GUIDELINES FOR ASSESSMENT AND INSTRUCTION
IN STATISTICS EDUCATION

COLLEGE REPORT 2016

2016 report: ROBERT CARVER • MICHELLE EVERSON (CHAIR) •
JOHN GABROSEK • NICHOLAS HORTON • ROBIN LOCK •
MEGAN MOCKO • ALLAN ROSSMAN • GINGER ROWELL •
PAUL VELLEMAN • JEFF WITMER • BEVERLY WOOD

2005 report: MARTHA ALIAGA • GEORGE COBB • CAROLYN CUFF •
JOAN GARFIELD (CHAIR) • ROB GOULD • ROBIN LOCK •
TOM MOORE • ALLAN ROSSMAN • BOB STEPHENSON •
JESSICA UTTS • PAUL VELLEMAN • JEFF WITMER

DRAFT: NOVEMBER 1, 2015

PUBLISHED BY THE AMERICAN STATISTICAL ASSOCIATION

TUFTE-LATEX.GOOGLECODE.COM

Initial draft of second edition, August 2015

Contents

Executive Summary

The American Statistical Association (ASA) funded the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project, which consists of two groups, one focused on K–12 education and one focused on introductory college courses. This report presents the recommendations developed by the college group.

The report includes a brief history of the introductory college course and summarizes the 1992 report[1] by George Cobb that has been considered a generally accepted set of recommendations for teaching these courses. Results of a survey on the teaching of introductory courses are summarized, along with a description of current versions of introductory statistics courses. We then offer a list of goals for students, based on what it means to be statistically literate. We present six recommendations for the teaching of introductory statistics that build on the previous recommendations from Cobb's report. Our six recommendations include the following:

1. Emphasize statistical literacy and develop statistical thinking
2. Use real data
3. Stress conceptual understanding, rather than mere knowledge of procedures
4. Foster active learning in the classroom
5. Use technology for developing conceptual understanding and analyzing data
6. Use assessments to improve and evaluate student learning

The report concludes with suggestions for how to make these changes and includes numerous examples in the appendices to illustrate details of the recommendations.

[1] George Cobb. *Heeding the Call for Change: Suggestions for Curricular Action (MAA Notes No. 22)*, chapter Teaching Statistics, pages 3–43. The Mathematical Association of America, Washington DC, 1992

Introduction

The GAISE project was funded by a member initiative grant from the ASA in 2003 to develop ASA-endorsed guidelines for assessment and instruction in statistics in the K–12 curriculum and for the introductory college statistics course.

Our work on the college course guidelines included many discussions over email and in-person small group meetings. Our discussions began by reviewing existing standards and guidelines, relevant research results from the studies of teaching and learning statistics, and recent discussions and recommendations regarding the need to focus instruction and assessment on the important concepts that underlie statistical reasoning.

History and Growth of the Introductory Course

The modern introductory statistics course has roots that go back a long way, to early books about statistical methods. R. A. Fisher's *Statistical Methods for Research Workers*, which first appeared in 1925, was aimed at practicing scientists. A dozen years later, the first edition of George Snedecor's *Statistical Methods* presented an expanded version of the same content, but there was a shift in audience. More than Fisher's book, Snedecor's became a textbook used in courses for prospective scientists who were still completing their degrees; statistics was beginning to establish itself as an academic subject, albeit with heavy practical, almost vocational emphasis. By 1961, with the publication of *Probability with Statistical Applications* by Fred Mosteller, Robert Rourke, and George Thomas, statistics had begun to make its way into the broader academic curriculum, but here again, there was a catch: In these early years, statistics had to lean heavily on probability for its legitimacy.

During the late 1960s and early 1970s, John Tukey's ideas of exploratory data analysis brought a near-revolutionary pair of changes to the curriculum: freeing certain kinds of data analysis from ties to probability-based models so that the analysis of data could begin to acquire status as an independent intellectual activity and introducing

a collection of "quick-and-dirty" data tools so students could analyze data without having to spend hours chained to a bulky mechanical calculator. Computers would later complete the "data revolution" in the beginning statistics curriculum, but Tukey's ideas of exploratory data analysis (EDA) provided both the first technical breakthrough and the new ethos that avoided invented examples.

Two influential books appeared in 1978: *Statistics*, by David Freedman, Robert Pisani, and Roger Purves, and *Statistics: Concepts and Controversies*, by David S. Moore. The publication of these two books marked the birth of what we regard as the modern introductory statistics course.

The evolution of content has been paralleled by other trends. One of these is a striking and sustained growth in enrollments. Two sets of statistics suffice here:

- At two-year colleges, according to the Conference Board of the Mathematical Sciences, statistics enrollments have grown from 27% of the size of calculus enrollments in 1970 to 74% of the size of calculus enrollments in 2000.
- The Advanced Placement exam in statistics was first offered in 1997. There were 7,500 students who took it that first year, more than in the first offering of an AP exam in any subject at that time. The next year, more than 15,000 students took the exam. The next year, more than 25,000, and the next, 35,000. In 2004, more than 65,000 students took the AP statistics exam.

Both the changes in course content and the dramatic growth in enrollment are implicated in a third set of changes, a process of democratization that has broadened and diversified the backgrounds, interests, and motivations of those who take the courses. Statistics has gone from being a course taught from a book like Snedecor's, for a narrow group of future scientists in agriculture and biology, to being a family of courses, taught to students at many levels, from pre-high school to post-baccalaureate, with very diverse interests and goals. A teacher in the 1940s, using Snedecor's *Statistical Methods*, could assume that most students were both quantitatively skilled and adequately motivated by their career plans. A teacher of today's beginning statistics courses works with a different group of students. Most take statistics earlier in their lives, increasingly often in high school; few are drawn to statistics by immediate practical need; and there is great variety in their levels of quantitative sophistication. As a result, today's teachers face challenges of motivation and exposition that are substantially greater than those of a half century ago.

Not only have the "what, why, who, and when" of introductory statistics been changing, but so has the "how." The last few decades

have seen an extraordinary level of activity focused on how students learn statistics, and on how we teachers can be more effective in helping them learn.

The 1992 Cobb Report

In the spring of 1991, George Cobb, in order to highlight important issues to the mathematics community, coordinated an email focus group on statistics education as part of the Curriculum Action Project of the Mathematics Association of America (MAA). The report was published in the MAA volume *Heeding the Call for Change*[2]. It included the following recommendations:

[2] George Cobb. *Heeding the Call for Change: Suggestions for Curricular Action (MAA Notes No. 22)*, chapter Teaching Statistics, pages 3–43. The Mathematical Association of America, Washington DC, 1992

Emphasize Statistical Thinking

Any introductory course should take as its main goal helping students to learn the basic elements of statistical thinking. Many advanced courses would be improved by a more explicit emphasis on those same basic elements, namely:

- *The need for data*. Recognizing the need to base personal decisions on evidence (data) and the dangers inherent in acting on assumptions not supported by evidence.
- *The importance of data production*. Recognizing that it is difficult and time-consuming to formulate problems and to get data of good quality that really deal with the right questions. Most people don't seem to realize this until they go through this experience themselves.
- *The omnipresence of variability*. Recognizing that variability is ubiquitous. It is the essence of statistics as a discipline and not best understood by lecture. It must be experienced.
- *The quantification and explanation of variability*. Recognizing that variability can been measured and explained, taking into consideration the following: (a) randomness and distributions; (b) patterns and deviations (fit and residual); (c) mathematical models for patterns; (d) model-data dialogue (diagnostics).

More Data and Concepts, Less Theory and Fewer Recipes

Almost any course in statistics can be improved by more emphasis on data and concepts, at the expense of less theory and fewer recipes. To the maximum extent feasible, calculations and graphics should be automated.

Foster Active Learning

As a rule, teachers of statistics should rely much less on lecturing and much more on alternatives such as projects, lab exercises, and group problem-solving and discussion activities. Even within the traditional lecture setting, it is possible to get students more actively involved.

The three recommendations were intended to apply quite broadly (e.g., whether or not a course has a calculus prerequisite and regardless of the extent to which students are expected to learn specific statistical methods). Although the work of the focus group ended with the completion of their report, many members of the group continued to work on these issues, especially on efforts at dissemination and implementation, as members of the joint ASA/MAA Committee on Undergraduate Statistics.

Current Status of the Introductory Statistics Course

OVER THE DECADE THAT FOLLOWED the publication of the Cobb report, many changes were implemented in the teaching of statistics. In recent years, many statisticians have become involved in the reform movement in statistical education aimed at the teaching of introductory statistics, and the National Science Foundation has funded numerous projects designed to implement aspects of this reform[3]. Moore[4] describes the reform in terms of changes in content (more data analysis, less probability), pedagogy (fewer lectures, more active learning), and technology (for data analysis and simulations).

In 1998 and 1999, Garfield[5] surveyed a large number of statistics instructors from mathematics and statistics departments and a smaller number of statistics instructors from departments of psychology, sociology, business, and economics to determine how the introductory course is being taught and to begin to explore the impact of the educational reform movement.

The results of this survey suggested that major changes were being made in the introductory course, that the primary area of change was in the use of technology, and that the results of course revisions generally were positive, although they required more time of the course instructor. Results were surprisingly similar across departments, with the main differences found in the increased use of graphing calculators, active learning and alternative assessment methods in courses taught in math departments in two-year colleges, the increased use of web resources by instructors in statistics departments, and the reasons cited for why changes were made (more math instructors were influenced by recommendations from statistics education). The results were also consistent in reporting that more changes were to be

[3] George Cobb. Reconsidering statistics education: A national science foundation conference. *Journal of Statistics Education*, 1(1), 1993. URL http://www.amstat.org/publications/jse/v1n1/cobb.html

[4] David Moore. New pedagogy and new content: The case of statistics. *International Statistical Review*, 65:123–165, 1997

[5] Joan Garfield. Evaluating the statistics education reform. Final report to the national science foundation, 2000. URL http://education.umn.edu/EdPsych/Projects/Impact.html

made, particularly as more technological resources became available.

Today's introductory statistics course is actually a family of courses taught across many disciplines and departments. The students enrolled in these courses have different backgrounds (e.g., in mathematics, psychology) and goals (e.g., some hope to do their own statistical analyses in research projects, some are fulfilling a general quantitative reasoning requirement).

As in the past, some of these courses are taught in large classes and some are taught in small classes (or even freshman seminars). Some students are taught statistics in computer labs, some students take the course using only a simple calculator, and some take the course via distance learning without ever seeing their classmates or instructor in person. Some classes are taught over a 10-week quarter and some are taught over a 15-week semester. Each of these classes might range from three to six hours per week.

Today's goals for students tend to focus more on conceptual understanding and attainment of statistical literacy and thinking, and less on learning a set of tools and procedures. While demands for dealing with data in an information age continue to grow, advances in technology and software make tools and procedures easier to use and more accessible to more people, thus decreasing the need to teach the mechanics of procedures, but increasing the importance of giving more people a sounder grasp of the fundamental concepts needed to use and interpret those tools intelligently. These new goals, described in the following section, reinforce the need to reexamine and revise many introductory statistics courses to help achieve the important learning goals for students.

Goals for Students in an Introductory Course: What it Means to be Statistically Educated

Some people teach courses that are heavily slanted toward teaching students to become statistically literate and wise consumers of data; this is somewhat similar to an art appreciation course. Some teach courses more heavily slanted toward teaching students to become producers of statistical analyses; this is analogous to the studio art course. Most courses are a blend of consumer and producer components, but the balance of that mix will determine the importance of each recommendation we present.

