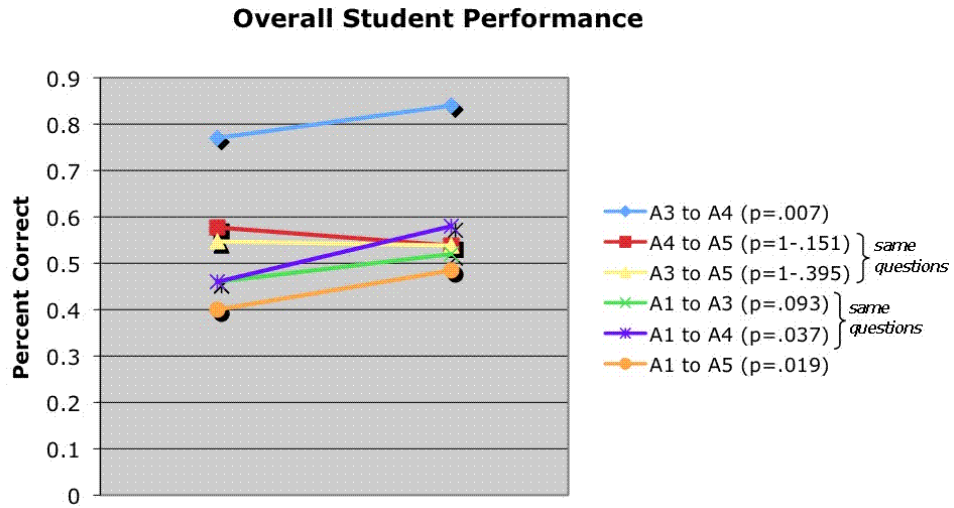


Figure 1: Exam score changes between subsets of questions

Although some of the increases were statistically significant, it's not clear that the increases were practically significant. In addition, with one exception, students rarely scored more than 60%. Although not significant, students' performance actually decreased on a set of items that appeared on assessments 3, 4, and 5.

Since these sets of questions are not all the same, we can't draw robust conclusions from these data. However, these data suggest that students seemed to slightly improve their understanding of the concepts, although students may have "lost" some of that understanding in the five weeks between using the modules and taking Assessment 5.

5.2 The Central Limit Theorem

To investigate students' understanding of the central limit theorem, we presented a distribution for a population and five potential sampling distributions (see Figure 2).

Students were asked to identify which histograms could correspond to sampling distributions for samples of increasing sizes and identify how these increasing sizes would affect the shape and spread of the distribution. This item appeared on Assessments 1, 2, and 5 with the only difference being the shape of the distributions. It should be noted that the differences between the histograms presented to the students may have been too subtle for the level of the course.

On assessments 2 and 5, roughly half of the students correctly responded that the sampling distributions should be shaped more like a normal

distribution and would have less variability when the sample size increased. Apart from this, students were generally unsuccessful at identifying how increasing the sample sizes would affect the distribution of sample means.

Not only did few students give correct answers, but there was very little improvement in performance between the three assessments. When asked to describe how increasing the sample size would change the distribution, students generally gave more correct answers on each successive assessment. However, these increases were not significant. When students were asked to identify the sampling distribution for samples of size 4 and 16, their performance actually decreased between the assessments; this decrease was significant in some cases.

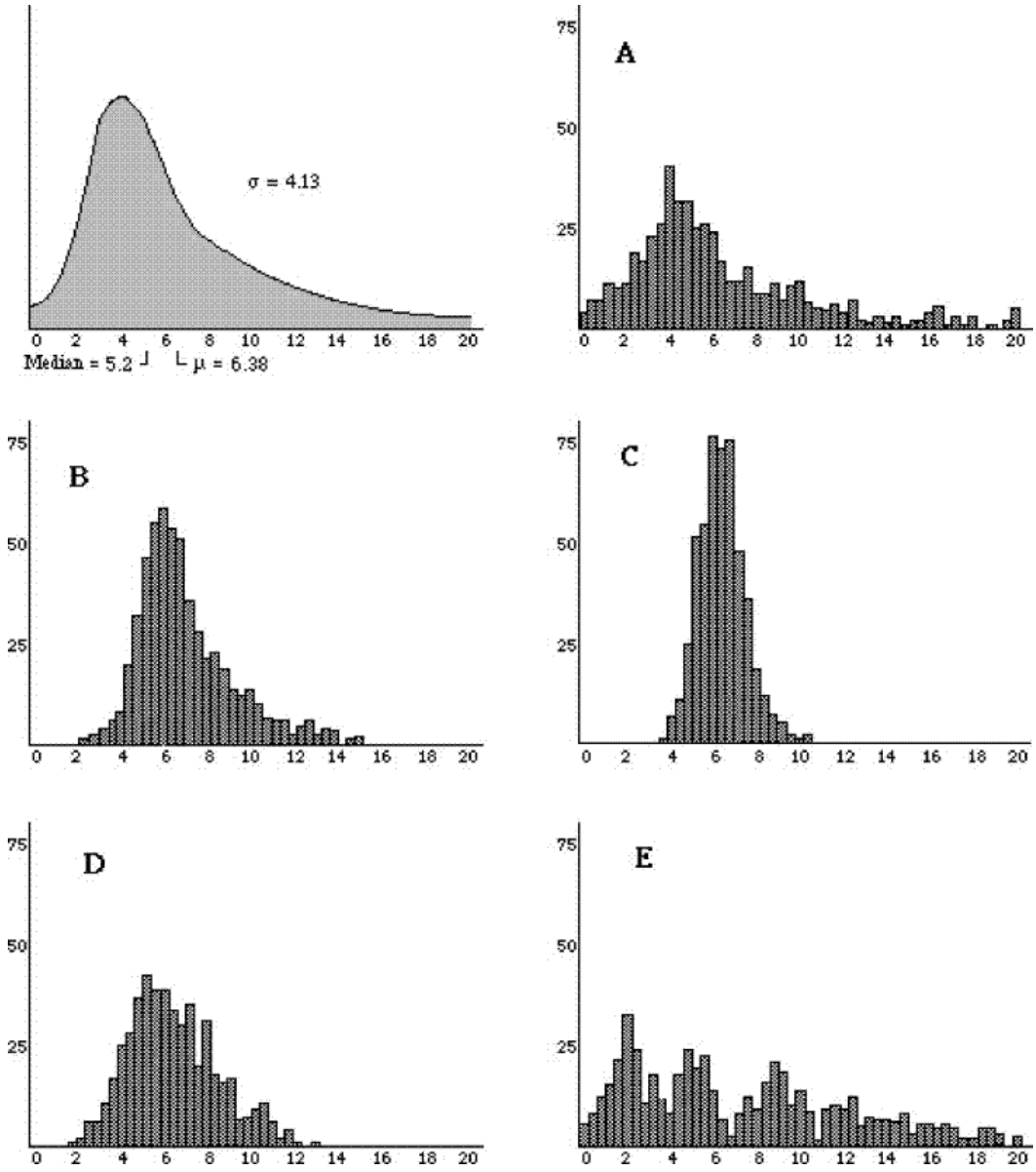


Figure 2: Population and Sampling Distributions

In addition, students reasoning about the effect of increasing the sample size was generally poor. Table 1 shows students' reasoning (using the categories described by delMas, Garfield, and Chance (1999))

Reasoning	Assessment 1	Assessment 2	Assessment 5
Correct	3	0	1
Good	6	2	5
Larger to Smaller	1	5	5
Incorrect	16	18	15

Table 1: Students' Reasoning About Increasing Sample Size

These data suggest that students not only failed to develop an understanding of the central limit theorem, but for some questions their original

intuition was apparently more accurate than their conception after the module and at the end of the course. In some sense, it appears that the students became more confused which may be a sign that they were thinking about the central limit theorem but had confounded multiple ideas about the concept.

This result—particularly for the assessment immediately following the module—was surprising. In the module, students individually gathered data and computed sample means for successively larger samples. The class predicted what the sampling distributions should look like, collectively plotted their data and discussed the results. The resulting distributions were reasonable illustrations of the central limit theorem, and we expected that the active prediction, reflection, and class discussion would help students develop an intuitive sense of how the sample size affected the distribution.

5.3 Confidence Intervals

Students in our class were generally able to successfully use a calculator to compute a confidence interval and then use this interval to decide whether or not to reject a null hypothesis. However, students had difficulty understanding the meaning of confidence intervals and how they are constructed. This is significant, for without this understanding it is not clear that students understand the accuracy and potential errors inherent in a confidence interval.

Students were somewhat successful at identifying that the sample statistic is guaranteed to lie in the confidence interval. Roughly half of the students identified this, with the percent increasing notably ($p = .072$) between Assessments 1 and 5. Since students relied primarily on their calculators instead of computing confidence intervals manually, perhaps it isn't surprising that they did not recognize the relationship between the sample statistic and the confidence interval. Despite this, it is troubling that some students thought that the sample size or standard deviation must lie within the interval, since both of these quantities have little to do with where a confidence interval is centered.

Table 2 is typical of our results. On the first, third, fourth, and fifth assessment (denoted A1, A3, A4, and A5), students were asked to select the meaning of a 95% confidence interval from the following four choices:

1. 95% of the intervals constructed using this process based on samples from this population will include the population mean
2. 95% of the time the interval will include the sample mean
3. 95% of the possible population means will be included by the interval
4. 95% of the possible sample means will be included by the interval

The first choice is the correct answer, although the second choice is not unreasonable. All four choices are similar, and this question requires a solid understanding of the meaning of a confidence interval (or memorizing what it means) to get it correct. Students' answers were classiffied as right or wrong (only the first choice was coded as right) and for each pair of assessments. Table 3 shows the percent of students who fall into each category of correctness.

As can be seen in Table 2 and 3 there were very few patterns in students' responses. Comparing A1, A3, and A4 to A5, they were slightly more likely to switch from a right to a wrong answer than they were a wrong to a right answer, although this was not significant. However, students were slightly more likely to switch from a wrong to a right answer than they were a right to a wrong answer when comparing A1 to A3 and A4 (which, again, was not significant).

	A1-6	A3-3	A4-3	A5-8
Correct Response	12	15	14	8
Incorrect Response	15	12	13	18

Table 2: Responses on A1 Problem 6, A3 Problem 3, A4 Problem 3, and A5 Problem 8

	A1-6:A3-3	A1-6:A4-3	A1-6:A5-8	A3-3:A4-3	A3-3:A5-8	A4-3:A5-8
n	27	27	26	27	26	26
wrong to right	30%	30%	15%	19%	12%	12%
right to wrong	19%	22%	31%	22%	35%	31%
wrong to wrong	26%	26%	38%	26%	35%	38%
right to right	26%	22%	15%	33%	19%	19%
McNemar's p-vlaue	0.291	0.396	$1 - 0.194$	0.500	$1 - .073$	$1 - 0.114$

Table 3: Response changes between A1 Problem 6, A3 Problem 3, A4 Problem 3, and A5 Problem 8

Students had varying success answering true/false questions about the meaning of a (95%) confidence interval, such as "If 200 confidence intervals were generated using the same process, about 10 of the confidence intervals would not include the population mean (μ)." While roughly 50-80% of students responded correctly on some of these questions, there were no distinct patterns of increase (or decrease) between the assessments. In addition, students generally had difficulty identifying what a 95% confidence interval means from a list of four options, and their performance decreased by the final assessment.

The assessments included multiple questions about confidence intervals, and it would seem reasonable that students' responses would demonstrate some consistency in their reasoning. For example, if a student thought that a 95% confidence interval meant that "95% of the intervals constructed using this process based on samples from this population will include the population mean," we would expect that student

to say that the statement: "If 200 confidence intervals were generated using the same process, about 10 of the confidence intervals would not include the population mean" was true. However, there was no significant association between these two responses (p=.756 on Assessment 3, p=.449 on Assessment 4, and p=.477 on Assessment 5). In fact, there were no significant associations between students' responses on pairs of questions like these. This suggests that the students had not developed a coherent conception of the meaning of confidence intervals. These results were troubling because a significant amount of time during the modules was spent having individual students compute confidence intervals and then comparing all of the intervals generated by the class. As expected, the centers of these intervals varied greatly and roughly 5% of them did not contain the true population parameter. We expected this hands-on method of data collection and statistic generation to give students a basic understanding of the meaning of a confidence interval.

5.4 Hypothesis Testing

Much like confidence intervals, students can frequently make a statement about a null hypothesis—whether or not it is false—but may not have a good understanding of what it is they are rejecting. On Assessments 3, 4, and 5, we asked students to indicate if it is possible to "prove the null hypothesis." On each assessment, roughly 80% of the students believed that, under certain conditions, it is possible to do so. However, based on informal conversations with students after all assessments were done, it appeared that some students had been thinking that they could prove the null hypotheses if they sampled the entire population.

Although most students in the class knew to reject a null hypothesis if the hypothesized value was not contained in the confidence interval, they had difficulty identifying a confidence interval that would allow them to reject a particular null hypothesis; there was no clear pattern to their errors.

Similar to the idea that different samples from a population can produce different confidence intervals, two samples from the same population in a well-designed study can produce different statistics. Students were asked to predict what *p*-values two researchers would get if they collected data in two random samples from the same population. Overall, students were relatively successful and generally improved from Assessment 1 to Assessment 4 (from roughly 65% getting the questions correct to 82% getting them correct). Since one of the choices for a response was "I'm not sure," it seems that roughly half of the students began with and maintained a reasonable intuition about this concept.

While some of the percentages are relatively high, we had hoped that our students would perform significantly better. As part of the modules, individual members of the class drew random samples from identical populations and computed their own *p*-values. These *p*-values were different for each person and some of the values were significant while others were not. Since students encountered and discussed this idea in multiple modules, we thought that their performance on these questions would be better than it was.

6. Conclusions

Our hands-on activities generally failed to help students develop a good understanding of the underlying statistical concepts. We had hoped that students would not only get a large percentage of the questions correct by the end of the course, but they would have also show significant improvement from their scores at the beginning of the course. Even though we thought we had designed and implemented these modules in a way that would help our students understand the "big ideas," our assessment showed that we did not accomplish our goal.

On a survey we gave at the end of the semester, students had a positive reaction to the modules. Many students (42%) listed the modules as the most interesting thing done in the course. When asked to indicate how beneficial the modules were on a scale of 1 (not beneficial) to 5 (extremely beneficial), the median score was a 4. Even though our modules did not effectively develop understanding, they did engage the students in the course.

These results have significant ramifications for teachers of statistics. No matter how innovative or stimulating a pedagogical idea may seem—and no matter how much the students seem to enjoy the class—it may not be sufficient to develop students' understanding. Numerous articles are published every year describing activities, pedagogical tools, and techniques that the authors believe will increase student understanding and engagement. However, it is imperative that these pedagogical innovations are tested and that we have empirical evidence of their effectiveness.

There are, of course, lurking variables in our study. It could be that our modules were .ne as written, but we did not implement them successfully or devote enough class time to their implementation. Conversely, our implementation may have been .ne but an aspect of the modules' design could have been confusing. Since the ideas we addressed in the modules were also discussed at other times during the semester, we can't distinguish between the effects of the modules and the effects of the other instruction; it could be the case that students did learn from doing the modules but then formed competing conceptions from other activities. Conversely, the modules were only implemented in one class period and students may have compartmentalized the class' discussion of the modules as distinct from the rest of the course. It also could be that our students were in some way unprepared to successfully reflect on their activities.

In addition, there are several methodological issues with our study. As previously mentioned, we could not separate the modules from the rest of the course, making it difficult to determine how the modules specifically affected student learning. Since we relied on written assessments and students wrote few explanations, we feel that we did not get a clear picture of our students' reasoning. It could be the case that many students could provide justifications for their answers that reflected a degree of understanding, or that students who provided correct answers were simply making educated guesses without really understanding the concepts.

Although our sample size was relatively small, it is unlikely that increasing the sample size would have shown significant overall improvement in student performance. Of the 76 comparisons for which we performed a McNemar test, students' performance decreased in 19 and did not change in 18. In addition, fewer than half of the tests that suggested an improvement in performance had *p*-values below .2. Consequently, a post-hoc analysis of power would not be particularly illuminating.

Our study design did not use a control group of students who either took an identical class (without using the modules) or—at least—a very

similar class. Such a group was not necessary to reach our conclusions— that the activities failed to substantially increase our students' understanding. However, when we conduct follow-up studies it will be important to have such a group to determine if the activities themselves are effective and if they are more effective than traditional instruction or CSMs.

An additional open question is the role that additional reflection and reinforcement might play in students learn from the modules. We did not assign any out of class work directly related to the modules. The questions in our assessments did not explicitly resemble the activities in the modules, and students were never directly tested on the content in the modules. As a result, we assume that the students spent minimal time outside of class reviewing and reflecting on the activities. Would this additional reflection help students develop and retain an understanding of the ideas addressed by the modules, and—if so—how much is needed?

Given these open questions, we can't conclude that the modules themselves were inadequate. However, this study has led us to rethink the design of the modules and helped us identify ways they might be improved. For example, when we use the modules in the future, we plan on giving students follow-up activities that have them spend more time describing key aspects of the concepts (such as the relationship between various confidence intervals and the associated parameter). Further, we plan to change and expand the way we assess our students' learning. For example, the way we actually used confidence intervals in class was to check whether or not the value of the null hypothesis was within the interval and—if it wasn't—to reject the null hypothesis; we plan on augmenting these tasks with writing assignments in which students explain their reasoning and the underlying statistical concepts. It is through this process of goal-setting, evaluation, assessment, and incremental improvement that we hope to not only help our students develop an understanding of statistics, but turn a reflective, critical eye on our own teaching to help us improve as educators.

Appendix A: Modules and Worksheets

Module 1

Description:

Each student gets a suit of (13) cards. We assign to each card the following value:

A	2	3	4	5	6	7	8	9	10	J	Q	K
1	2	3	4	5	6	7	8	9	10	10	0	0

1. Students predict the shape of the sampling distribution (for their individual suit of cards) if they draw 30 cards.
2. Students compute μ and s for their population (they should get μ =5.0 and s =3.5) and discuss why their answers make sense.
3. Each student draws 1 card and compute the mean value of this sample.
4. Continue drawing cards, and have each student record the sample mean of their first 10, 20, and 30 cards.
5. Each student records their results in a class spreadsheet (using Excel, Minitab, or any other statistical spreadsheet program).
6. As a class, have students predict what the distributions of each of these should look like, describing their shape, spread, and center. Then compute the mean and standard deviation and graph each distribution.
7. The class describes what they see and discuss these results.

Worksheet

Activity 1 Worksheet

The Central Limit Theorem

Here is our system for assigning values to each card:

A	2	3	4	5	6	7	8	9	10	J	Q	K
1	2	3	4	5	6	7	8	9	10	10	0	0

You will be drawing 30 cards (with replacement) and computing several means and standard deviations.
Before you begin, predict:

1. When you compute the means of the first 1, 10, 20, and 30 cards, how do you think will these means will compare to each other? Why?
2. When you compute the standard deviations of the first 1, 10, 20, and 30 cards, how do you think will these standard deviations will compare to each other? Why?

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30

	Mean	Std. Dev.
First 1		
First 10		
First 20		
All 30		

Module 2

Description:

In this activity, students will try to determine if a bag contains equal numbers of two different colors of bingo chips. Before beginning the experiment, each student must successfully complete the first 5 questions on the worksheet. This ensures that the students understand what hypothesis they are testing before they begin and forces them to use formal notation and language to describe the situation.

1. Each student then gets a small canvas bag containing 70 bingo chips of one color and 30 chips of another color
2. Each student samples 45 chips with replacement and uses this to compute a *p*-value and a 90% con.dence interval. Each student then enters their results into a spreadsheet.
3. As a class, examine a graph of the distribution and collectively decide whether the bags had equal numbers of the different chip colors; if they decide that the bags were not equally split, then they should try to estimate what the actual split was.
4. Reveal to the class that there was a 70/30 split
5. Students investigate the connection between this split and the confidence intervals.
6. Students discuss the *p*-values and the cases in which they rejected H_0.
7. Discuss the number of confidence intervals that captured the true proportion, how the true proportion can land anywhere in the interval, and the connection between the *p*-values and the confidence interval.

Worksheet

Activity 2 Worksheet

The One Proportion Test and Interval

Directions: Each student will be given a bag that contains two different colors of bingo chips (blue and purple). We will take a simple random sample of n = 45 chips (with replacement) to try and determine the if there is an equal amount of blue and purple chips.

- What is the population?
- What is the parameter of interest? You must include both the appropriate notation to represent the parameter as well as a clear written definition.
- What is the relevant statistic? Again provide notation and a definition.
- State H_0 and Ha for our test. Also, write a sentence explaining the meaning of both H_0 and Ha.
- Before gathering any data, do you expect H_0 to be false? Why?
- You may now obtain your sample. What are the values of X and $\hat{p}$?
- Find a 90% confidence interval.
- What is the *p*-value for the hypothesis test?
- What is your conclusion? Did this match your prediction? If it didn't, why do you think your prediction was off?
- What do you expect the class distribution of X, $\hat{p}$, the confidence intervals, and *p*-values will look like? Explain for each.
- Now put your data in the class database. Compare the actual results to your predictions.

Module 3

Description:

While students used a single bag of chips in the previous activity, here they will use two bags of chips to determine if the bags have identical

proportions. As before, each student must successfully complete the first 5 questions on the worksheet prior to starting the experiment.