The desired result of all introductory statistics courses is to produce statistically educated students, which means that students should develop statistical literacy and the ability to think statistically. The following goals represent what such a student should know and understand. Achieving this knowledge will require learning some statistical techniques, but the specific techniques are not as important as the knowledge that comes from going through the process of learning them. Therefore, we are not recommending specific topical coverage.

Students should believe and understand why:

- Data beat anecdotes
- Variability is natural, predictable, and quantifiable
- Random sampling allows results of surveys and experiments to be extended to the population from which the sample was taken
- Random assignment in comparative experiments allows cause-and-effect conclusions to be drawn
- Association is not causation
- Statistical significance does not necessarily imply practical importance, especially for studies with large sample sizes

- Finding no statistically significant difference or relationship does not necessarily mean there is no difference or no relationship in the population, especially for studies with small sample sizes

Students should recognize:

- Common sources of bias in surveys and experiments
- How to determine the population to which the results of statistical inference can be extended, if any, based on how the data were collected
- How to determine when a cause-and-effect inference can be drawn from an association based on how the data were collected (e.g., the design of the study)
- That words such as “normal," “random,” and “correlation” have specific meanings in statistics that may differ from common usage

Students should understand the parts of the process through which statistics works to answer questions, namely:

- How to obtain or generate data
- How to graph the data as a first step in analyzing data, and how to know when that’s enough to answer the question of interest
- How to interpret numerical summaries and graphical displays of data—both to answer questions and to check conditions (to use statistical procedures correctly)
- How to make appropriate use of statistical inference
- How to communicate the results of a statistical analysis

Students should understand the basic ideas of statistical inference, including:

- The concept of a sampling distribution and how it applies to making statistical inferences based on samples of data (including the idea of standard error)
- The concept of statistical significance, including significance levels and p-values
- The concept of confidence interval, including the interpretation of confidence level and margin of error

Finally, students should know:

- How to interpret statistical results in context
- How to critique news stories and journal articles that include statistical information, including identifying what's missing in the presentation and the flaws in the studies or methods used to generate the information
- When to call for help from a statistician

Recommendations

We endorse the ideas in the three original goals found in the Cobb report[6] and have expanded them in light of today's situation. The intent of these recommendations is to help students attain the list of learning goals described in the previous section.

[6] George Cobb. *Heeding the Call for Change: Suggestions for Curricular Action (MAA Notes No. 22)*, chapter Teaching Statistics, pages 3–43. The Mathematical Association of America, Washington DC, 1992

Recommendation 1: Emphasize statistical literacy and develop statistical thinking.

We define statistical literacy as understanding the basic language of statistics (e.g., knowing what statistical terms and symbols mean and being able to read statistical graphs) and fundamental ideas of statistics. For readings on statistical literacy, see Gal[7], Rumsey[8], and Utts[9].

Statistical thinking has been defined as the type of thinking that statisticians use when approaching or solving statistical problems. Statistical thinking has been described as understanding the need for data, the importance of data production, the omnipresence of variability, and the quantification and explanation of variability[10]. We provide illustrations of statistical thinking in the following example and analogy.

[7] Iddo Gal. Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70:1–51, 2002

[8] D. J. Rumsey. Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, 10(3), 2002. URL http://www.amstat.org/publications/jse/v10n3/rumsey2.html

[9] Jessica Utts. What educated citizens should know about statistics and probability? *The American Statistician*, 57 (2):74–79, 2003

[10] George Cobb. *Heeding the Call for Change: Suggestions for Curricular Action (MAA Notes No. 22)*, chapter Teaching Statistics, pages 3–43. The Mathematical Association of America, Washington DC, 1992

The Funnel Example

Think of a funnel that is wide at the top, corresponding to a great many situations, and narrow at the bottom, corresponding to a few specialized cases. Statisticians are practical problem-solvers. When a client presents a problem (e.g., Is there a treatment effect present?), the statistician tries to provide a practical answer that addresses the problem efficiently. Quite often, a simple graph is sufficient to tell the story. Perhaps a more detailed plot will answer the question at hand. If not, then some calculations may be needed. A simple test based on a gross simplification of the situation may confirm that a treatment effect is present. If simplifying the situation is troublesome, then a more refined test may be used, capturing more of the specifics of the

modeling situation at hand. Different statisticians may come up with somewhat different analyses of a given set of data, but will usually agree on the main conclusions and only worry about minor points if those points matter to the client. If there is no standard procedure to answer the question, then and only then will the statistician use first principles to develop a new tool. *We should model this type of thinking for our students, rather than showing them a set of skills and procedures and giving them the impression that, in any given situation, there is one best procedure to use and only that procedure is acceptable.*

The Carpentry Analogy

In week 1 of the carpentry (statistics) course, we learned to use various kinds of planes (summary statistics). In week 2, we learned to use different kinds of saws (graphs). Then, we learned about using hammers (confidence intervals). Later, we learned about the characteristics of different types of wood (tests). By the end of the course, we had covered many aspects of carpentry (statistics). But I wanted to learn how to build a table (collect and analyze data to answer a question) and I never learned how to do that. *We should teach students that the practical operation of statistics is to collect and analyze data to answer questions.*

SUGGESTIONS FOR TEACHERS:

- ✓ Model statistical thinking for students, working examples and explaining the questions and processes involved in solving statistical problems from conception to conclusion.
- ✓ Use technology and show students how to use technology effectively to manage data, explore data, perform inference, and check conditions that underlie inference procedures.
- ✓ Give students practice developing and using statistical thinking. This should include open-ended problems and projects.
- ✓ Give students plenty of practice with choosing appropriate questions and techniques, rather than telling them which technique to use and merely having them implement it.
- ✓ Assess and give feedback on students' statistical thinking.

In the appendices, we present examples of projects, activities, and assessment instruments that can be used to develop and evaluate statistical thinking.

Recommendation 2: Use real data.

IT IS IMPORTANT TO USE REAL DATA in teaching statistics to be authentic to consider issues related to how and why the data were produced or collected, and to relate the analysis to the problem context. Using real data sets of interest to students is also a good way to engage them in thinking about the data and relevant statistical concepts. There are many types of real data, including archival data, classroom-generated data, and simulated data. Sometimes, hypothetical data sets may be used to illustrate a particular point (e.g., The Anscombe data illustrates how four data sets can have the same correlation but strikingly different scatterplots.) or to assess a specific concept. It is important to only use created or realistic data for this specific purpose and not for general data analysis and exploration. An important aspect of dealing with real data is helping students learn to formulate good questions and use data to answer them appropriately based on how the data were produced.

SUGGESTIONS FOR TEACHERS:

- ✓ Search for good, raw data to use from web data repositories, textbooks, software packages, and surveys or activities in class. If there is an opportunity, seek out real data directly from a practicing research scientist (through a journal or at one's home institution). Using such data can enliven your class and increase the store of good data sets for other teachers by communicating the newly found data to others. Search for and use summaries based on real data, from data summary websites, journal articles, websites with surveys and polls, and textbooks.
- ✓ Use data to answer questions relevant to the context and generate new questions.
- ✓ Make sure questions used with data sets are of interest to students—if no one cares about the questions, it's not a good data set for the introductory class. (Example: physical measurements on species no one has heard of.) *Note: Few data sets interest all students, so one should use data from a variety of contexts.*
- ✓ Use class-generated data to formulate statistical questions and plan uses for the data before developing the questionnaire and collecting the data. (Example: Ask questions likely to produce different shaped histograms, use interesting categorical variables to investigate relationships.) It is important that data gathered from students in class not contain information that

could be embarrassing to students and that students' privacy is maintained.

- ✓ Get students to practice entering raw data using a small data set or a subset of data, rather than spending time entering a large data set. Make larger data sets available electronically.
- ✓ Use subsets of variables in different parts of the course, but integrate the same data sets throughout. (Example: Do side-by-side boxplots to compare two groups, then do two-sample t-tests on the same data. Use histograms to investigate shape, then to verify conditions for hypothesis tests.)

The appendices include examples of good ways (and not-so-good ways) to use data in homework, projects, tests, etc.

Recommendation 3: Stress conceptual understanding, rather than mere knowledge of procedures.

MANY INTRODUCTORY COURSES contain too much material, and students end up with a collection of ideas that are understood only at surface level, are not well-integrated, and are quickly forgotten. If students don't understand the important concepts, there's little value in knowing a set of procedures. If they understand the concepts well, then particular procedures will be easy to learn. In the student's mind, procedural steps too often claim attention that an effective teacher could otherwise direct toward concepts.

Recognize that giving more attention to concepts than to procedures may be difficult politically, both with students and client disciplines. However, students with a good conceptual foundation from an introductory course are well-prepared to study additional statistical techniques such as research methods, regression, experimental design, or statistical methods in a second course.

SUGGESTIONS FOR TEACHERS:

- ✓ View the primary goal as not to cover methods, but to discover concepts.
- ✓ Focus on students' understanding of key concepts, illustrated by a few techniques, rather than covering a multitude of techniques with minimal focus on underlying ideas.
- ✓ Pare down content of an introductory course to focus on core concepts in more depth. *Examples of syllabi focused on concepts, compared to a syllabus focused on a list of topics, are in the appendices.*

Perform routine computations using technology to allow greater emphasis on interpretation of results. Although the language of mathematics provides compact expression of key ideas, use formulas that enhance the understanding of concepts, and avoid computations that are divorced from understanding. For example, $s = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}}$ helps students understand the role of standard deviation as a measure of spread and to see the impact of individual y values on s, whereas $s = \sqrt{\frac{\Sigma y^2 - \frac{1}{n}(\Sigma y)^2}{n-1}}$ has no redeeming pedagogical value.

Recommendation 4: Foster active learning in the classroom.

USING ACTIVE LEARNING METHODS IN CLASS is a valuable way to promote collaborative learning, allowing students to learn from each other. Active learning allows students to discover, construct, and understand important statistical ideas and to model statistical thinking. Activities have an added benefit in that they often engage students in learning and make the learning process fun. Other benefits of active learning methods are the practice students get communicating in the statistical language and learning to work in teams. Activities offer the teacher an informal method of assessing student learning and provide feedback to the instructor on how well students are learning. It is important that teachers not underestimate the ability of activities to teach the material or overestimate the value of lectures, which is why suggestions are provided for incorporating activities, even in large lecture classes.

TYPES OF ACTIVE LEARNING INCLUDE:

- Group or individual problem-solving, activities, and discussion
- Lab activities (physical and computer-based)
- Demonstrations based on data generated on the spot from the students

SUGGESTIONS FOR TEACHERS:

- ✓ Ground activities in the context of real problems. Therefore, data should be collected to answer a question, not "collect data to collect data" (without a question).
- ✓ Mix lectures with activities, discussions, and labs.
- ✓ Precede computer simulations with physical explorations (e.g., die rolling, card shuffling).

- ✓ Collect data from students (anonymously).
- ✓ Encourage predictions from students about the results of a study that provides the data for an activity before analyzing the data. This motivates the need for statistical methods. (If all results were predictable, we wouldn't need either data or statistics.)
- ✓ Do not use activities that lead students step by step through a list of procedures, but allow students to discuss and think about the data and the problem.
- ✓ Plan ahead to make sure there is enough time to explain the problem, let the students work through the problem, and wrap up the activity during the same class. It is hard to complete the activity in the next class period. Make sure there is time for recap and debriefing, even if at the beginning of the next class period.
- ✓ Provide a lot of feedback to students on their performance and learning.
- ✓ Include assessment as an important component of an activity.

SUGGESTIONS FOR IMPLEMENTING ACTIVE LEARNING IN LARGE CLASSES:

- ✓ Take advantage of large classes providing opportunities for large sample sizes for student-generated data.
- ✓ In large classes, it may be easier to have students work in pairs, rather than in larger groups.
- ✓ Use a separate lab/discussion section for activities, if possible.

Recommendation 5: Use technology for developing concepts and analyzing data.

TECHNOLOGY HAS CHANGED the way statisticians work and should change what and how we teach. For example, statistical tables such as a normal probability table are no longer needed to find p-values, and we can implement computer-intensive methods. We think technology should be used to analyze data, allowing students to focus on interpretation of results and testing of conditions, rather than on computational mechanics.Technology tools should also be used to help students visualize concepts and develop an understanding of abstract ideas by simulations. Some tools offer both types of uses,

See the Appendices for an example illustrating technology uses.

while, in other cases, a statistical software package may be supplemented by web applets. Regardless of the tools used, it is important to view the use of technology not just as a way to compute numbers but as a way to explore conceptual ideas and enhance student learning as well. We caution against using technology merely for the sake of using technology (e.g., entering 100 numbers in a graphing calculator and calculating statistical summaries) or for pseudo-accuracy (carrying out results to multiple decimal places). Not all technology tools will have all desired features. Moreover, new ones appear all the time.

TECHNOLOGIES AVAILABLE:

- Graphing calculators
- Statistical packages
- Educational software
- Applets
- Spreadsheets
- Web-based resources, including data sources, online texts, and data analysis routines
- Classroom response systems

SUGGESTIONS FOR TEACHERS ON WAYS TO USE TECHNOLOGY:

- ✓ Access large, real data sets
- ✓ Automate calculations
- ✓ Generate and modify appropriate statistical graphics
- ✓ Perform simulations to illustrate abstract concepts
- ✓ Explore "what happens if . . . " questions
- ✓ Create reports

CONSIDERATIONS FOR TEACHERS WHEN SELECTING TECHNOLOGY TOOLS:

- Ease of data entry, ability to import data in multiple formats
- Interactive capabilities
- Dynamic linking between data, graphical, and numerical analyses

- Ease of use for particular audiences
- Availability to students, portability

Recommendation 6: Use assessments to improve and evaluate student learning.

STUDENTS WILL VALUE WHAT YOU ASSESS. Therefore, key ideas need to be aligned with learning goals. Assessments need to focus on demonstrating understanding key ideas, not just on skills, procedures, and computed answers. A course should include formative assessments (e.g., quizzes, midterm exams, and small projects) along with summative evaluations (e.g., exams and course grades). Useful and timely feedback is essential for assessments to lead to learning. Types of assessment may be more or less practical in different types of courses. However, it is possible, even in large classes, to implement a variety of useful assessments.

CAUSEweb.org hosts a repository of learning outcomes for introductory statistics courses.

TYPES OF ASSESSMENT:

- Homework
- In-class quizzes and exams
- Online quizzes and activities
- Minute papers
- Projects
- Activities
- Oral presentations
- Written reports
- Videos reports
- Article critiques

See the Appendices for examples of model assessment items and suggestions for improving weak items. Other rich sources of items include the ARTIST project (Assessment Resource Tools for Improving Statistical Thinking, https://apps3.cehd.umn.edu/artist) and LOCUS (Levels of Conceptual Understanding in Statistics, https://locus.statisticseducation.org)

SUGGESTIONS FOR TEACHERS:

- ✓ Integrate assessment as an essential component of the course. Assessment tasks that are well-coordinated with what the teacher is doing in class are more effective than tasks that focus on what happened in class two weeks earlier.
- ✓ Use a variety of assessment methods to provide a more complete evaluation of student learning.

- ✓ Assess statistical literacy using assessments such as interpreting or critiquing articles in the news and graphs in media.
- ✓ Use items that focus on choosing good interpretations of graphs or selecting appropriate statistical procedures.
- ✓ Assess statistical thinking using assessments such as student projects and open-ended investigative tasks.

SUGGESTIONS FOR STUDENT ASSESSMENT IN LARGE CLASSES:

- ✓ Use small group projects instead of individual projects.
- ✓ Use peer review of projects to provide feedback and improve projects before grading.
- ✓ Use discussion sections for student presentations (or a special poster session outside of class hours).

Making It Happen

Statistics education has come a long way since Fisher and Snedecor XX add citation in margin?. Moreover, teachers of statistics across the country have generally been enthusiastic about adopting modern methods and approaches. Nevertheless, changing the way we teach isn't always easy. In a way, we are all teachers and learners, a bit like hermit crabs: To grow, we must first abandon the protective shell of what we are used to and endure a period of vulnerability until we can settle into a new and larger set of habits and expectations.

We have presented many ideas in this report. We advise readers to move in the directions suggested by taking small steps at first. Examples of small steps include the following:

- Adding an activity to your course
- Having your students do a small project
- Integrating an applet into a lecture
- Demonstrating the use of software to your students
- Increasing the use of real data sets
- Deleting a topic from the list you currently try to cover to focus more on understanding concepts

Your teaching philosophy will inform your choice of textbook, but the recommendations in this report are not about choosing a text. They are about a way of teaching.

There are many resources available, including the MAA Notes volumes that deal with teaching statistics, the Consortium to Advance Undergraduate Statistics Education (CAUSE) (causeweb.org), the Isostat discussion list (`www.lawrence.edu/fac/jordanj/isostat.html`), the SIGMAA- Stat Ed group within the MAA (`www.pasles.org/sigmaastat`), and the ASA website, especially the Center for Statistics Education (`www.amstat.org/education`) and the Statistical Education Section (`www.amstat.org/sections/educ`).

GAISEing into the Future

A good deal of progress has been made, but there is still plenty of room to improve the introductory statistics course. Moreover, this course must be flexible and adapt to change as more students enter college having learned aspects of statistics in elementary and secondary school. The Advanced Placement course continues to change the statistics education landscape. Although we have been addressing the general introductory course, we must be mindful of other courses, such as business statistics and mathematical statistics, and of the content and goals of good second courses in statistics that build on the solid conceptual understanding developed in the first course.

APPENDICES

EXAMPLES AND COMMENTARY IN THESE APPENDICES ARE PROVIDED FOR ADDITIONAL GUIDANCE, CLARIFICATION, AND ILLUSTRATION OF THE GUIDELINES IN THE MAIN REPORT.

APPENDIX A: EXAMPLES OF ACTIVITIES AND PROJECTS

1. Desirable Characteristics of Class Activities
2. Activities That Could Be Improved
    a. Pepsi vs. Coke Activity
    b. A Central Limit Theorem Activity
3. Additional Examples of Activities and Projects
    a. Data Gathering and Analysis: A Class of Projects
    b. Team constructed questions about relationships
    c. Comparing Manual Dexterity under Two Conditions

APPENDIX B: EXAMPLES OF ASSESSMENT ITEMS

1. Examples of Assessment Items with Problems and Commentary About the Nature of the Difficulty
2. Examples Showing Ways to improve Assessment Items
3. Additional Examples of Good Assessment Items

APPENDIX C: EXAMPLE OF USING TECHNOLOGY

A Technology-Based Simulation to Examine the Effectiveness of Treatments for Cocaine Addiction

APPENDIX D: EXAMPLES OF NAKED, REALISTIC, AND REAL DATA

1. Naked Data (not recommended)
2. Realistic Data (better, but still not the best)
3. Real Data (recommended)

Appendix A: Examples of Activities and Projects

Desirable Characteristics of Class Activities

- The activity should mimic a real-world situation. It should not seem like "busy work." For instance, if you use coins or cards to conduct a binomial experiment, explain real-world binomial experiments they could represent.

- The class should be involved in some of the decisions about how to conduct the activity. Students don't learn much from following a detailed "recipe" of steps.

- The decisions made by the class should require knowledge learned in the class. For instance, if they are designing an experiment, they should consider principles of good experimental design learned in class, rather than "intuitively" deciding how to conduct the experiment.

- If possible, the activity should include design, data collection, and analysis so students can see the whole process at work.

- It is sometimes better to have students work in teams to discuss how to design the activity and then reconvene the class to discuss how it will be done, but it is sometimes better to have the class work together for the initial design and other decisions. It depends on how difficult the issues to be discussed are and whether each team will need to do things in exactly the same way.

- The activity should begin and end with an overview of what is being done and why.

- The activity should be fun!

Activities That Could Be Improved

Pepsi vs. Coke Activity

Today, we will test whether Pepsi or Coke tastes better. Divide into groups of four. Choose one person in your group to be the experimenter. *Note:* If you are not the experimenter, please refrain from looking at the front of the classroom.

1. On the table in the front of the classroom are two large soda bottles, one of Pepsi and one of Coke. There are also cups labeled A and B. The experimenter should go to the table and flip a coin. If it's heads, then pour Pepsi into a cup labeled A and Coke into a cup labeled B. If it's tails, pour Pepsi into cup B and Coke into cup A. Remember which is which. Bring the cups back to your team.
2. Have a team member taste both drinks. Record which one he or she prefer—the one in cup A or the one in cup B.
3. The experimenter should now reveal to the team member if it was Coke or Pepsi that was preferred.
4. The experimenter should repeat this process for each team member once. Then, one of the other team members should give the taste test to the experimenter so each student will have done it once.
5. Come together as a class. Your teacher will ask how many of you preferred Coke.
6. Look up the formula in your book for a confidence interval for a proportion. Construct a confidence interval for the proportion of students in the class who prefer Coke.
7. Do a hypothesis test for whether either drink was preferred by the class.

Critique: The test is not double blind. There is no reason why the experimenter can't be blind to which drink is which. The person who initially sets up the experiment could cover or remove the labels from the drink containers and call them drinks 1 and 2. The drinks could then be prepared in advance into cups labeled A and B. The order of presentation should be randomized for each taster.

Central Limit Theorem Activity

The purpose of this exercise is to verify the Central Limit Theorem. Remember that this theorem tells us that the mean of a large sample is:

- Approximately bell-shaped
- Has mean equal to the mean of the population
- Has standard deviation equal to the population standard deviation/ sqrt(n) — $\sigma/\sqrt{n}$

Critique: This is not a good activity for at least two reasons. First, it has absolutely no real-world motivation and reinforces the myth that statistics is boring and useless. Second, the instructions are too complete. There is no room for exploration on the part of the students; they are simply given a "recipe" to follow.

Please follow these instructions to verify that the Central Limit Theorem holds.

1. Divide into pairs. Each pair should have 1 die.
2. Take turns rolling the die, 25 times each, so you will have 50 rolls. Keep track of the number that lands face up each time.
3. Draw a histogram of the results. The die faces are equally likely, so the histogram should have a "uniform" shape. Verify that it does.
4. Find the mean and standard deviation for the 50 rolls.
5. The mean and standard deviation for rolling a single die are 3.5 and 1.708, respectively. Is the mean for your 50 rolls close to 3.5? Is the standard deviation close to 1.708?
6. Come together as a class. Draw the theoretical curve that the mean of 50 rolls should have. Remember that it's bell-shaped and has a mean equal to the population mean, so that's 3.5 in this case, and the standard deviation in this case should be $1.708/\sqrt{50} = .24$.
7. Have each pair mark their mean for the 50 rolls on the curve. Notice whether they seem reasonable, given what is expected using the Central Limit Theorem.