- Each student gets two small canvas bags. The first bag contains 40 blue chips and 60 purple chips; the second contains 65 blue chips and 35 purple chips.
- Students sample 45 chips with replacement from each bag and use this to compute a *p*-value and a 90% confidence interval.
- Each student then enters their results into a spreadsheet.
- The class examines a graph of the distribution and collectively decides whether their bags had identical proportions; if they decide that the proportions were different, they should estimate what the difference was.
- Reveal that the actual proportion was 40/60 and 65/35.
- Students investigate the connection between this split and the confidence intervals.
- Discuss the *p*-values and the cases in which they rejected H_0.
- Discuss the number of confidence intervals that captured the true proportion, how the true proportion can land anywhere in the interval, and the connection between the *p*-values and the confidence interval.

Worksheet

Activity 3 Worksheet

The Two Proportion Test and Interval

Directions: Each student (or pair of students) will be given two bags (we will call bag two the bag with an X on the outside) that contains two different colors of bingo chips. We will take a simple random sample of $n = 45$ chips (with replacement) from each bag to try and determine if the ratio of blue chips is the same in each bag.

1. What is the population?
2. What are the parameters of interest? You must include both the appropriate notation to represent the parameter as well as a clear definition.
3. What are the relevant statistics? Again provide notation and a definition.
4. State H_0 and Ha for our test. Also, write a sentence explaining the meaning of both H_0 and H_a.
5. Before gathering any data, do you expect H_0 to be false? Why?
6. You may now obtain your sample (work in pairs with each person choosing a bag). What are the values of X_1, X_2, $\hat{p}_1$ and $\hat{p}_2$? Remember the Group 2 is the bag with an X on it.
7. Find a 90% confidence interval.
8. What is the *p*-value for the hypothesis test?
9. What is your conclusion? Did this match your prediction? If it didn't, why do you think your prediction was off?
10. What do you expect the class distribution of X, $\hat{p}$ the confidence intervals, and *p*-values will look like? Explain for each.
11. Now put your data in the class database. Compare the actual results to your predictions.

Module 4

Description:

In this activity, students will try to determine if the sum of the means of three dice is the mean of the sum of the dice. Before beginning the experiment, each student must successfully complete the first 5 questions on the worksheet. This ensures that the students understand what hypothesis they are testing before they begin and forces them to use formal notation and language to describe the situation.

1. Each student gets three dice: A six-sided, eight-sided, and 12-sided die.
2. Students roll the three dice together 30 times and record the sum and then use this to compute a *p*-value and a 90% confidence interval for the mean of the sum.
3. Each student then enters their results into a spreadsheet.
4. The class examines a graph of the distribution and collectively decides whether the sum of the means of three dice is the mean of the sum of the dice.
5. Discuss the *p*-values and the cases in which they rejected H_0.
6. Discuss the number of confidence intervals that captured the true mean, how the true mean can land anywhere in the interval, and the connection between the *p*-values and the confidence interval.

Worksheet

Activity 4 Worksheet

The T-test and T-Interval

Directions: Each student gets three dice: A six-sided, eight-sided, and 12-sided die. We want to decide if the mean of the sum of the three dice is the same as the mean we obtain by summing the mean for each of the die. We will take a simple random sample of n = 30 rolls of the three dice.

1. What is the mean μ for each of the three dice?
2. What is the population?
3. What are the parameters of interest? You must include both the appropriate notation to represent the parameter as well as a clear definition.
4. What are the relevant statistics? Again provide notation and a definition.

5. State H_0 and H_a for our test. Also, write a sentence explaining the meaning of both H_0 and H_a.
6. Before gathering any data, do you expect H_0 to be false? Why?
7. You may now obtain your sample. What are the values of $\overline{X}$ and s?
8. Find a 90% confidence interval for the mean of the sum.
9. What is the p-value for the hypothesis test?
10. What is your conclusion? Did this match your prediction? If it didn't, why do you think your prediction was off?
11. What do you expect the class distribution of X, $\hat{p}$, the confidence intervals, and p-values will look like? Explain for each.
12. Now put your data in the class database.

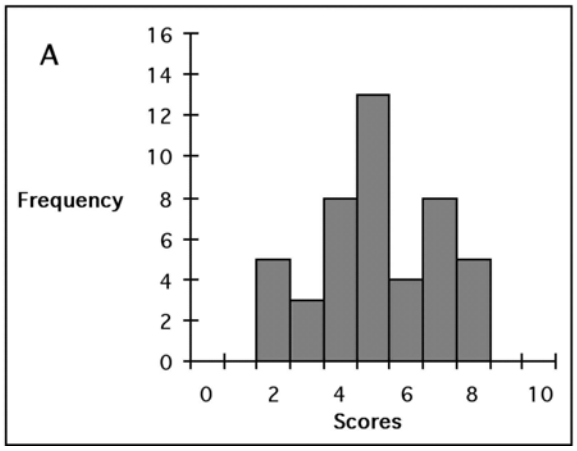
Appendix B: Assessments

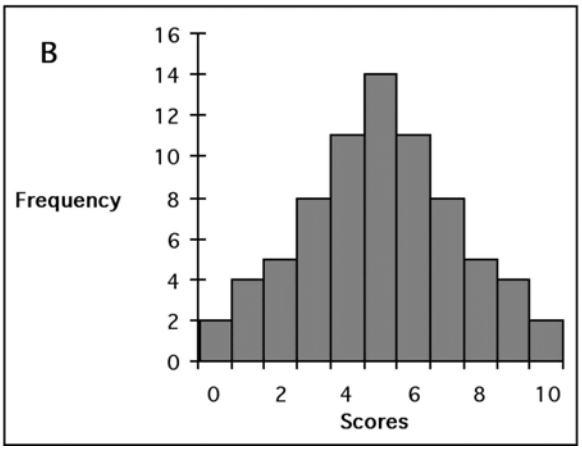
Assessment 1

Preliminary Survey

Name_____ Date _____

1. Which of the following distributions shows MORE variability?
A has more variability_____ B has more variability_____

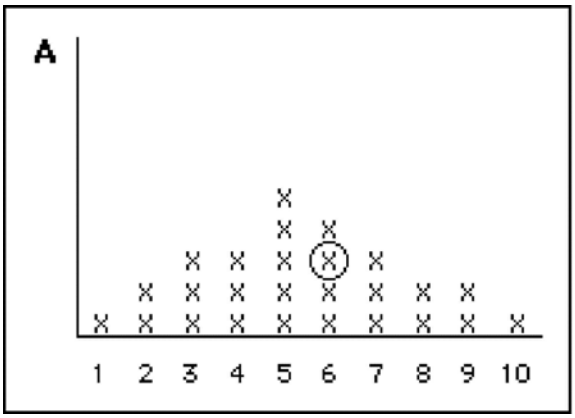


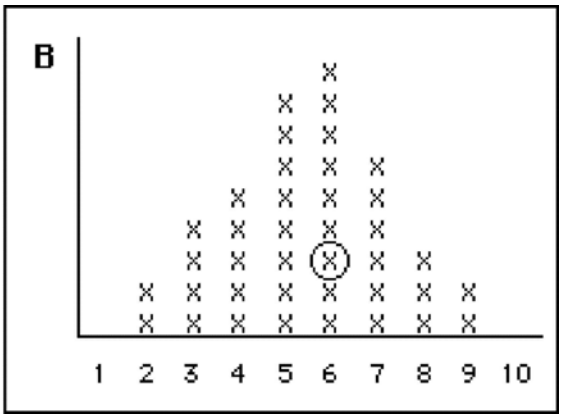


Circle the statement (or statements) that led you to select your answer above.

(a) Because it's bumpier
(b) Because it's more spread out
(c) Because it has a larger number of different scores
(d) Because the values differ more from the center
(e) Other (please explain)

2. Figure A represents a sample of 26 weights and Figure B represents a sampling distribution of mean weights for samples of size 3. One value is circled in each distribution.





Is there a difference between what is REPRESENTED by the X circled in A and the X circled in B? Explain why or why not .

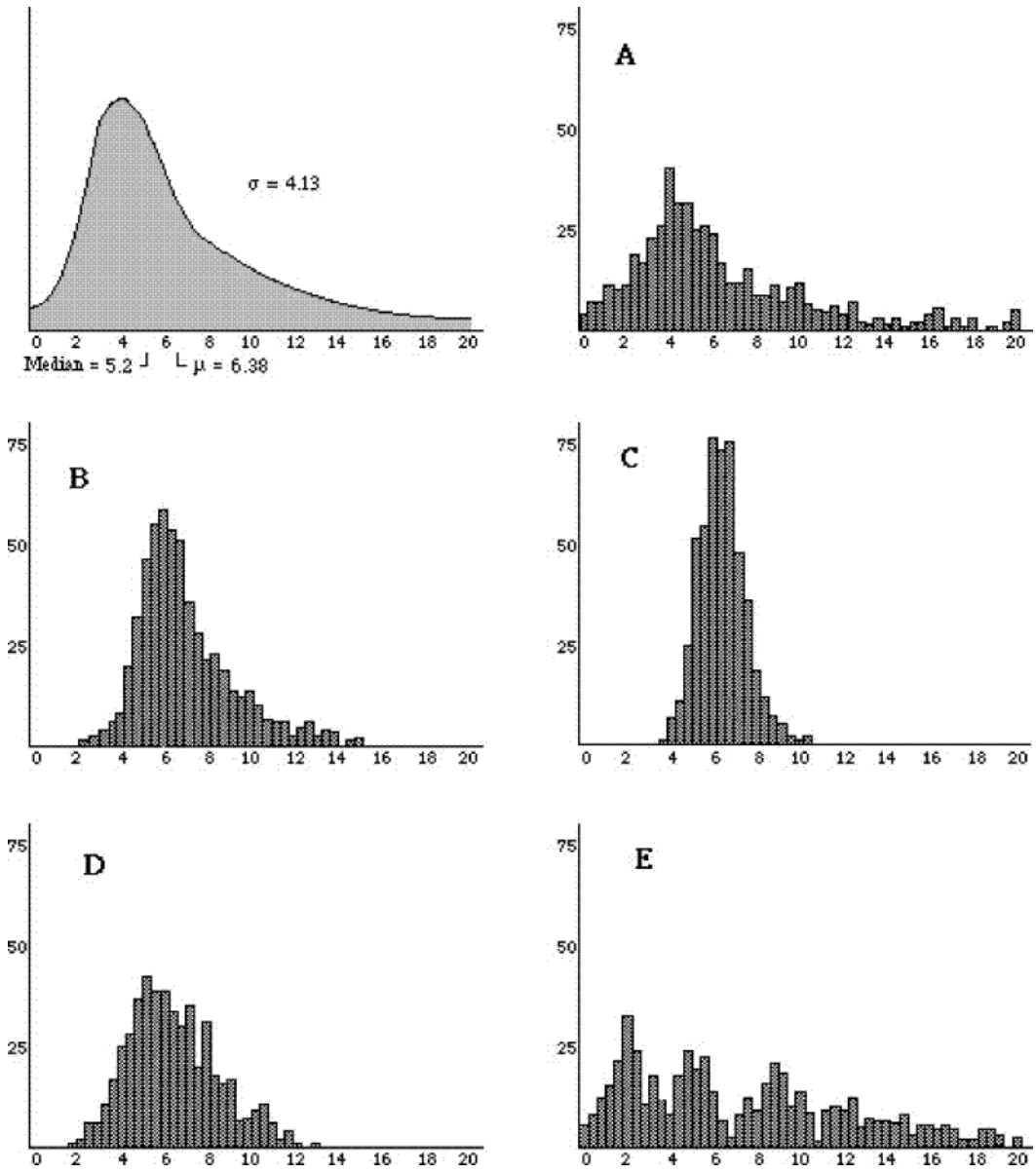
3. A sample of 50 data measurements is selected from a population of temperatures. A sample mean of 20 degrees is obtained. What would be your best estimate of μ, the population mean?