HOW TO IMPROVE ON THIS ACTIVITY?

The "Cents and the Central Limit Theorem" activity from *Activity-Based Statistics* (Scheaffer et al.) provides an example for illustrating the Central Limit Theorem that is more aligned with the guidelines. Some other good examples from *Activity-Based Statistics*:

- The introduction to hypothesis testing activity (where you draw cards at random from a deck and always get the same color) works well.
- Matching Graphs to Variables generates a lot of discussion and learning.
- Random Rectangles has become a standard, for good reason.
- Randomized Response is not central to the introductory course, but it does involve some statistical thinking.

Additional Examples of Activities and Projects

Data Gathering and Analysis: A Class of Projects

The idea for projects such as the ones described here comes from Robert Wardrop's *Statistics: Learning in the Presence of Variability* (Dubuque, IA: William C. Brown, 1995). These projects, in turn, are based on a study by cognitive psychologists Daniel Kahneman and Amos Tversky.

Consider two versions of the "General's Dilemma":

> Version 1: Threatened by a superior enemy force, the general faces a dilemma. His intelligence officers say his soldiers will be caught in an ambush in which 600 of them will die unless he leads them to safety by one of two available routes. If he takes the first route, 200 soldiers will be saved. If he takes the second, there is a two-thirds chance that 600 soldiers will be saved and a two-thirds chance that none will be saved. Which route should he take?

> Version 2: Threatened by a superior enemy force, the general faces a dilemma. His intelligence officers say his soldiers will be caught in an ambush in which 600 of them will die unless he leads them to safety by one of two available routes. If he takes the first route, 400 soldiers will die. If he takes the second, there is a one-third chance that no soldiers will die and a two-thirds chance that 600 will die. Which route should he take?

Both versions of the question have the same two answers; both describe the same situation. The two questions differ only in their wording: One speaks of lives lost, the other of lives saved.

A pair of questions of this form leads easily to a simple randomized comparative experiment with the two questions as "treatments": Recruit a set of subjects, sort them into two groups using a random number table, and assign one version of the question to each group. The results can be summarized in a 2x2 table of counts:

	Answer	
Question	A	B
Version 1		
Version 2		

The data can be analyzed by comparing the two proportions. Using Fisher's exact test or the chi-square test with continuity correction, for example.

Exercise Set 1.2 in Wardrop's book lists a large number of variations on this structure, many of them carried out by students. Here are abbreviated versions of just four:

- Ask people in a history library whether they find a particular argument from a history book persuasive; the argument was presented with and without a table of supporting data.
- Ask women at the student union whether they would accept if approached by a male stranger and invited to have a drink; the male was/was not described as "attractive."
- Ask customers ordering an ice cream cone whether they want a regular or waffle cone; the waffle cone was/was not described as "homemade."
- Ask college students either (1) Would you recommend the counseling service for a friend who was depressed? Or (2) Would you go to the counseling service if you were depressed?

Projects based on two versions of a two-answer question offer a number of advantages:

- Data collection can be completed in a reasonable length of time.
- Randomization ensures that the results will be suitable for formal inference.
- Randomization makes explicit the connection between chance in data gathering and the use of a probability model for analysis.
- The method of analysis is comparatively simple and straightforward.
- The structure (a 2x2 table of counts) is one with very broad applicability.
- Finally, the format is very open-ended, which affords students a wide range of areas of application from which to choose and offers substantial opportunities for imagination and originality in choosing subjects and the pair of questions.

Team-Constructed Questions About Relationships

Adapted from Project 2.2, Instructors' Resource Manual, *Mind On Statistics*, Utts and Heckard

These instructions are for the teacher. Instructions for students are on the Project 4 Team Form.

Goal: Provide students with experience in formulating a research question, then collecting and describing data to help answer it

Supplies: (N = number of students; T = number of teams)

- N index cards or slips of paper of each of T colors (or use board space; see below)
- T or 2T overhead transparencies and pens (see Step 3 for the reason for 2T of them)
- T calculators

Students should work in teams of 4 to 6. See the Sample Project 4 Team Form.

Step 1: Each team formulates two categorical variables for which they want to know if there is a relationship, such as whether someone is a firstborn (or only) child and whether they prefer indoor or outdoor activities (recent research suggests that firstborns prefer indoor activities and later births prefer outdoor activities); male/female and opinion on something; class (senior, junior, etc.), and whether they own a car, etc. To make it easier to finish in time, you may want to restrict them to two categories per variable.

NOTE: This can also be done with one categorical and one quantitative variable and the data retained for use when doing two-sample inference.

There are two possible methods for collecting data—using index cards (or paper) or using the board. Each of the next few steps will be described for both methods.

Step 2: Cards: Each team is assigned a color from the T colors of index cards. For instance, Team 1 might be blue, Team 2 is pink, and so on. Board: Assign each team space on the chalkboard to write their questions.

Step 3: Each team asks the whole class its two questions. Cards: The team writes the questions on an overhead transparency and displays them, with each team taking a turn to go to the front of the room. Students write their answers on the index card corresponding to that team's color and the team collects them. For instance, all students in the class write their answers to Team 1's questions on the blue index card, their answers to Team 2's questions on the pink card, and so on. Board: A team member writes the questions on the board

along with a two-way table where each student can put a hash mark in the appropriate cell.

Step 4: Cards: After each team has asked its questions and students have written their answers, the cards are collected and given to the appropriate team. For instance, Team 1 receives all the blue cards. Board: All class members go to each segment of the board and put a hash mark in the cell of the table that fits them.

Step 5: Each team tallies, summarizes, and prepares a graphical display of the data for their questions. The results are written on an overhead transparency.

Step 6: Each team presents the results to the class.

Step 7: Results can be retained for use when covering chi-square tests for independence if you are willing to pretend that the data are a random sample from a larger population.

PROJECT 4: TEAM FORM

TEAM MEMBERS:

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____

INSTRUCTIONS:

1. Create two categorical variables for which you think there might be an interesting relationship for class members. If you prefer, you may turn a quantitative variable into a categorical one, such as GPA—high or low (using a cut-off such as GPA $\geq$ 3.0). Each variable should have two categories to make it easier to finish in the allotted time.
2. List the two variables below, designating which is the explanatory variable and which is the response variable, if that makes sense for your situation.
    - Explantory variable:
    - Response variable:
3. Each team will be assigned one segment of the chalk board. One team member is to go to the board and write your two questions. Also, write a "two-way" table on the board in which people will put a "hash mark" into the square that describes them.
4. Everyone will now go to the board and fill in a hash mark in the appropriate box for each team's set of questions.
5. After everyone has gone to the board and filled in all their data, enter the totals in the table below for your team's questions. Also enter what the categories are for each variable.

	Response Variable		
Explanatory Variable	Category 1	Category 2	Total
Category 1			
Category 2			
Total			

6. Create appropriate numerical and graphical summaries to display on your team's overhead transparency. Write a brief summary of your findings below and on the back if needed.
7. A member of each team will present the team's result to the class, using the overhead transparency.
8. Turn in this sheet and the overhead transparency.

Comparing Manual Dexterity Under Two Conditions

Adapted from Project 12.2, Instructors' Resource Manual, *Mind On Statistics*, Utts and Heckard

These instructions are for the teacher. Instructions for students are on the "Project 5 Team Form."

Goal: Provide students with experience in designing, conducting, and analyzing an experiment

Supplies: (N = number of students, T = number of teams)

NOTE: A variation is to have them do the task with and without wearing a latex glove instead of with the dominant and nondominant hand. In that case, you will need *N pairs of latex gloves*

- T bowls filled with about 30 of each of two distinct colors of dried beans
- 2T empty paper cups or bowls
- T stop watches or watches with a second hand

The Story: A company has many workers whose job is to sort two types of small parts. Workers are prone to get repetitive strain injury, so the company wonders if there would be a big loss in productivity if the workers switch hands, sometimes using their dominant hand and sometimes using their nondominant hand. (Or if you are using latex gloves, the story can be that, for health reasons, they might want to require gloves.) Therefore, you are going to design, conduct, and analyze an experiment making this comparison. Students will be timed to see how long it takes to separate the two colors of beans by moving them from the bowl into the two paper cups, with one color in each cup. A comparison will be done after using dominant and nondominant hands. An alternative is to time students for a fixed time, such as 30 seconds, and see how many beans can be moved in that amount of time.

Step 1: As a class, discuss how the experiment will be done. This could be done in teams first. See below for suggestions.

- What are the treatments? What are the experimental units?
- Principles of experimental design to consider are as follows. Use as many of them as possible in designing and conducting this experiment. Discuss why each one is used.
  - Blocking or creating matched-pairs
  - Randomization of treatments to experimental units, or randomization of order of treatments
  - Blinding or double blinding
  - Control group
  - Placebo

    - Learning effect or getting tired
- What is the parameter of interest?
- What type of analysis is appropriate—hypothesis test, confidence interval, or both?

The class should decide that each student will complete the task once with each hand. Why is this preferable to randomly assigning half of the class to use their dominant hand and the other half to use their nondominant hand? How will the order be decided? Should it be the same for all students? Will practice be allowed? Is it possible to use a single or double-blind procedure?

Step 2: Divide into teams and carry out the experiment.
The Project 5 Team Form shows one way to assign tasks to team members.

Step 3: Descriptive statistics and preparation for inference.
Convene the class and create a stemplot of the differences. Discuss whether the necessary conditions for this analysis are met. Were there any outliers? If so, can they be explained? Have someone compute the mean and standard deviation for the differences.

Step 4: Inference.
Have teams reconvene. Each team is to find a confidence interval for the mean difference and conduct the hypothesis test.

Step 5: Reconvene the class and discuss conclusions.

SUGGESTIONS FOR HOW TO DESIGN AND ANALYZE THE EXPERIMENT IN SAMPLE PROJECT 5

Design issues:

- Blocking or creating matched-pairs: Each student should be used as a matched-pair, doing the task once with each hand.
- Randomization of treatments to experimental units, or randomization of order of treatments: Randomize the order of which hand to use for each student.
- Blinding or double blinding: Obviously, the student knows which hand is being used, but the time-keeper doesn't need to know.
- Control group: Not relevant for this experiment.

- Placebo: Not relevant for this experiment.

- Learning effect or getting tired: There is likely to be a learning effect, so you may want to build in a few practice rounds. Also, randomizing the order of the two hands for each student will help with this.

One possible design:

Have each student flip a coin. Heads, start with dominant hand. Tails, start with nondominant hand. Time them to see how long it takes to separate the beans. The person timing them could be blind to the condition by not watching.

Analysis:

• WHAT IS THE PARAMETER OF INTEREST? •

Answer: Define the random variable of interest for each person to be a "manual dexterity difference" of

d = number of extra seconds required with nondominant hand
d = time with non-dominant hand – time with dominant hand

Define μ_d = population mean manual dexterity difference.

• WHAT ARE THE NULL AND ALTERNATIVE HYPOTHESES? •

$$H_0 : \mu_d = 0$$
$$H_A : \mu_d > 0 \text{ (faster with dominant hand)}$$

• IS A CONFIDENCE INTERVAL APPROPRIATE? •

Yes, it will provide information about how much faster workers can accomplish the task with their dominant hands. The formula for the confidence interval is

$$\bar{d} \pm t* \left(\frac{s_d}{\sqrt{n}} \right)$$

where $t*$ is the critical t-value with $df = n - 1$ and s_d is the standard deviation of the difference scores. To carry out the test, compute $t = \frac{\bar{d}-0}{s_d/\sqrt{n}}$, then compare to the critical t-value to find the p-value.

PROJECT 5: TEAM FORM

TEAM MEMBERS:

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____

INSTRUCTIONS:

You will work in teams. Each team should take a bowl of beans and two empty cups. You are each going to separate the beans by moving them from the bowl to the empty cups, with one color to each cup. You will be timed to see how long it takes. You will each do this twice, once with each hand, with order randomly determined.

1. Designate these jobs. You can trade jobs for each round if you wish.

    - *Coordinator* — runs the show
    - *Randomizer* — flips a coin to determine which hand each person will start with, separately for each person
    - *Time-keeper* — must have watch with second hand to time each person for the task
    - *Recorder* — records the results in the table below

2. Choose who will go first. The randomizer tells the person which hand to use first. Each person should complete the task once before moving to the second hand for the first person. That gives everyone a chance to rest between hands.

3. The time-keeper times the person while they move the beans one at a time from the bowl to the cups, separating colors.

4. The recorder notes the time and records it in the table.

5. Repeat this for each team member.

6. Each person then goes a second time, with the hand not used the first time.

7. Calculate the difference for each person.

RESULTS FOR THE CLASS

Record the data here:

NAME	Time for *non dominant* hand	Time for *dominant* hand	d = difference nondominant − dominant hand

Parameter to be tested and estimated is:

Confidence interval:

Hypothesis test—hypotheses and results:

Appendix B: Examples of Assessment Items

Well-designed assessment items help to determine whether students understand key statistical concepts. Since the original GAISE report was written in 2005, there have been many improvements in the ways that instructors and institutions determine whether students have met the learning outcomes for introductory statistics courses.

Students value that which is assessed, so it is important that we assess student learning in a manner consistent with our stated goals. Good items assess the development of statistical thinking and conceptual understanding, preferably using technology and real data.

Below we present exemplary assessment items, some of which include commentary. We also present a few items that are not strong, with suggestions on how they can be improved. Finally, we present advice on constructing a rubric when assessing a project report or presentation.

In general, we want students to interpret results more than we want them to produce results. If we ask a True/False question, we want the student to explain why *a statement is true or is false, so that we can assess the thinking that lead to the answer chosen. However, sometimes the practicalities of teaching a large class mean that an appropriate exam question might be a multiple choice item that does not ask for explanation.*

Examples of Exemplary Assessment Items

We begin by providing examples of exemplary assessment items with commentary about the items.

Item 1

Are metal bands used for tagging harmful to penguins? Researchers investigated this question with a sample of 100 penguins near Antarctica. All of these penguins had already been tagged with RFID chips, and the researchers randomly assigned 50 of them to receive a metal band on their flippers in addition to the RFID chip. The other 50 penguins did not receive a metal band. Researchers then kept track of which penguins survived for the 4.5-year study and which did not. They found that 16 of the 50 penguins with a metal band survived, compared to 31 of the 50 penguins without a metal band.

1. Calculate the difference in the proportions who survived between the two groups.
2. The p-value for comparing the two group's survival propor-

tions turns out to be 0.005. Explain (as if to someone who has not studied statistics) what this p-value means: This is the probability of ...

3. Summarize your conclusion from this p-value. Be sure to address the issue of causation as well as the issue of significance. Also justify your conclusion.

Item 2

Suppose that 20% of undergraduate students at a university own an iPad and 60% of graduate students at the university own an iPad. Is it reasonable to conclude that 40% (the average of 20% and 60%) of all students at the university (undergraduate and graduate students combined) own an iPad? Explain why or why not, as if to a college student who has not taken a statistics class.

Item 3

Suppose that you take a random sample of 100 houses currently for sale in California. Does the Central Limit Theorem suggest that a histogram of the house prices in the sample will display an approximately normal distribution? Explain briefly.

Item 4

Does everyone who scores below the median on this exam necessarily have a negative z-score for this exam? Explain.

Item 5

Describe a situation where a third variable could be masking the relationship between two variables.

Sample solution: For an observational study which assessed the association between coffee drinking and cancer, smoking status could mask (or "confound") the relationship, since smoking could be associated with both coffee drinking and cancer (see also Appendix D, Multivariable thinking).

Item 6

Let Y denote the amount a student spends on textbooks for one semester. Suppose Nancy, who is statistically savvy, wants to know how fall, semester 1, and spring, semester 2, compare. In particular, suppose she is interested in the averages μ_1 and μ_2. You may assume that Nancy has taken several statistics courses and knows a lot about statistics, including how to interpret confidence intervals and hypothesis tests. You have random samples from each semester and are to analyze the data and write a report. You seek advice from four persons:

1. **Rudd says,** "Conduct an $\alpha = 0.05$ test of $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$ and tell Nancy whether you reject H_0."
2. **Linda says,** "Report a 95% confidence interval for $\mu_1 - \mu_2$."
3. **Steve says,** "Conduct a test of $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$ and report to Nancy the p-value from the test."
4. **Gloria says,** "Compare $\bar{y}_1$ to $\bar{y}_2$. If $\bar{y}_1 > \bar{y}_2$, then test $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 > \mu_2$ using $\alpha = 0.05$ and tell Nancy whether you reject H_0. If $\bar{y}_1 < \bar{y}_2$, then test $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 < \mu_2$ using $\alpha = 0.05$ and tell Nancy whether you reject H_0."

Rank the four pieces of advice from worst to best and explain why you rank them as you do. That is, explain what makes one better than another.

Examples of Assessment Items with Problems and Commentary

We next give some examples of assessment items with problems and commentary about the nature of the difficulty.

Assessment items to avoid using on tests: traditional True/False, pure computation without a context or interpretation, items with too much data to enter and compute or analyze, or items that only test memorization of definitions or formulas.

Item 7

A teacher taught two sections of elementary statistics last semester, each with 25 students, one at 8:00 a.m. and one at 4:00 p.m. The means and standard deviations for the final exams were 78 and 8 for the 8:00 a.m. class and 75 and 10 for the 4:00 p.m. class. In examining these numbers, it occurred to the teacher that the better students probably sign up for 8:00 a.m. class. So she decided to test whether the mean final exam scores were equal for her two groups of students. State the hypotheses and carry out the test.

Critique: The teacher has all the population data so there is no need to do statistical inference. In addition, the proposed design has serious flaws in terms of statistical practice.

Item 8

An economist wants to compare the mean salaries for male and female CEOs. He gets a random sample of 10 of each and does a t-test. The resulting p-value is 0.045.

1. State the null and alternative hypotheses.

Critique: The question doesn't address the conditions necessary for a t-test, and with the small sample sizes, they are almost surely violated here. Salaries are almost surely skewed.

2. Make a statistical conclusion.
3. State your conclusion in words that would be understood by someone with no training in statistics.

Item 9

Which of the following gives the definition of a *p*-value?

A. It's the probability of rejecting the null hypothesis when the null hypothesis is true.
B. It's the probability of not rejecting the null hypothesis when the null hypothesis is true.
C. It's the probability of observing data as extreme as that observed.
D. It's the probability that the null hypothesis is true.

Critique: None of these answers is quite correct. Answers B and D are clearly wrong; answer A is the level of significance; and answer C would be correct if it continued "…or more extreme, given that the null hypothesis is true."

Examples Showing Ways to Improve Assessment Items

Item 9 (revisited)

Which of the following gives the definition of a *p*-value?

CHANGED TO:

A randomized trial of the use of bednets to prevent malaria in sub-Saharan Africa yielded a p-value of 0.001. Without resorting to jargon, interpret this result in the context of the problem to someone without background knowledge of statistics.

Sample solution: If bednets were not associated with malaria prevalence then we'd only be likely to see a result this extreme or more extreme one time out of a thousand. Therefore we conclude that bednets must be effective in preventing malaria.

True/False items, even when well-written, do not provide much information about student knowledge because there is always a 50% chance of getting the item right without any knowledge of the topic. One approach is to change the items into forced-choice questions with three or more options.

Item 10

The size of the standard deviation of a data set depends on where the center is. True or False

CHANGED TO:

Does the size of the standard deviation of a data set depend on where the center is located?

A. Yes, the higher the mean, the higher the standard deviation.
B. Yes, because you have to know the mean to calculate the

standard deviation.

C. No, the size of the standard deviation is not affected by the location of the distribution.

D. No, because the standard deviation only measures how the values differ from each other, not how they differ from the mean.

Item 11

A correlation of +1 is stronger than a correlation of −1. True or False

REWRITTEN AS:

A recent article in an educational research journal reports a correlation of +0.8 between math achievement and overall math aptitude. It also reports a correlation of −0.8 between math achievement and a math anxiety test. Which of the following interpretations is the most correct?

A. The correlation of +0.8 indicates a stronger relationship than the correlation of −0.8.

B. The correlation of +0.8 is just as strong as the correlation of −0.8.

C. It is impossible to tell which correlation is stronger.

Context is important for helping students see and deal with statistical ideas in real-world situations.

Item 12

Once it is established that X and Y are highly correlated, what type of study needs to be done to establish that a change in X causes a change in Y?

A CONTEXT IS ADDED:

A researcher is studying the relationship between an experimental medicine and T4 lymphocyte cell levels in HIV/AIDS patients. The T4 lymphocytes, a part of the immune system, are found at reduced levels in patients with the HIV infection. Once it is established that the two variables, dosage of medicine, and T4 cell levels are highly correlated, what type of study needs to be done to establish that a change in dosage causes a change in T4 cell levels?

A. correlational study

B. controlled experiment

C. prediction study

D. survey

Try to avoid repetitious/tedious calculations on exams that may become the focus of the problem for the students at the expense of concepts and interpretations.

Item 13

It was claimed that 1 out of 5 cardiologists takes an aspirin a day to prevent hardening of the arteries. Suppose the claim is true. If 1,500 cardiologists are selected at random, what is the probability that at least 275 of the 1,500 take an aspirin a day?

Critique: This problem requires use of software to calculate the exact binomial or use of the binomial approximation to the normal. Computer output might be provided to augment this question and facilitate solution.

Item 14

A first-year program course used a final exam that contained a 20-point essay question asking students to apply Darwinian principles to analyze the process of expansion in major league sports franchises. To check for consistency in grading among the four professors in the course, a random sample of six graded essays were selected from each instructor. The scores are summarized in the table below. Construct an ANOVA table to test for a difference in means among the four instructors.

Critique: The version of the question above requires a fair amount of pounding on the calculator to get the results and never even asks for an interpretation. The revision below still requires some calculation (which can be adjusted depending on the amount of computer output provided) but the calculations can be done relatively efficiently—especially by students who have a good sense of what the computer output is providing.

Instructor	**Scores**					
Affinger	18	11	10	12	15	12
Beaulieu	14	14	11	14	11	14
Cleary	19	20	15	19	19	16
Dean	17	14	17	15	18	15

Rewritten as:

A first-year program course . . . (same intro as above) . . . The scores are summarized in the table below, along with some descriptive statistics for the entire sample and a portion of the one-way ANOVA output.