(a) It would be exactly 20 degrees
(b) It would be close to 20 degrees
(c) I wouldnt be able to make an estimate. I know nothing about μ. Its an unknown parameter and this is just one sample.
(d) Other:

4. For the quantities listed below, circle the ones that vary from sample to sample and explain why you chose these:

- Population standard deviation
- Sample standard deviation
- Population mean
- Sample mean

5. The distribution for a population of test scores is displayed below on the left. Each of the other five graphs labeled A to E represent possible distributions of sample means for random samples drawn from the population.



Please read each question carefully.

(a) Which graph represents a distribution of sample means for 500 samples of size 4?

(circleone) A B C D E

Answer each of the following questions regarding the sampling distribution you chose for the above question:

(b) What do you expect for the shape of the sampling distribution? (check only one)
___ D Shaped more like a NORMAL DISTRIBUTION
___ D Shaped more like the POPULATION
___ D Shaped like some OTHER DISTRIBUTION

(c) Circle the word between the two vertical lines that comes closest to completing the following sentence.

I expect the sampling distribution to have	less the same more	VARIABILITY than/as the population

Please explain your reasoning:

(d) Which graph do you think represents a distribution of sample means for 500 samples of size 16?
(circleone) A B C D E

Answer each of the following questions regarding the sampling distribution you chose for the above question

(e) What do you expect for the shape of the sampling distribution? (check only one)
___ D Shaped more like a NORMAL DISTRIBUTION
___ D Shaped more like the POPULATION D
___ Shaped like some OTHER DISTRIBUTION

Circle the word between the two vertical lines that comes closest to completing each of the following sentences.

(f)

I expect the sampling distribution to have	less the same more	VARIABILITY than/as the population

Please explain your reasoning:

(g)

I expect the sampling distribution I chose for the	less the same more	VARIABILITY than/as the sampling distribution I chose for the first question

Please explain your reasoning:

6. A 95% confidence interval indicates that:

a. 95% of the intervals constructed using this process based on samples from this population will include the population mean
b. 95% of the time the interval will include the sample mean
c. 95% of the possible population means will be included by the interval
d. 95% of the possible sample means will be included by the interval

7. Researchers ask a random sample of apartment dwellers in a large city their ideal air temperatures. They find the sample mean ($\overline{X}$) is 72 degrees. Using a two-tailed test, they reject H_0 : $\mu = 68$ at the 5% significance level. Which of the following could be a 95% confidence interval for μ, the average ideal temperature for all apartment dwellers in the city?

a. 70-73
b. 69-75
c. 68-76
d. 66-78
e. 66-70
f. Not enough information is given to answer this question

Two different pollsters, A and B, are trying to decide if a senators favorability ratings are above 50%. They each do their own random sample 1000 people in the senators state to perform a hypothesis test with.

8. Which of the following scenarios is most likely?

a. Both pollsters will get the same *p*-value.
b. Both pollsters will get *p*-values below 0.05.
c. Both pollsters will get *p*-values somewhat close to each other.
d. One pollster will get a value close to 1 while the other will get a *p*-value close to 0.
e. I'm not sure.

9. Assume the alternative hypothesis is true and that pollster A gets a *p*-value of 0.055 and that B gets a 0.032. The differences in the *p*-value is explained by

a. pollsters A did something wrong.
b. pollster B did something wrong.
c. one of the two pollsters did something wrong but we dont know which one.
d. neither pollster did anything wrong this is due to sampling variability.
e. I'm not sure.

10. Which of the following values will always be within the upper and lower limits of a confidence interval?

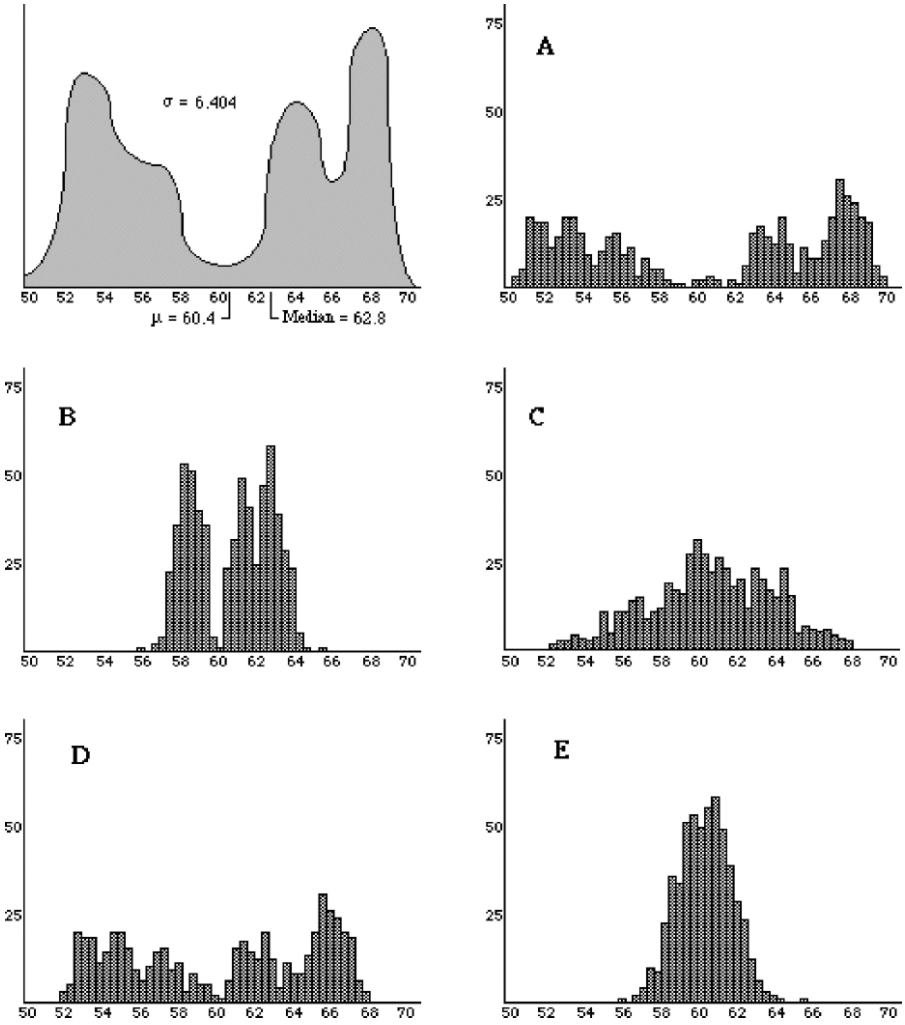
a. the population mean
b. the sample mean
c. the sample size
d. the standard deviation of the sample

Assessment 2

Assessment 2: Central Limit Theorem

Name_____ Date _____

The distribution for a population of test scores is displayed below on the left. Each of the other five graphs labeled A to E represent possible distributions of sample means for random samples drawn from the population.



1. Please read each question carefully.

(a) Which graph represents a distribution of sample means for 500 samples of size 4?
(circleone) A B C D E
Please explain your reasoning:

Answer each of the following questions regarding the sampling distribution you chose for the above question

(b) What do you expect for the shape of the sampling distribution? (check only one)
___ D Shaped more like a NORMAL DISTRIBUTION
___ D Shaped more like the POPULATION
___ D Shaped like some OTHER DISTRIBUTION
Please explain your reasoning:

(c) Circle the word between the two vertical lines that comes closest to completing the following sentence.

I expect the sampling distribution to have	less the same more	VARIABILITY than/as the population

Please explain your reasoning:

(d) Which graph do you think represents a distribution of sample means for 500 samples of size 16?
(circleone) A B C D E
Please explain your reasoning:

Answer each of the following questions regarding the sampling distribution you chose for the above question.

(e) What do you expect for the shape of the sampling distribution? (check only one)
___ D Shaped more like a NORMAL DISTRIBUTION
___ D Shaped more like the POPULATION
___ D Shaped like some OTHER DISTRIBUTION
Please explain your reasoning:

Circle the word between the two vertical lines that comes closest to completing each of the following sentences.

(f)

I expect the sampling distribution to have	less the same more	VARIABILITY than/as the population

Please explain your reasoning:

(g)

I expect the sampling distribution I chose for the	less the same more	VARIABILITY than/as the sampling distribution I chose for the first question

Please explain your reasoning:

2. Which of the following statements is NOT true according to the Central Limit Theorem? Select all that apply.
___ An increase in sample size from $n = 16$ to $n = 25$ will produce a sampling distribution with a smaller standard deviation.
___ The mean of a sampling distribution of sample means is equal to the population mean divided by the square root of the sample size.
___ The larger the sample size, the more the sampling distribution of sample means resembles the shape of the population.
___ The mean of the sampling distribution of sample means for samples of size $n = 15$ will be the same as the mean of the sampling distribution for samples of size $n = 100$.
___ The larger the sample size, the more the sampling distribution of sample means will resemble a normal distribution.

Explain your reasoning:

3. If sampling distributions of sample means are examined for samples of size 1, 5, 10, 16 and 50, you will notice that as n increases in size, the shape of the sampling distribution appears more like that of the:

a. normal distribution
b. population distribution
c. uniform distribution
d. even distribution

Explain your reasoning:

4. The amount of money college students spend each semester on textbooks is normally distributed with a mean of \$195 and a standard deviation of \$20. Suppose you take a random sample of 100 college students from this population. There would be a 68% chance that the sample mean ($\bar{X}$) amount spent on textbooks would be between:

a. \$191 and \$199.
b. \$193 and \$197.
c. \$175 and \$215.
d. \$155 and \$235.

Explain your reasoning:

Assessment 3

Assessment 3: Confidence Intervals and Hypothesis Tests

Name_____ Date _____

1. Two researchers are going to take a sample of data from the same population of chemistry students. Researcher A's sample will consist only of the students in her class. Researcher B will select a random sample of students from among all students taking chemistry. Both researchers will construct a 95% confidence interval for the mean score on the chemistry final exam using their own sample data. Which researcher's method has a 95% chance of capturing the true mean of the population of all students taking chemistry?

a. Both methods have a 95% chance of capturing the true mean
b. Researcher A
c. Researcher B
d. Neither

Please explain your reasoning:

2. A 95% confidence interval is calculated for a set of weights and the resulting confidence interval is 22 to 28 pounds. Indicate whether EACH of the following statements is True or False.

_____ 95% of the individual weights are between 22 and 28 pounds.
_____ Most of the individual weights are between 22 and 28 pounds.
_____ The probability that the interval includes the population mean (μ) is 95%.
_____ The probability that the interval includes the sample mean ($\bar{X}$) is 95%.
_____ If 200 confidence intervals were generated using the same process, about 10 of the confidence intervals would not include the population mean (μ).

Please explain your reasoning:

3. A 95% confidence interval indicates that:

a. 95% of the possible population means will be included by the interval
b. 95% of the intervals constructed using this process based on samples from this population will include the population mean
c. 95% of the possible sample means will be included by the interval
d. 95% of the time the interval will include the sample mean

Please explain your reasoning:

4. Which of the following is true?

a. It is possible to prove the null hypothesis under certain conditions.
b. It is always possible to prove the null hypothesis.
c. It is impossible to prove the null hypothesis.

Please explain your reasoning:

5. Researchers ask a random sample of apartment dwellers in a large city their ideal air temperatures. They find the sample mean ($\bar{X}$) is 72 degrees. Using a two-tailed test, they reject H_0 : μ = 68 at the 5% significance level. Which of the following could be a 95% confidence interval for μ, the average ideal temperature for all apartment dwellers in the city?

a. 70-73
b. 69-75
c. 68-76
d. 66-78
e. 66-70
f. Not enough information is given to answer this question

Please explain your reasoning:

The following situation is for problems 6 and 7: The average number of fruit candies in a large bag is estimated. The .95 confidence interval is [40-48].

6. Based on this information, you know that the best estimate of the population mean is

a. 40
b. 41
c. 42
d. 43
e. 44
f. 45

Please explain your reasoning:

7. Based on this information, you know that you can reject $H_0 : \mu = 38$ at $p =$

a. .85
b. .15
c. .10
d. .05
e. .01
f. .001

Please explain your reasoning:

8. Which of the following values will always be within the upper and lower limits of a confidence interval for the mean?

a. the population mean
b. the sample mean
c. the sample size
d. the standard deviation of the sample

Please explain your reasoning:

The following situation is used for problems 9 and 10: Two different pollsters, A and B, are trying to decide if a mayor's favorability ratings are above 50%. They each do their own random sample 1000 people in the senator's state to perform a hypothesis test with.

9. Which of the following scenarios is most likely?

a. Both pollsters will get the same *p*-value.
b. Both pollsters will get *p*-values below 0.05.
c. Both pollsters will get *p*-values somewhat close to each other.
d. One pollster will get a value close to 1 while the other will get a *p*-value close to 0.
e. I'm not sure.

Please explain your reasoning:

10. Assume the alternative hypothesis is true and that pollster A gets a *p*-value of 0.032 and that B gets a 0.055. The differences in the *p*-value is explained by

a. pollster A did something wrong.
b. pollster B did something wrong.
c. one of the two pollsters did something wrong but we don't know which one.
d. neither pollster did anything wrong this is due to sampling variability.
e. I'm not sure.

Please explain your reasoning:

Assessment 4

Assessment 4: Confidence Intervals and Hypothesis Tests

Name_____ Date _____

1. Two researchers are going to take a sample of data from the same population of physics students. Researcher A will select a random sample of students from among all students taking physics. Researcher B's sample will consist only of the students in her class. Both researchers will construct a 95% confidence interval for the proportion of scores on the physics final exam above 80% using their own sample data. Which researcher's method has a 95% chance of capturing the true mean of the population of all students taking physics?

a. Both methods have a 95% chance of capturing the true proportion
b. Researcher A
c. Research B
d. Neither

Please explain your reasoning:

2. A 95% confidence interval is calculated for a set of weights and the resulting confidence interval is 42 to 48 pounds. Indicate whether EACH of the following statements is True or False.

_____ 95% of the individual weights are between 42 and 48 pounds.
_____ Most of the individual weights are between 42 and 48 pounds
_____ The probability that the interval includes the population mean (μ) is 95%.
_____ The probability that the interval includes the sample mean ($\overline{X}$) is 95%.
_____ If 200 confidence intervals were generated using the same process, about 10 of the confidence intervals would not include the population mean (μ).

Please explain your reasoning:

3. A 95% confidence interval indicates that:

a. 95% of the intervals constructed using this process based on samples from this population will include the population proportion
b. 95% of the time the interval will include the sample proportion
c. 95% of the possible population proportions will be included by the interval
d. 95% of the possible sample proportions will be included by the interval

Please explain your reasoning:

4. Which of the following is true?

a. It is impossible to prove the null hypothesis.
b. It is always possible to prove the null hypothesis.
c. It is possible to prove the null hypothesis under certain conditions.

Please explain your reasoning:

5. Researchers ask a random sample of apartment dwellers in a large city their ideal air temperatures. They find the sample proportion ($\hat{p}$) of apartment dwellers that prefer the temperature above 70 degrees is 0.65. Using a two-tailed test, they reject $H_0 : p = 0.50$ at the 5% significance level. Which of the following could be a 95% confidence interval for p, the average ideal temperature for all apartment dwellers in the city?

a. 0.55 -0.7
b. 0.6 -0.7
c. 0.5 -0.8
d. 0.45 -0.85
e. 0.4 -0.6
f. Not enough information is given to answer this question

Please explain your reasoning:

The following situation is for problems 6 and 7: The proportion of red fruit candies in a large bag is estimated. The .95 confidence interval is [0.30-0.50].

6. Based on this information, you know that the best estimate of the population proportion is

a. 0.30
b. 0.35
c. 0.40
d. 0.45
e. 0.50
f. 0.55

Please explain your reasoning:

7. Based on this information, you know that you can reject $H_0 : p = .025$ at a p-value =

a. .85
b. .15
c. .10
d. .05
e. .01
f. .001

Please explain your reasoning:

8. Which of the following values will always be within the upper and lower limits of a confidence interval?

a. the population proportion
b. the sample proportion
c. the sample size
d. the standard deviation of the sample

Please explain your reasoning:

The following situation is used for problems 9 and 10: Two different pollsters, A and B, are trying to decide if a senator's favorability ratings are above 33%. They each do their own random sample 1000 people in the senator's state to perform a hypothesis test with.

9. Which of the following scenarios is most likely?

a. Both pollsters will get the same *p*-value.
b. Both pollsters will get *p*-values below 0.05.
c. Both pollsters will get *p*-values somewhat close to each other.
d. One pollster will get a value close to 1 while the other will get a *p*-value close to 0.
e. I'm not sure.

Please explain your reasoning:

10. Assume the alternative hypothesis is true and that pollster A gets a *p*-value of 0.055 and that B gets a 0.032. The differences in the *p*-value is explained by

a. pollster A did something wrong.
b. pollster B did something wrong.
c. one of the two pollsters did something wrong but we don't know which one.
d. neither pollster did anything wrong this is due to sampling variability.
e. I'm not sure.

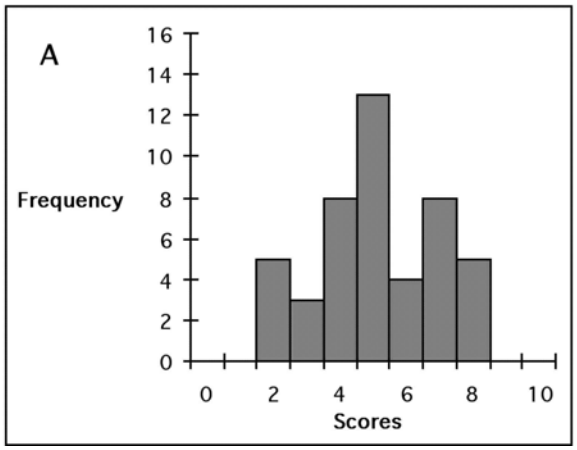
Please explain your reasoning:

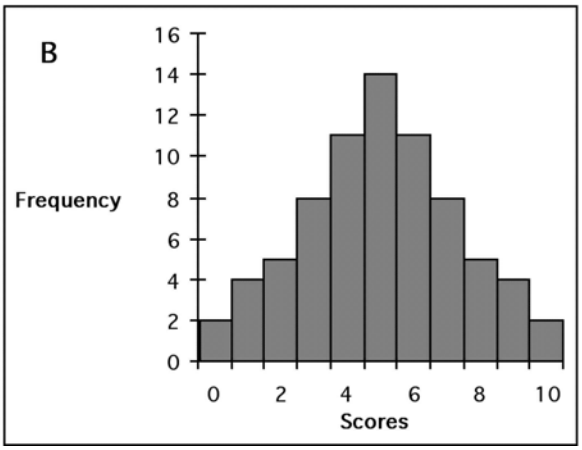
Assessment 5

Assessment 5: Summary

Name_____ Date _____

1. Which of the following distributions shows MORE variability?
A has more variability _____ B has more variability_____



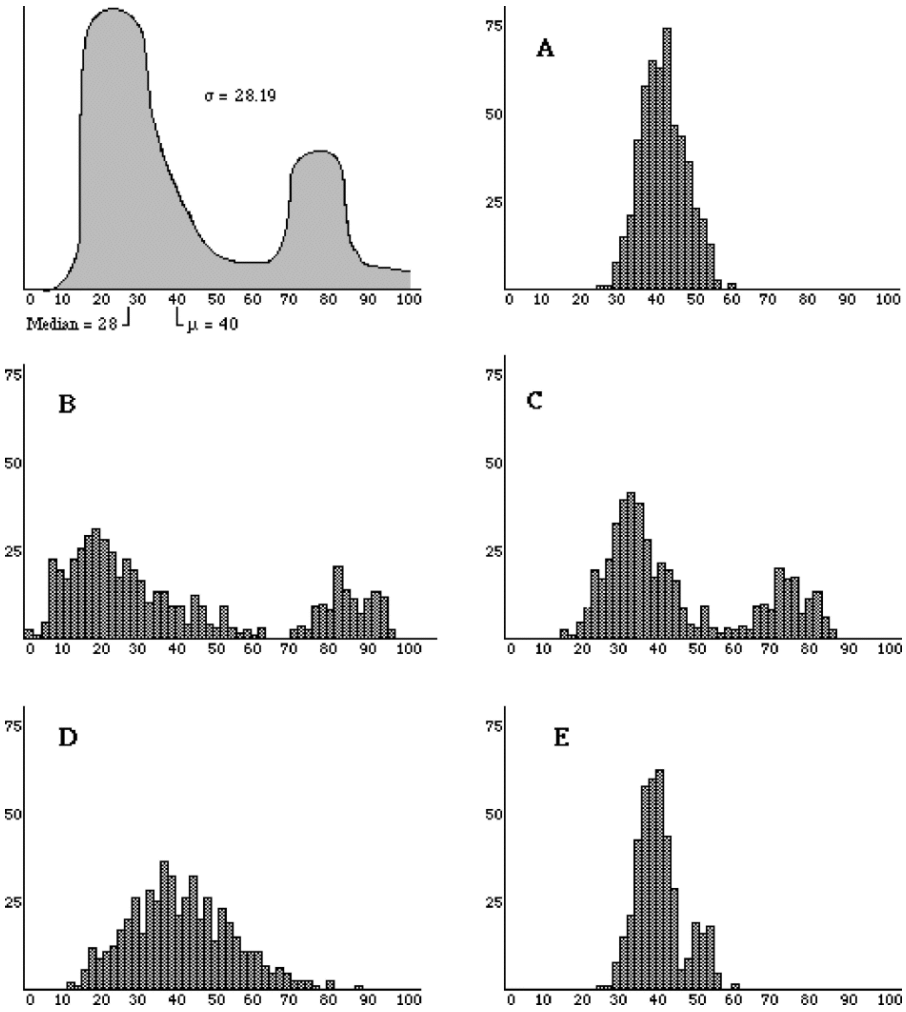


Circle the statement (or statements) that led you to select your answer above.
(a) Because it's bumpier
(b) Because it's more spread out
(c) Because it has a larger number of different scores
(d) Because the values differ more from the center
(e) Other (please explain)

2. Which of the following is true?

a. It is impossible to prove the null hypothesis.
b. It is always possible to prove the null hypothesis.
c. It is possible to prove the null hypothesis under certain conditions.