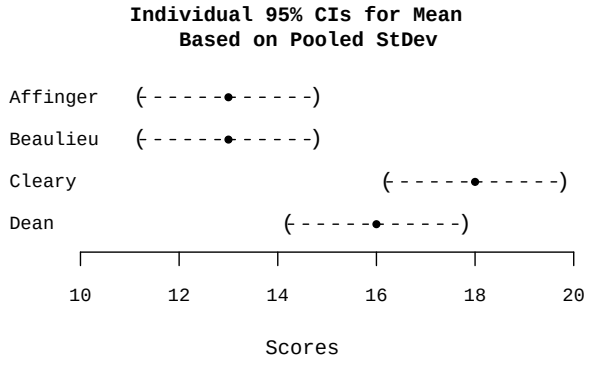
1. Unfortunately, we are missing the ANOVA table from the output. Use the information given above to construct the ANOVA table and conduct a test (5% level) for any significant differences among the average scores assigned by the four instructors. Be sure to include hypotheses and a conclusion. If you have trouble getting one part of the table that you need to complete the rest (or the next question), make a reasonable guess or ask for assistance (for a small point fee).

```
Descriptive Statistics
Variable      N    Mean   Median   TrMean   StDev   SEMean
Score     24.00   15.00    15.00    15.00    2.92     0.60

One-way Analysis of Variance
          ***ANOVA TABLE OMITTED ***

Level         N    Mean    StDev
Affinger      6   13.00     2.97
Beaulieu      6   13.00     1.55
Cleary        6   18.00     2.00
Dean          6   16.00     1.55

Pooled StDev = 2.098
```



2. After completing the ANOVA table, construct a 95% confidence interval for the average score given by Dr. Affinger. *Note: Your answer should be consistent with the graphical display.*

Additional Examples of Good Assessment Items

Item 15

A study found that individuals who lived in houses with more than two bathrooms tended to have higher blood pressure than individuals who lived in houses with two or fewer bathrooms. Can cause-and-effect be determined from this? Why or why not?

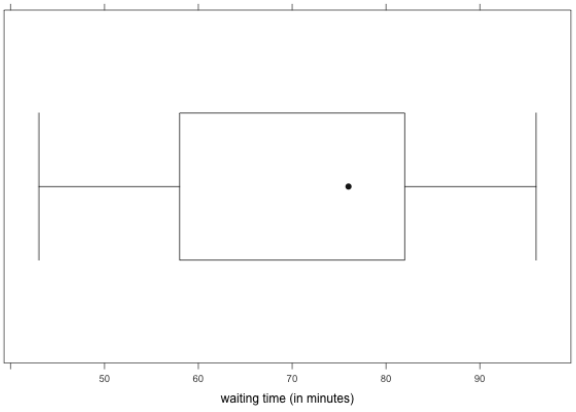
Item 16

Researchers took random samples of subjects from two populations and applied a test to the data; the p-value for the test, using a nondirectional alternative, was 0.06. For each of the following, say whether the statement is true or false and why.

1. There is a 6% chance that the two population distributions really are the same.
2. If the two population distributions really are the same, then a difference between the two samples as extreme as the difference that these researchers observed would only happen 6% of the time.
3. If a new study were done that compared the two populations, there is a 6% probability that H_0 would be rejected again.
4. If $\alpha = 0.05$ and a directional alternative were used, and the data departed from H_0 in the direction specified by the alternative hypothesis, then H_0 would be rejected.

Item 17

As the name suggests, the Old Faithful geyser in Yellowstone National Park has eruptions that come at fairly predictable intervals, making it particularly attractive to tourists. Here is a boxplot of the times between eruptions recorded by an observer.



You are a busy tourist and have only 10 minutes to sit around and watch the geyser. But you can choose when to arrive. If the last eruption occurred at noon, what time should you arrive at the geyser to maximize your chances of seeing an eruption?

1. 12:50pm
2. 1:00pm

3. 1:05pm
4. 1:15pm
5. 1:25pm

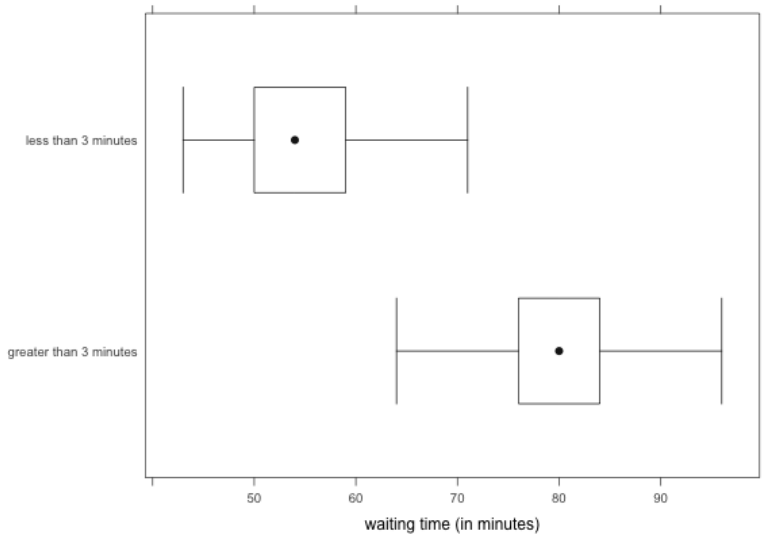
Roughly, what is the probability that in the best 10-minute interval, you will actually see the eruption:

1. 5%
2. 10%
3. 20%
4. 30%
5. 50%
6. 75%

A simple measure of how faithful is Old Faithful is the interquartile range. What is the interquartile range, according to the boxplot above?

1. 10 minutes
2. 15 minutes
3. 25 minutes
4. 35 minutes
5. 50 minutes
6. 75 minutes

Not only are you a busy tourist, you are a smart tourist. Having read about Old Faithful, you understand that the time between eruptions depends on how long the previous eruption lasted. Here's a box plot indicating the distribution of inter-eruption times when the previous eruption duration was less than three minutes.



You can easily ask the ranger what was the duration of the previous eruption.

What is the best 10-minute interval to return (after a noon eruption) so that you will be most likely to see the next eruption, given that the previous eruption was less than three minutes in duration?

1. 12:30 to 12:40
2. 12:40 to 12:50
3. 12:50 to 1:00
4. 1:15 to 1:25
5. 1:25 to 1:35

How likely are you to see an eruption if you return for the most likely 10-minute interval?

1. About 5%
2. About 10%
3. About 20%
4. About 30%
5. About 50%
6. About 75%

Item 18

An article on the CNN web page begins with the sentence, "Family doctors overwhelmingly believe that religious faith can help patients

heal, according to a survey released Monday." Later, the article states, "Medical researchers say the benefits of religion may be as simple as helping the immune system by reducing stress," and Dr. Harold Koenig is reported to say that "people who regularly attend church have half the rate of depression of infrequent churchgoers."

Use the language of statistics to critique the statement by Dr. Koenig and the claim, suggested by the article, that religious faith and practice help people fight depression. You will want to select some of the following words in your critique: observational study, experiment, blind, double-blind, precision, bias, sample, spurious, confounding, causation, association, random, valid, reliable.

Item 19

A student weighed 100 industrial diamonds. She found that the sample average weight was 4.80 grams and the SD was 0.28 grams. *In the context of this setting*, explain what is meant by the sampling distribution of an average.

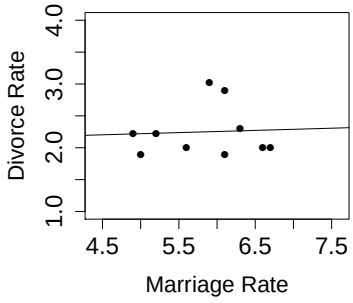
Item 20

A gardener wishes to compare the yields of three types of pea seeds—type A, type B, and type C. She randomly divides the type A seeds into three groups and plants some in the east part of her garden, some in the central part of the garden, and some in the west part of the garden. Then, she does the same with the type B seeds and type C seeds.

1. *What kind* of experimental design is the gardener using?
2. *Why* is this kind of design used in this situation? (Explain in the *context of the situation*.)

Item 21

The scatterplot shows how divorce rate, y, and marriage rate, x, are related for a collection of 10 countries. The regression line has been added to the plot.



1. The U.S. is not one of the 10 points in the original collection of countries. It happens that the U.S. has a higher marriage rate than any of the 10 countries. Moreover, the divorce rate for the U.S. is higher than one would expect, given the pattern of the other countries. How would adding the U.S. to the data set affect the regression line? Why?

2. Think about the scatterplot and regression line after the U.S. has been added to the data set. Provide a sketch of the residual plot. Label the axes and identify the U.S. on your plot with a triangle.

Item 22

Researchers wanted to compare two drugs, formoterol and salbutamol, in aerosol solution to a placebo for the treatment of patients who suffer from exercise-induced asthma. Patients were to take a drug or the placebo, do some exercise, and then have their "forced expiratory volume" measured. There were 30 subjects available.

1. Should this be an experiment or an observational study? Why?
2. Within the context of this setting, what is the placebo effect?
3. Briefly explain how to set up a randomized blocks design (RBD) here.
4. How would an RBD be helpful? That is, what is the main advantage of using an RBD in a setting like this?

Item 23

For each of the following three settings, state the type of analysis you would conduct (e.g., one-sample *t*-test, regression, chi-square test of independence, chi-square goodness-of-fit test, etc.) if you had all the raw data and specify the roles of the variable(s) on which you would perform the analysis, but *do not actually carry out the analysis*.

1. A student measured the effect of exercise on pulse for each of 13 students. She measured pulse before and after exercise (doing 30 jumping jacks) and found that the average change was 55.1 and the SD of the changes was 18.4. How would you analyze the data?
2. Three HIV treatments were tested for their effectiveness in preventing progression of HIV in children. Of 276 children given drug A, 259 lived and 17 died. Of 281 children given drug B, 274 lived and seven died. Of 274 children given drug C, 264 lived and 10 died. How would you analyze the data?
3. A researcher was interested in the relationship between blood pressure and physical activity. He measured the blood pressure and weekly total number of steps from a Fitbit for 125 women. How would you analyze these data?

Item 24

I constructed parallel dotplots of the data from four samples. I then conducted a test of $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ and rejected H_0 at the $\alpha = 0.05$ level. I also tested $H_0 : \mu_1 = \mu_2 = \mu_3$ and rejected H_0 at the $\alpha = 0.05$ level. However, when I tested $H_0 : \mu_2 = \mu_3$ using $\alpha = 0.05$, I did *not* reject H_0. Likewise, when I tested $H_0 : \mu_1 = \mu_4$ using $\alpha = 0.05$, I did *not* reject H_0.

1. Your job is to sketch a graph of the parallel dotplots of the data. That is, based on what I told you about the tests, you should have an idea of how the data look. Use that idea to draw a graph. Indicate the sample means with triangles that you add to the dotplots.
2. It is possible to get data with the same sample means that you graphed in part 1, but for which the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ is not rejected at the $\alpha = 0.05$ level. Provide a graph of this situation. That is, keep the same sample means (triangles) you had from part 1, but show how the data would have been different if H_0 were not to be rejected.