Please explain your reasoning:



3. Please read each question carefully.

(a) Which graph represents a distribution of sample means for 500 samples of size 4?
(circleone) A B C D E

Please explain your reasoning:

Answer each of the following questions regarding the sampling distribution you chose for the above question

(b) What do you expect for the shape of the sampling distribution? (check only one)
_____ D Shaped more like a NORMAL DISTRIBUTION
_____ D Shaped more like the POPULATION
_____ D Shaped like some OTHER DISTRIBUTION

Please explain your reasoning:

(c) Circle the word between the two vertical lines that comes closest to completing the following sentence.

I expect the sampling distribution to have	less the same more	VARIABILITY than/as the population

Please explain your reasoning:

(d) Which graph do you think represents a distribution of sample means for 500 samples of size 16?
(circleone) A B C D E

Please explain your reasoning:

Answer each of the following questions regarding the sampling distribution you chose for the above question

(e) What do you expect for the shape of the sampling distribution? (check only one)
_____ Shaped more like a NORMAL DISTRIBUTION
_____ Shaped more like the POPULATION
_____ Shaped like some OTHER DISTRIBUTION

Please explain your reasoning:

Circle the word between the two vertical lines that comes closest to completing each of the following sentences.
(f)

I expect the sampling distribution to have	less the same more	VARIABILITY than/as the population

Please explain your reasoning:

(g)

I expect the sampling distribution I chose for the	less the same more	VARIABILITY than/as the sampling distribution I chose for the first question

Please explain your reasoning:

The following situation is for problems 4 and 5: Two different pollsters, A and B, are trying to decide if a governor's favorability ratings are above 60%. They each do their own random sample 1000 people in the senators state to perform a hypothesis test with.

4. Which of the following scenarios is most likely?

a. Both pollsters will get the same *p*-value.
b. Both pollsters will get *p*-values below 0.05.
c. Both pollsters will get *p*-values somewhat close to each other.
d. One pollster will get a value close to 1 while the other will get a *p*-value close to 0.
e. I'm not sure.

Please explain your reasoning:

5. Assume the alternative hypothesis is true and that pollster A gets a *p*-value of 0.055 and that B gets a 0.032. The differences in the *p*-value is explained by

a. pollster A did something wrong.
b. pollster B did something wrong.
c. one of the two pollsters did something wrong but we dont know which one.
d. neither pollster did anything wrong this is due to sampling variability.
e. I'm not sure.

Please explain your reasoning:

6. Which of the following values will always be within the upper and lower limits of a confidence interval?

a. the standard deviation of the sample
b. the population mean
c. the sample size
d. the sample mean

Please explain your reasoning:

7. Which of the following values will always be within the upper and lower limits of a confidence interval?

a. the population mean
b. the sample mean
c. the sample size
d. the standard deviation of the sample

Please explain your reasoning:

8. A 95% confidence interval indicates that:

a. 95% of the intervals constructed using this process based on samples from this population will include the population proportion
b. 95% of the time the interval will include the sample proportion
c. 95% of the possible population proportions will be included by the interval
d. 95% of the possible sample proportions will be included by the interval

Please explain your reasoning:

9. Two researchers are going to take a sample of data from the same population of biology students. Researcher A's sample will consist only of the students in her class. Researcher B will select a random sample of students from among all students taking biology. Both researchers will construct a 95% confidence interval for the mean score on the biology final exam using their own sample data. Which researcher's method has a 95% chance of capturing the true mean of the population of all students taking biology?

a. Researcher A
b. Researcher B
c. Both methods have a 95% chance of capturing the true mean
d. Neither

Please explain your reasoning:

10. A 95% confidence interval is calculated for a set of weights and the resulting confidence interval is 52 to 58 pounds. Indicate whether EACH of the following statements is True or False.
_____ 95% of the individual weights are between 52 and 58 pounds.
_____ Most of the individual weights are between 52 and 58 pounds.
_____ The probability that the interval includes the population mean (μ) is 95%.
_____ The probability that the interval includes the sample mean ($\overline{X}$) is 95%.
_____ If 200 confidence intervals were generated using the same process, about 10 of the confidence intervals would not include the population mean (μ).

Please explain your reasoning:

Appendix C: Statistical Results

NOTE: All *p*-values are for a one-side McNemar's test unless stated otherwise. Items are denoted by the assessment on which they appeared followed by a dash and the item number on that particular assessment. For example, question 1 on Assessment 1 is denoted A1-1.

A1-1 to A5-1

This problem looks at two histograms and asks which has more variability and why. Twenty of the 27 (74%) students got the question of which graph has more variability correct on the pre-test and 22 (81%) did on Assessment 5, with two missing responses. All students who got the

correct answer on A1 also responded correctly on A5. Of the seven who incorrectly answered the question on A1, four got the correct answer on A5. A one-sided McNemar test gives a *p*-value of 0.0625.

	A1-1 to A5-1
wrong to right	4/25 (16%)
right to wrong	0
wrong to wrong	3/25 (12%)
right to right	18/25 (72%)
McNemar's p-vlaue	0.0625

Table 4: A1-1 to A5-1

On the why question there where two correct response. On the pretest 11 of the 27 (41%) correctly identified one of the two responses and one student correctly identified both. On Assessment 5, 17(63%) correctly identified one of the correct answers and again only one identified both, with one missing response. Of the 11 correct responses, three gave incorrect responses on A5; of the 16 incorrect response 9 gave a correct response on A5. A one-sided McNemar test gives a *p*-value of 0.073.

	A1-1w to A5-1w
wrong to right	9/26 (35%)
right to wrong	3/26 (12%)
wrong to wrong	5/26 (19%)
right to right	9/26 (35%)
McNemar's p-vlaue	0.073

Table 5: A1-1 to A5-1 Explanation

A1-3 to A3-6 to A4-6

This set of problems involve estimating a parameter based on sample information. Problem A1-3 is somewhat different than the other two in that it just gives the results of a sample whereas the other two provide the reader with just a confidence interval.

	A1-3 to A3-6	A1-3 to A4-6	A3-6 to A4-6
wrong to right	9/27 (33%)	10/27 (37%)	3/27 (11%)
right to wrong	2/27 (7%)	0	0
wrong to wrong	1/27 (4%)	0	0
right to right	15/27 (56%)	17/27 (63%)	24/27 (89%)
McNemar's p-vlaue	0.0325	0.001	0.125

Table 6: A1-3 to A3-6 to A4-6

A1-5 to A2-1 to A5-3 a-g

This set of problems pertains to understanding of the central limit theorem. All problems are exactly the same except for a different graph. Partial credit was converted to full credit for parts (a) and (d).

	A1-5a to A2-1a	A1-5a to A5-3a	A2-1a to A5-3a
wrong to right	2/26 (8%)	0	7/25 (28%)
right to wrong	13/26 (50%)	13/26 (50%)	9/25 (36%)
wrong to wrong	2/26 (8%)	3/26 (12%)	7/25 (28%)
right to right	9/26 (35%)	10/26 (38%)	2/25 (8%)
McNemar's p-vlaue	$1 - 0.0035$	$1 - 0.000$	$1 - 0.402$

Table 7: A1-5 to A2-1 to A5-3 a

	A1-5b to A2-1b	A1-5b to A5-3b	A2-1b to A5-3b
n	25	25	25
wrong to right	7 (28%)	11 (44%)	8 (32%)
right to wrong	4 (16%)	5 (20%)	4 (16%)
wrong to wrong	9 (36%)	5 (20%)	5 (20%)
right to right	5 (20%)	4 (16%)	8 (32%)
McNemar's p-vlaue	0.275	0.105	0.388

Table 8: A1-5 to A2-1 to A5-3 b

	A1-5c to A2-1c	A1-5c to A5-3c	A2-1c to A5-3c
n	25	25	25
wrong to right	8 (32%)	8 (32%)	5 (20%)
right to wrong	4 (16%)	2 (8%)	5 (20%)
wrong to wrong	11 (44%)	14 (56%)	11 (44%)
right to right	2 (8%)	3 (12%)	4 (16%)
McNemar's p-vlaue	0.194	0.145	0.5

Table 9: A1-5 to A2-1 to A5-3 c

NEED TABLE 10

Table 10: A1-5 to A2-1 to A5-3 d

Table 10: A1-5 to A2-1 to A5-3 d

	A1-5d to A2-1d	A1-5d to A5-3d	A2-1d to A5-3d
n	24	25	24
wrong to right	3 (13%)	2 (8%)	2 (8%)
right to wrong	9 (38%)	9 (36%)	3 (13%)
wrong to wrong	11 (46%)	13 (52%)	19 (79%)
right to right	1 (4%)	1 (4%)	0
McNemar's p-vlaue	$1 - 0.073$	$1 - 0.033$	0.5

Table 11: A1-5 to A2-1 to A5-3 d

NEED TABLE 12

Table 12: A1-5 to A2-1 to A5-3 e

	A1-5e to A2-1e	A1-5e to A5-3e	A2-1e to A5-3e
n	23	25	23
wrong to right	9 (39%)	5 (20%)	3 (13%)
right to wrong	6 (26%)	5 (20%)	7 (30%)
wrong to wrong	2 (9%)	8 (32%)	4 (17%)
right to right	6 (4%)	7 (28%)	9 (39%)
McNemar's p-vlaue	0.304	0.5	$1 - 0.172$

Table 13: A1-5 to A2-1 to A5-3 e

	A1-5f to A2-1f	A1-5f to A5-3f	A2-1f to A5-3f
n	24	26	23
wrong to right	7 (29%)	8 (31%)	7 (30%)
right to wrong	3 (13%)	2 (8%)	4 (17%)
wrong to wrong	13 (54%)	14 (54%)	9 (39%)
right to right	1 (4%)	2 (8%)	3 (13%)
McNemar's p-vlaue	0.172	0.055	0.275

Table 14: A1-5 to A2-1 to A5-3 f

	A1-5g to A2-1g	A1-5g to A5-3g	A2-1g to A5-3g
n	25	27	25
wrong to right	10 (40%)	8 (30%)	3 (12%)
right to wrong	4 (16%)	5 (19%)	6 (24%)
wrong to wrong	4 (16%)	8 (30%)	5 (20%)
right to right	7 (28%)	6 (22%)	11 (44%)
McNemar's p-vlaue	0.090	0.291	$1 - 0.254$

Table 15: A1-5 to A2-1 to A5-3 g

A1-6 to A3-3 to A4-3 to A5-8

This problem evaluates the students understanding of the meaning of a 95% confidence interval. All four problems are exactly the same.For the purpose of this analysis partial credit is coded as correct.

	A1-6:A3-3	A1-6:A4-3	A1-6:A5-8	A3-3:A4-3	A3-3:A5-8	A4-3:A5-8
n	27	27	26	27	26	26
wrong to right	8 (30%)	8 (30%)	4 (15%)	5 (19%)	3 (12%)	3 (12%)
right to wrong	5 (19%)	6 (22%)	8 (31%)	6 (22%)	9 (35%)	8 (31%)
wrong to wrong	7 (26%)	7 (26%)	10 (38%)	7 (26%)	9 (35%)	10 (38%)
right to right	7 (26%)	6 (22%)	4 (15%)	9 (33%)	5 (19%)	5 (19%)
McNemar's p-vlaue	0.291	0.396	$1 - 0.194$	0.500	$1 - .073$	$1 - 0.114$

Table 16: A1-6 to A3-3 to A4-3 to A5-8

A1-7 to A3-5 to A4-5

This problem relates hypothesis testing to confidence intervals. All three problems are very similar. For the purpose of this analysis partial credit is coded as correct.

	A1-7 to A3-5	A1-7 to A4-5	A3-5 to A4-5
n	27	26	26
wrong to right	3 (11%)	10 (38%)	8 (31%)
right to wrong	4 (15%)	6 (23%)	4 (15%)
wrong to wrong	10 (37%)	3 (12%)	5 (19%)
right to right	10 (37%)	7 (27%)	9 (35%)
McNemar's p-vlaue	0.500	0.227	0.194

Table 17: A1-7 to A3-5 to A4-5

A1-8 to A3-9 to A4-9 to A5-4

This problem involves interpreting the *p*-values from two samples from the same population. All four problems are very similar.

	A1-8:A3-9	A1-8:A4-9	A1-8:A5-4	A3-9:A4-9	A3-9:A5-4	A4-9:A5-4
n	27	27	26	27	26	26
wrong to right	4 (15%)	6 (22%)	8 (31%)	4 (15%)	4 (15%)	4 (15%)
right to wrong	3 (11%)	3 (11%)	3 (12%)	2 (7%)	1 (4%)	3 (12%)
wrong to wrong	6 (22%)	4 (15%)	2 (8%)	5 (19%)	4 (15%)	2 (8%)
right to right	14 (52%)	14 (52%)	13 (50%)	16 (59%)	17 (65%)	17 (65%)
McNemar's p-vlaue	0.500	0.254	0.114	0.344	0.188	0.500

Table 18: A1-8 to A3-9 to A4-9 to A5-4

A1-9 to A3-10 to A4-10 to A5-5

This problem involves interpreting the *p*-values from two samples from the same population. All four problems are very similar.

	A1-9:A3-10	A1-9:A4-10	A1-9:A5-5	A3-10:A4-10	A3-10:A5-5	A4-10:A5-5
n	27	26	26	26	26	25
wrong to right	5 (19%)	5 (19%)	6 (23%)	2 (8%)	3 (12%)	5 (20%)
right to wrong	3 (11%)	5 (19%)	2 (8%)	4 (15%)	1 (4%)	1 (4%)
wrong to wrong	4 (15%)	3 (12%)	3 (12%)	4 (15%)	4 (15%)	3 (12%)
right to right	15 (56%)	13 (50%)	15 (58%)	16 (62%)	18 (70%)	16 (64%)
McNemar's p-vlaue	0.364	0.500	0.145	$1 - .0344$	0.313	0.110

Table 19: A1-9 to A3-10 to A4-10 to A5-5

A1-10 to A3-8 to A4-8 to A5-7

These questions ask what is always in a confidence interval. Allfour problems are exactly the same. We used A5-7 instead of A5-6 because its answer order is the same as the other three questions whereas A5-6 changes the order.

	A1-10:A3-8	A1-10:A4-8	A1-10:A5-7	A3-8:A4-8	A3-8:A5-7	A4-8:A5-7
n	26	26	25	27	26	26
wrong to right	5 (19%)	7 (27%)	12 (48%)	5 (19%)	7 (27%)	5 (19%)
right to wrong	3 (11%)	3 (11%)	5 (20%)	3 (11%)	2 (8%)	2 (8%)
wrong to wrong	14 (54%)	12 (46%)	6 (24%)	13 (48%)	10 (39%)	10 (39%)
right to right	4 (15%)	4 (15%)	2 (8%)	6 (22%)	7 (27%)	9 (35%)
McNemar's p-vlaue	0.364	0.172	0.072	0.364	0.090	0.227

Table 20: A1-10 to A3-8 to A4-8 to A5-7

A3-1 to A4-1 to A5-9

This problem relates to understanding bias and random samples. The three problems are essentially the same.

	A3-1 to A4-1	A3-1 to A4-1	A4-1 to A5-9
n	27	26	26
wrong to right	1 (4%)	2 (8%)	1 (4%)
right to wrong	0 (0%)	2 (8%)	2 (8%)
wrong to wrong	1 (4%)	0 (0%)	0 (0%)
right to right	25 (93%)	22 (85%)	23 (88%)
McNemar's p-vlaue	0.500	0.500	0.500

Table 21: A3-1 to A4-1 to A5-9

A3-2 to A4-2 to A5-10 a-e

This is a set of true/false questions checking the understanding of confidence intervals.

	A3-2a to A4-2a	A3-2a to A5-10a	A4-2a to A5-10a
n	26	26	25
wrong to right	1 (4%)	1 (4%)	3 (12%)
right to wrong	3 (12%)	3 (12%)	3 (12%)
wrong to wrong	11 (42%)	11 (42%)	10 (40%)
right to right	11 (42%)	11 (42%)	9 (36%)
McNemar's p-vlaue	$1 - 0.313$	$1 - 0.313$	0.500

Table 22: A3-2 to A4-2 to A5-10 a

	A3-2b to A4-2b	A3-2b to A5-10b	A4-2b to A5-10b
n	25	25	23
wrong to right	6 (24%)	5 (20%)	3 (13%)
right to wrong	0 (0%)	1 (4%)	6 (26%)
wrong to wrong	15 (6%)	16 (64%)	10 (44%)
right to right	4 (16%)	3 (12%)	4 (17%)
McNemar's p-vlaue	0.016	0.110	$1 - 0.254$

Table 23: A3-2 to A4-2 to A5-10 b

	A3-2c to A4-2c	A3-2c to A5-10c	A4-2c to A5-10c
n	27	25	25
wrong to right	2 (7%)	2 (8%)	5 (20%)
right to wrong	6 (22%)	7 (28%)	5 (20%)
wrong to wrong	5 (19%)	4 (16%)	6 (24%)
right to right	14 (52%)	12 (48%)	9 (36%)
McNemar's p-vlaue	$1 - 0.145$	$1 - 0.090$	0.500

Table 24: A3-2 to A4-2 to A5-10 c

	A3-2d to A4-2d	A3-2d to A5-10d	A4-2d to A5-10d
n	26	24	23
wrong to right	5 (19%)	5 (21%)	2 (9%)
right to wrong	3 (12%)	5 (21%)	5 (22%)
wrong to wrong	9 (35%)	9 (38%)	9 (39%)
right to right	9 (35%)	5 (21%)	7 (30%)
McNemar's p-vlaue	0.364	0.500	$1 - 0.227$

Table 25: A3-2 to A4-2 to A5-10 d

	A3-2e to A4-2e	A3-2e to A5-10e	A4-2e to A5-10e
n	26	24	23
wrong to right	6 (23%)	4 (17%)	2 (9%)
right to wrong	2 (8%)	3 (13%)	4 (17%)
wrong to wrong	1 (4%)	2 (8%)	1 (4%)
right to right	17 (65%)	15 (63%)	16 (70%)
McNemar's p-vlaue	0.145	0.500	$1 - 0.344$

Table 26: A3-2 to A4-2 to A5-10 e

A3-4 to A4-4 to A5-2

This question asks if it is possible to prove the null hypothesis. All three questions are identical.

	A3-4 to A4-4	A3-4 to A5-2	A4-4 to A5-2
n	26	27	26
wrong to right	1 (4%)	3 (11%)	1 (4%)
right to wrong	1 (4%)	2 (7%)	1 (4%)
wrong to wrong	21 (81%)	20 (74%)	21 (81%)
right to right	3 (12%)	2 (7%)	3 (12%)
McNemar's p-vlaue	0.500	0.500	0.500

Table 27: A3-4 to A4-4 to A5-2

A3-7 to A4-7

This questions gives a confidence interval and asks if a particular null hypothesis can be rejected. Both questions are similar.

	A3-7 to A4-7
n	27
wrong to right	6 (22%)
right to wrong	4 (15%)
wrong to wrong	11 (41%)
right to right	6 (22%)
McNemar's p-vlaue	0.377

Table 28: A3-7 to A4-7

A3 grade to A4 grade (10 questions)

Comparing Assessment 3 to Assessment 4 including partial credit given for explanations using a paired *t*-test yields a one-sided *p*-value of 0.059, with means of 11.4 and 12.0 respectively.

Comparing Assessment 3 to 4 without counting credit given for explanations (just the multiple choice problems) yields a one sided *p*-value of 0.007, with means of 7.7 and 8.4 respectively.

A1 (6,7,8,9,10) grade to A3 & A4 (3,5,8,9,10) grade

Comparing questions 6, 7, 8, 9, and 10 on Assessment 1 with questions 3, 5, 8, 9, 10 on Assessment 3 (all questions similar)yields a one sided *p*-value of 0.093, with means of 2.3 and 2.6 respectively. Comparing questions 6, 7, 8, 9, and 10 on Assessment 1 with questions 3, 5, 8, 9, 10 on Assessment 4 (all questions similar)yields a one sided *p*-value of 0.037, with means of 2.3 and 2.9 respectively.

A1 (1,5,6,8,9,10) grade to A5 (1,3,4,5,7,8) grade (13 questions)

We get a one sided *p*-value of 0.019 with means of 5.2 and 6.3, respectively.