Item 25

A student collected data on a random sample of 12 breakfast cereals. They recorded x = fiber (in grams/ounce) and y = price (in cents/ounce). A scatterplot of the data shows a linear relationship. The fitted regression model is

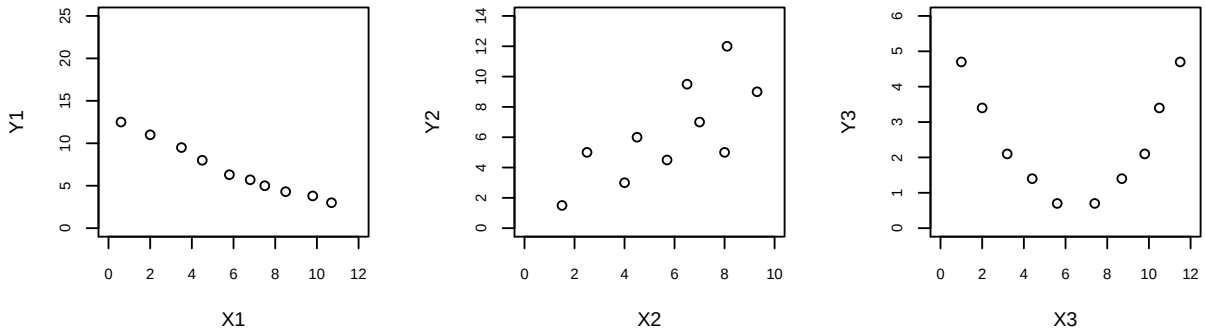
$$\hat{y} = 17.42 + 0.62x$$

The sample correlation coefficient, r, is 0.23. The SE of b_1 is 0.81. Also, $s_{y|x} = 3.1$.

1. Find r^2 and interpret r^2 in the context of this problem.
2. Suppose a cereal has 2.63 grams of fiber/ounce and costs 17.3 cents/ounce. What is the residual for this cereal?
3. Interpret the value of $s_{y|x}$ in the context of this problem. That is, what does it mean to say that $s_{y|x} = 3.1$?
4. In the context of this problem, explain what is meant by "the regression effect."

Item 26

Give a rough estimate of the sample correlation for the data in each of the scatterplots below.



Item 27

Identify whether a scatterplot would or would not be an appropriate visual summary of the relationship between the variables. In each case, explain your reasoning.

1. Blood pressure and age
2. Region of country and opinion about stronger gun control laws
3. Verbal SAT and math SAT score
4. Handspan and gender (male or female)

Item 28

The paragraphs that follow each describe a situation that calls for some type of statistical analysis. For each, you should:

1. Give the name of an appropriate statistical procedure to apply (from the list below). You may use the same procedure more than once, and some questions might have more than one correct answer.
2. In some problems, you will also be given a p-value. Use it to reach a conclusion for that specific problem. Be sure to say something more than just Reject H_0 or Fail to Reject H_0. (Assume a 5% significance level.)

Some statistical procedures you might choose:

A. Anthropologists have found two burial mounds in the same region. They know several tribes lived in the region and that the tribes have been classified according to different lengths of skulls. They measure a random sample of skulls found in each

Confidence interval (for a mean, p, ...)	Normal distribution
Determining sample size	Correlation
Test for a mean	Simple linear regression
Test for proportion	Multiple regression
Difference in means (paired data)	Two-way table (chi-square test)
Difference in means (two independent samples)	ANOVA for difference in means
Difference in proportions	Two-way ANOVA for means

burial mound and wish to determine if the two mounds were made by different tribes. (p-value = 0.0082)

B. The Hawaiian Planters Association is developing three new strains of pineapple (call them A, B, and C) to yield pulp with higher sugar content. Twenty plants of each variety (60 plants in all) are randomly distributed into a two-acre field. After harvesting, the resulting pineapples are measured for sugar content and the yields are recorded for each strain. Are there significant differences in average sugar content between the three strains? (p-value = 0.987)

C. Researchers were commissioned by the Violence In Children's Television Investigative Monitors (VICTIM) to study the frequency of depictions of violent acts in Saturday morning TV fare. They selected a random sample of 40 shows that aired during this time period over a 12-week period. Suppose 28 of the 40 shows in the sample were judged to contain scenes depicting overtly violent acts. How should they use this information to make a statement about the population of all Saturday morning TV shows?

D. The Career Planning Office is interested in seniors' plans and how they might relate to their majors. A large number of students are surveyed and classified according to their MAJOR (Natural Science, Social Science, Humanities) and FUTURE plans (Graduate School, Job, Undecided). Are the type of major and future plans related? (p-value = 0.047)

E. *Sophomore Magazine* asked a random sample of 15 year olds if they were sexually active (yes or no). They would like to see if there is a difference in the responses between boys and girls. (p-value = 0.029)

F. Every week during the Vietnam War, a body count (number of enemy killed) was reported by each army unit. The last digits of these numbers should be fairly random. However, suspicions arose that the counts might have been fabricated. To test this, a large random sample of body count figures was examined and the frequency with which the last digit was a 0 or a 5 was recorded. Psychologists have shown that people making up their own random numbers will use these digits less often than random chance would suggest (i.e., 103 sounds like a more "real" count than 100). If the data were authentic counts, the proportion of numbers ending in 0 or 5 should be about 0.20. (p-value=0.002)

G. In one of his adventures, Sherlock Holmes found footprints made by the criminal at the scene of a crime and measured the distance between them. After sampling many people, measuring their height and length of stride, he confidently announced that he could predict the height of the suspect. How?

Item 29

Some of the statistical inference techniques we have studied include:

A. One-sample z-procedures for a proportion

B. Two-sample z-procedures for comparing proportions

C. One-sample t-procedures for a mean

D. Two-sample t-procedures for comparing means

E. Paired-sample t-procedures

F. Chi-square procedures for two-way tables

G. ANOVA procedures

H. Linear regression procedures

For each of the following research questions (on the next page), indicate (by letter) the appropriate statistical inference procedure for investigating the question.

The list of methods or examples can be shortened.

A. Economists compared starting salaries of new employees across three different groups: those with graduate degrees, those with only bachelor's degrees, and those with no higher education degrees.

B. A researcher investigated whether laughter increases blood flow by having subjects watch a humorous movie and a stressful movie, randomly deciding which movie the subject would see first, measuring the blood flow through the person's blood vessels while watching the movie.

C. Student researchers investigated whether balsa wood is less elastic after it has been immersed in water. They took 44 pieces of balsa wood and randomly assigned half to be immersed in water and the other half not to be. They measured the elasticity by seeing how far (in inches) the piece of wood would project a dime into the air.

D. Do more than two-thirds of students at a particular university have at least one class on Fridays during this term?

E. Are people more likely to fill in the missing letter in F A I _ with an L if they are given a red pen rather than a blue pen?

F. Is there an association between a college student's level of drinking alcohol (classified as none, some, or considerable) and her/his residence situation (classified as living on-campus, off-campus with parents, or off-campus but not with parents)?

G. A researcher used data from the American Time Use Survey (ATUS) to investigate whether high school math teachers tend to spend more time working per day than high school history teachers.

H. Biologists recorded the frequency of a cricket's chirps (in chirps per minute) and also the temperature (in degrees Fahrenheit) when the cricket measurement was recorded. They investigated whether chirp frequency is a significant predictor of temperature.

Item 30

How accurate are radon detectors of a type sold to homeowners? To answer this question, university researchers placed 12 detectors in a chamber that exposed them to 105 picocuries per liter of radon. The detector readings found are below. A printout of the descriptive statistics from Minitab follows.

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

This item might be improved by providing more output (e.g., 95% confidence interval) to allow students to tackle it without calculation or use of a table

Variable	N	Mean	Median	TrMean	StDev	SE Mean	Minimum	Maximum	Q1	Q3
readings	12	104.13	102.75	103.54	9.40	2.71	91.90	122.30	96.90	109.90

1. Is there convincing evidence that the mean 20 readings of all detectors of this type differs from the true value of 105? Perform the appropriate hypothesis test with $\alpha = 0.05$.
2. What is the Type I error associated with this problem?
3. What is the Type II error associated with this problem?
4. What is the probability of a Type II error if the reading of the detectors is too low by 5 picocuries (really 100 when it should read 105)?

Item 31

According to a U.S. Food and Drug Administration (FDA) study, a cup of coffee contains an average of 115 mg of caffeine, with the amount per cup ranging from 60 to 180 mg depending on the brewing method. Suppose you want to repeat the FDA experiment to obtain an estimate of the mean caffeine content to within 5 mg with 95% using your favorite brewing method. In problems such as this, we can estimate the standard deviation of the population to be $\frac{1}{4}$ of the range. How many cups of coffee must you brew to be 95% confident?

Item 32

An internet company is planning to test which of two online ad campaigns is more effective in generating clicks on their site. Outline the design of an experiment you would use to examine this claim. Assume you have money to place 500 ads for each of the two possible campaigns.

Item 33

A study of iron deficiency among infants compared samples of infants following different feeding regimens. One group contained breast-fed infants, while the children in another group were fed a standard baby formula without any iron supplements. A graphical display indicates that the blood hemoglobin levels in children (both breast-fed and formula-fed) are approximately normally distributed in each group. Here are the summary results on blood hemoglobin

levels at 12 months of age:

Group	N	$\bar{X}$	s
Breast-fed	230	13.3	1.7
Formula	230	12.4	1.8

The two sample t-test yielded a test statistics of 5.51 with 458 degrees of freedom. This is associated with a two-sided p-value that was less than 0.0001. Interpret the results from the test statistic and p-value that are provided. Be sure to report the observed difference in groups in the context of the problem.

Item 34

Which implies a stronger linear relationship, a correlation of +0.4 or a correlation of −0.6? Briefly explain your choice.

Item 35

A group of physicians subjected the polygraph to the same careful testing given to medical diagnostic tests. They found that if 1,000 people were subjected to the polygraph and 500 told the truth and 500 lied, the polygraph would indicate that approximately 185 of the truth-tellers were liars and 120 of the liars were truth-tellers. In the application of the polygraph test, an individual is presumed to be a truth-teller until indicated that s/he is a liar. What is a Type I error in the context of this problem? What is the probability of a Type I error in the context of this problem? What is a Type II error in the context of this problem? What is the probability of a Type II error in the context of this problem?

Item 36

Audiologists recently developed a rehabilitation program for hearing-impaired patients in a Canadian program for senior citizens. A simple random sample of the 30 residents of a particular senior citizens home and the seniors were diagnosed for degree and type of sensorineural hearing loss which was coded as follows: 1 = hear within normal limits, 2 = high-frequency hearing loss, 3 = mild loss, 4 = mild-to-moderate loss, 5 = moderate loss, 6 = moderate-to-severe loss, and 7 = severe-to-profound loss. The data are as follows:

6 7 1 1 2 6 4 6 4 2 5 2 5 1 5
4 6 6 5 5 5 2 5 3 6 4 6 6 4 2

1. Create a boxplot of the data.
2. Give a good description of the data.
3. Find a 95% confidence interval for the mean hearing loss of senior citizens in this Canadian program. The mean and standard deviation of the above data are 4.2 and 1.808, respectively. Interpret the interval.

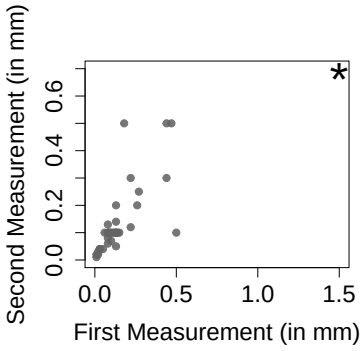
Item 37

A utility company was interested in knowing if agricultural customers would use less electricity during peak hours if their rates were different during those hours. Customers were randomly assigned to continue to get standard rates or to receive the time-of-day structure. Special meters were attached that recorded usage during peak and off-peak hours; the technician who read the meter did not know what rate structure each customer had.