A3/A4 (1,2,3,4,8,9,10) grade to A5 (2,4,5,7,8,9,10) grade (11 questions)

Comparing question 1, 2, 3, 4, 8, 9, 10 on Assessment 3 withquestions 2, 4, 5, 7, 8, 9, 10 on Assessment 5 yields a *p*-value of 1 - 0.395 with means of 6.02 and 5.92, respectively. Comparing question 1, 2, 3, 4, 8, 9, 10 on Assessment 4 withquestions 2, 4, 5, 7, 8, 9, 10 on Assessment 5 yields a *p*-value of 1 - 0.151 with means of 6.35 and 5.92, respectively.

Descriptive Information on A2 2,3,4

On A2-2, 7 out of 26 students (27%) received at least partial credit with only 1 receiving full credit. On A2-3, 9 out of 26 students were correct (35%) while 21 out of 26 (81%) were correct on A2-4.

Association Between Same-Assessment Items

Note: Students' responses were coded into two categories for each item: The response of interest and the "other" category, which included all other responses. The number-letter pairs refer to the question number and the response of interest (e.g. 6a/c refers to question 6, choices a and c). All *p*-values are for a .2 test resulting form a 2×2 contingency table.

- Assessment 1
  – 6a/c and 10a: p = .918
  – 6b/d and 10b: p = .116
  – 7a/b/c/d and 10b: p = .143
  – 8c and 9d: p = .159
- Assessment 3
  – 2c and 8a: p = .334
  – 2d and 8b: p = .247
  – 2d and 5a/b/c/d: p = .432
  – 2e and 3b: p = .756
  – 2e and 3b/d: p = .055
  – 2e and 8a: p = .791
  – 3a/b and 8a: p = .150
  – 3c/d and 8b: p = .611
  – 5b/c/d and 6e: p = .057
  – 5a/b and 7d: p = .816
  – 5a/b and 7a/b/c/d: p = .081
  – 5a/b/c/d and 8b: p = .148
  – 9c and 10d: p = .534

- Assessment 4
  – 2c and 8a: p = .946
  – 2d and 8b: p = .191
  – 2d and 5a/b/c/d: p = .208
  – 2e and 3a: p = .449
  – 2e and 3a/b: p = .960
  – 2e and 8a: p = .220
  – 3a/c and 8a: p = .732
  – 3b/d and 8b: p = .135
  – 5b/c/d and 6c: p =N/A (two cells of size zero)
  – 5a/b and 7d: p = .100
  – 5a/b and 7a/b/c/d: p = .778
  – 5a/b/c/d and 8b: p = .683
  – 9c and 10d: p = .535
- Assessment 5
  – 6b and 8a/c: p = .664
  – 6b and 10c: p = .633
  – 6b and 10e: p = .772
  – 6d and 8b/d: p = .691
  – 8a and 10e: p = .477
  – 8a/b and 10e: p = .897

Endnotes

[1]Problems that received partial credit of 0.5 are Assessment 1 problem 5a answer d (A1-5a d), A1-5d c, A1-6 d, A1-7 a, A2-1a a, A3-3 c, A3-5 a, A4-3 d, A4-5 a, and A5-3a b, while problems A3-4a, A4-4 c, and A5-2 c received a 0.25 partial credit. In addition, on A1-1 and A5-1 full credit was only given to students who answered b and d, while partial credit was given to those who answered either b or d (but not both).

[2]The sampling distributions questions were problem 5 on Assessment 1, problem 1 on Assessment 2, and problem 3 on Assessment 5.

[3]delMas, Garfield, and Chance (1999) describe "good reasoning" as "When a student chose a histogram for the larger sample size that was shaped like a normal distribution and that had less variability than the histogram chosen for the smaller sample size." They describe "larger to smaller reasoning" as when "students chose a histogram with less variability for the larger sample size."

[4]The results of the statistical tests can be found in Appendix C.

References

Bradley, D. R., Hemstreet, R. L., and Ziegenhagen, S. T. (1992), "A Simulation Laboratory for Statistics," Behavior Research Methods, Instruments, and Computers, 24, 190-204.

Bradstreet, 1996. T.E. Bradstreet, Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician* 50 (1996), pp. 69–78.

Chance, B. L. (1997), "Experiences with Authentic Assessment Techniques in an Introductory Statistics Course," Journal of Statistics Education, [Online], 5(3). (www.amstat.org/publications/jse/v5n3/chance.html)

Dambolena, I. G. (1986), "Using Simulation in Statistics Courses," Collegiate Microcomputer, 4, 339344.

delMas, R. C., Garfield, J., & Chance, B. (1999), "A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning," Journal of Statistics Education 7(3). (www.amstat.org/publications/jse/secure/v7n3/delmas.cfm)

Dietz, E. J. (1993), "A Cooperative Learning Activity on Methods of Selecting a Sample," The American Statistician, 47, 104-108.

Fillebrown, S. (1994), "Using Projects in an Elementary Statistics Course for Non-Science Majors," Journal of Statistics Education, [Online], 2 (2). (www.amstat.org/publications/jse/v2n2/.llebrown.html)

Giesbrecht, N. (1996), "Strategies for Developing and Delivering effective Introductory-Level Statistics and Methodology Courses," ERIC Document Reproduction Service, No. 393-668, Alberta, BC.

Gnanadesikan, M., Scheaffer, R., Watkins, A. & Witmer, J. (1997) "An Activity-Based Statistics Course," Journal of Statistics Education, [Online], 5(2). (www.amstat.org/publications/jse/v5n2/gnanadesikan.html)

Goodman, T. A. (1986), "Using the Microcomputer to Teach Statistics," Mathematics Teacher, 79, 210-215.

Gordon, F. (1987), "Computer Graphics Simulation of the Central Limit Theorem," Mathematics and Computer Education, 2, 48-55.

Gordon, F. S., and Gordon, S. P. (1989), "Computer Graphics Simulations of Sampling Distributions," Collegiate Microcomputer, 7, 185-189.

Gratz, Z. S., Volpe, G. D., and Kind, B. M. (1993), "Attitudes and Achievement in Introductory Psychological Statistics Classes: Traditional versus Computer Supported Instruction," ERIC Document Reproduction Service No. 365 405, Ellenville, NY.

Halley, F. S. (1991), "Teaching Social Statistics with Simulated Data," Teaching Sociology, 19, 518-525.

Hesterberg, T. C. (1998), "Simulation and Bootstrapping for Teaching Statistics," American Statistical Association Proceedings of the Section on Statistical Education, Alexandria, VA: American Statistical Association, 44-52.

Hodgson, T. R. (1996), "The Effects of Hands-On Activities on Students' Understanding of Selected Statistical Concepts," in *Proceedings of the Eighteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, eds. E. Jakubowski, D. Watkins, and H. Biske, Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education, pp. 241-246.

Hogg, R.V. (1991), "Statistical Education: Improvements are Badly Needed," The American Statistician, 45, 342-343.

Hubbard, R. (1992), "Teaching Statistics With MINITAB," Australian Mathematics Teacher, 48, 8-10.

Hunter, W.G. (1977), "Some Ideas about Teaching Design of Experiments, with 25 Examples of Experiments Conducted by Students," The American Statistician, 31, 12-20.

Karley, L. M. (1990), "Using Computer Graphics in Statistics," Mathematics and Computer Education, 24, 232-239.

Ledolter, J. (1995), "Projects in Introductory Statistics Courses," The American Statistician, 49, 364 367.

Lunsford, M. L., Rowell, G. H., & Goodson-Espy, T. (2006), "Classroom Research: Assessment of Student Understanding of Sampling Distributions of Means and the Central Limit Theorem in Post-Calculus Probability and Statistics Classes," Journal of Statistics Education, [Online], 14(3). (www.amstat.org/publications/jse/v14n3/lunsford.html)

Mackisack, M. (1994), "What is the Use of Experiments Conducted by Statistics Students?" Journal of Statistics Education [Online], 2(1). (http://www.amstat.org/publications/jse/v2n1/mackisack.html)

Marasinghe, M. G., Meeker, W. Q., Cook, D., and Shin, T. (1996), "Using Graphics and Simulation to Teach Statistical Concepts," The American Statistician, 50, 342-351.

Maxwell, N. (1994), "A Coin-Flipping Exercise to Introduce the *p*-value," Journal of Statistics Education, [Online], 2(1). (www.amstat.org/publications/jse/v2n1/maxwell.html)

McBride, A. B. (1996), "Creating a Critical Thinking Learning Environment: Teaching Statistics to Social Science Undergraduates," Political Science and Politics, 29, 517-521.

Mills, J. D. (2002), "Using Computer Simulation Methods to Teach Statistics: A Review of the Literature," Journal of Statistics Education, [Online], 10(1). (www.amstat.org/publications/jse/v10n1/mills.html)

Mittag, K. C. (1992), "Using Computers to Teach the Concepts of the Central Limit Theorem," ERIC Document Reproduction Service No. 349 947, San Francisco, CA.

National Council for Teachers of Mathematics, (2000), Principles and standards for school mathematics, Reston, VA: Author.

National Council for Teachers of Mathematics, (2006), Curriculum focal points for prekindergarten through grade 8 mathematics:A quest for coherence, Reston, VA: Author.

Packard, A. L., Holmes, G. A., and Fortune, J. C. (1993), "A Comparison of Three Presentation Methods of Teaching Statistics," ERIC Document Reproduction Service No. 365 696, Chicago, IL.

Pulley, L. B., and Dolbear, F. T. (1984), "Computer Simulation Exercises for Economics Statistics," Journal of Economics Education, 3, 77-87.

Schwartz, D. L., Goldman, S. R., Vye N. J., Barron, B. J., and The Cognition Technology Group at Vanderbilt (1997), "Aligning Everyday and Mathematical Reasoning: The Case of Sampling Assumptions," in *Reflections on Statistics: Agendas for Learning, Teaching and Assessment in K-12*, ed. S. Lajoie, Hillsdale, NJ: Erlbaum.

Snee, R. D.(1993), "What's Missing in Statistical Education?", The American Statistician, 47, 149-154.

Sullivan, M. M. (1993), "Students Learn Statistics When They Assume a Statistician's Role," ERIC Document Reproduction Service No. 368 547, Boston, MA.

Velleman, P. F., and Moore, D. S. (1996), "Multimedia for Teaching Statistics: Promises and Pitfalls," The American Statistician, 50, 217-225.

Von Glaserfeld, E. (1987), "Learning as a Constructive Activity," in Problems of Representation in the Teaching and Learning of Mathematics, Hillsdale, NJ: Lawrence Erlbaum Associates, 3-17.

Wood , M. (2005), "The Role of Simulation Approaches in Statistics," Journal of Statistics Education, [Online], 13(3). (www.amstat.org/publications/jse/v13n3/wood.html)

Thomas J. Pfaff
Ithaca College
953 Danby Rd.
Ithaca, NY, 14850
tpfaff@ithaca.edu
(607) 274-7066

Aaron Weinberg
Ithaca College
953 Danby Rd.
Ithaca, NY, 14850
aweinberg@ithaca.edu
(607) 2747081