1. Is this an observational study or experiment? Defend your answer.
2. What are the explanatory and response variables?
3. List a potential confounding variable in this work.
4. Is this a matched-pair design? Defend your answer.

Item 38

At the beginning of the semester, we measured the width of a page in our statistics book. Below is the scatterplot of the first measurement vs. the second measurement.



1. Describe the distribution.
2. What effect does the starred point have on the correlation? That is, if the starred point were removed, how would be correlation change, if at all?

Item 39

A study in the *Journal of Leisure Research* investigated the relationship between academic performance and leisure activities. Each in a sample of 159 high-school students was asked to state how many leisure activities they participated in weekly. From the list, activities that involved reading, writing, or arithmetic were labeled "academic leisure activities." Some of the results are in the table below:

	Mean	Standard Deviation
GPA	2.96	0.71
Number of leisure activities	12.38	5.07
Number of academic leisure activities	2.77	1.97

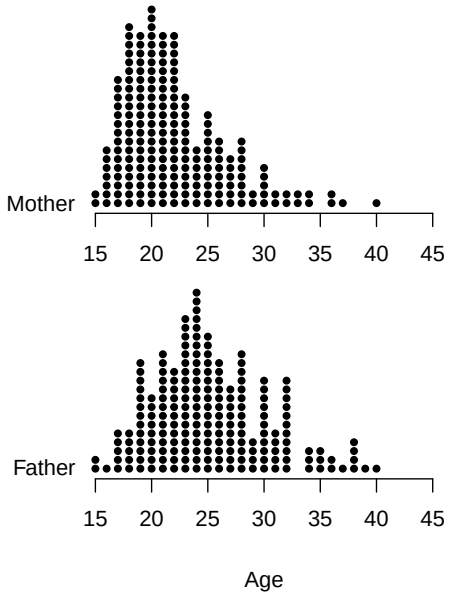
Based on these numbers (and knowing that the GPA is a value between 0 and 4 and the number of activities cannot be negative), discuss the potential skewness of each of the above variables.

Item 40

Events A and B are disjoint. Discuss whether or not A and B can be independent.

Item 41

A sample of 200 mothers and a sample of 200 fathers were taken. The age of the mother when she had her first child and the age of the father when he had his first child were recorded.



1. Describe the data for the mothers' age.
2. Describe the data for the fathers' age.
3. Compare the distributions.
4. A suggestion is made to check the correlation between the ages if we wish to compare the two populations. Is this a good suggestion? Why or why not?

Item 42

When doing a randomized experiment, the original randomization of units to treatment groups breaks the association between

1. The explanatory variable and the response variable
2. The explanatory variable and confounding variables
3. The response variable and confounding variables

Item 43

When doing a randomization test, the simulated re-randomization of units to treatment groups breaks the association between

1. The explanatory variable and the response variable

2. The explanatory variable and confounding variables
3. The response variable and confounding variables

Item 44

For each of the following, circle your answer to indicate whether the quantity can NEVER be negative or can SOMETIMES be negative:

1. z-score SOMETIMES NEVER
2. Probability SOMETIMES NEVER
3. Test statistic SOMETIMES NEVER
4. Sample proportion SOMETIMES NEVER
5. Standard deviation SOMETIMES NEVER
6. Inter-quartile range SOMETIMES NEVER
7. Standard error SOMETIMES NEVER
8. p-value SOMETIMES NEVER
9. Slope coefficient SOMETIMES NEVER
10. Correlation coefficient SOMETIMES NEVER

Item 45

A high school statistics class wants to estimate the average number of chocolate chips in a generic brand of chocolate chip cookies. They collect a random sample of cookies, count the chips in each cookie, and calculate a 95% confidence interval for the average number of chips per cookie (18.6 to 21.3). Indicate if each is VALID or INVALID.

1. We are 95% certain that the confidence interval of 18.6 to 21.3 includes the true average number of chocolate chips per cookie. VALID INVALID
2. We are 95% certain that each cookie for this brand has approximately 18.6 to 21.3 chocolate chips. VALID INVALID
3. We expect 95% of the cookies to have between 18.6 and 21.3 chocolate chips. VALID INVALID

Multiple True/False items of this sort can provide very useful information. If there is a single correct understanding for a statistical concept, but several known misunderstandings for the same concept, a multiple T/F item can provide information on whether or not a student correctly recognizes each of the misunderstandings as false or invalid.

Item 46

Suppose that you have a very big container with 1000 candies in it; 600 are red and 400 are yellow. The candies are all mixed up in the

container. With eyes closed, students draw 10 candies, one at a time. They record whether each candy was red or yellow, and then replace and remix the candies. The teacher asked each student to do this six times. Kate, Dana, and Minh decided to play a trick on their teacher. Only one of them actually did the experiment, while the other two just made up their data. Below are their reports, where R = red candy and Y = yellow candy.

It's important to note that while students like candy (particularly if they can be eaten after the class), they may have had experiences of this sort beginning in elementary or middle school. Rob Gould's *Statistics and the Modern Student* stresses the importance of real and complex data that have broader meaning.

Kate:	Dana:	Minh:
RRYRYRRYYR (6 red)	RYYRRRYRRY (6 red)	RRYRYRRRRY (7 red)
YRRYRRRYRY (6 red)	YRYRRYYRRY (3 red)	RYYYRRRRRR (7 red)
RRYRYRRRYY (6 red)	RRYRRRYRRR (8 red)	YRRRYYRRYY (5 red)
RYRRRYRYRY (6 red)	YRYYYYYYYY (1 red)	RRRRRYRRYR (8 red)
YRRYRRYYRR (6 red)	RRRRRRRRYR (9 red)	RYRRYYYRRR (6 red)
RRRYRYYRYR (6 red)	RYYRYYRRYY (4 red)	YRYRYRRRRR (7 red)

Which student do you think is most likely to have done the experiment?

Item 47

Consider an observational study of the effects of second-hand smoke on health in which we want to compare non-smokers (i) who live with a smoker to (ii) those who do not live with a smoker. There are two ways in which independence is relevant in the sampling and data collection process. (a) Give an example in which one type of independence is met but the other is not; (b) give an example in which the other type of independence is met but the first is not.

Item 48

A terse report of a statistical test is given below:

> The P-value for a hypothesis test with hypotheses $H_0 : \mu = 3$ versus $H_1 : \mu \neq 3$ is 0.04.

Critique the following responses for clarity, completeness and correctness.

1. This means that the probability of getting our test statistic is 0.04.
2. This means that the probability of getting a test statistic at least as extreme as ours is 0.04.
3. This means that if the null hypothesis is true, the probability of getting a test statistics at least as extreme as ours is 0.04

4. This means that if the null hypothesis is true, the probability of getting a test statistic less than or equal to the one we got is 0.04
5. This means that it is very unlikely that the result that was used to compute this P-value would have happened by pure chance alone, assuming that Ho is true. Therefore we could conclude that the evidence is against the Null Hypothesis , and Ho is probably not true.
6. The sentence means that assuming the population average is equal to three, the likelihood of getting an average as large or larger than we got for our sample is about 4 percent.
7. The p-value is the probability that the data will be as extreme or more extreme as the alternate hypothesis suggests.

Item 49

Another exam question asked for an explanation of what the following sentence means:

The interval (2.25, 2.75) is a 99% confidence interval for the mean GPA of UT students having between 45 and 60 credit hours.

Critique the following responses for clarity and correctness.

1. A 99% confidence interval is used to show that 99% of the time when you pick a sample from the population (students having between 45 and 60 credit hours) you will find a mean GPA in the interval (2.25, 2.75).
2. There is a 99% chance that $2.25 \leq \mu \leq 2.75$.
3. This means that if we took many, many simple random samples and constructed a confidence interval based on each sample, 99% of the resulting confidence intervals would contain the true mean.

Item 50

For each part, draw a scatter plot satisfying the conditions given, or else explain why the conditions are impossible:

1. Regression line has small positive slope and correlation is high and positive.
2. Regression line has large positive slope and correlation is high and positive.
3. Regression line has small positive slope and correlation is low and positive.

4. Regression line has large positive slope and correlation is low and positive.

5. Regression line has positive slope and correlation is negative.

Examples of Assessments for Presentations and Projects

Halvorsen (ICOTS, 2010, http://iase-web.org/documents/papers/icots8/ICOTS8_4G3_HALVORSEN.pdf provides motivation for the use of projects as well as details of specific deliverables.

Projects and presentations are an increasingly common component of introductory statistics courses. Projects provide an opportunity for students to learn statistics by doing statistics. They demonstrate that statistical practice includes formulating a statistical question, designing a plan for collecting relevant data, using appropriate statistical methods for analyzing the data, and presenting results in a public setting such as a poster, oral presentation, or a paper (Halvorsen, ICOTS 2010). Students have the opportunity to develop statistical questions that arise from broader research questions, to design data analysis plans, and to communicate results. We provide a basic rubric for presentations and projects along with a sample numeric grading scheme.

Core Competency	Needs Improvement	Basic	Surpassed
Computation Perform computations	Computations contain errors and extraneous code	Computations are correct but contain extraneous/unnecessary computations	Computations are correct and properly identified and labeled
Analysis Choose and carry out analysis appropriate for data and context	Choice of analysis is overly simplistic, irrelevant, or missing key component	Analysis appropriate, but incomplete, or not important features and assumptions not made explicit	Analysis appropriate, complete, advanced, relevant, and informative
Synthesis Identify key features of the analysis, and interpret results (including context)	Conclusions are missing, incorrect, or not made based on results of analysis	Conclusions reasonable, but is partially correct or partially complete	Make relevant conclusions explicitly connected to analysis and to context
Visual presentation Communicate findings graphically clearly, precisely, and concisely	Inappropriate choice of plots; poorly labeled plots; plots missing	Plots convey information correctly but lack context for interpretation	Plots convey information correctly with adequate/appropriate reference information
Verbal Communicate findings in writing clearly, precisely, and concisely	Explanation is illogical, incorrect, or incoherent.	Explanation is partially correct but incomplete or unconvincing	Explanation is correct, complete, and convincing

If needed, the competencies can be converted into a numeric score. One might begin by giving a score of 85 for achieving basic competency in all 5 categories. Then we add to this score for competencies

that surpass the basic level and subtract for those that need improvement. Three points might be added (subtracted) for each of the first three competencies that have surpassed the basic (need improvement), with four points added (subtracted) for the fourth competency that is surpassed (needs improvement) and five points for the fifth competency. In other words, it is increasingly challenging to surpass the basic competency, and it is increasingly problematic to not achieve basic competency. For example, if all five competencies are rated "surpassed", the score is $85 + 3 * 3 + 4 + 5 = 100$; if 4 competencies are rated "surpassed" and the fifth is "basic" then the score is $85 + 3 * 3 + 4 = 95$; and for 3 "surpassed", 1 "needs improvement", and 1 "basic", the score is $85 + 3 * 3 - 3 = 91$. If a competency is missing, then 15 points are subtracted regardless of how many competencies are categorized as needing improvement.

Appendix C: Example of Using Technology

This example starts with a real-world situation, has students do a physical simulation using cards, and then brings in computer technology to automate the simulation.

A Technology-Based Simulation to Examine the Effectiveness of Treatments for Cocaine Addiction

A study on the treatment of cocaine addiction described the results of an experiment comparing two drugs for helping addicts stay off cocaine (D.M. Barnes, "Breaking the Cycle of Cocaine Addiction", *Science*, Vol. 241, 1988, pp. 1029-1030). A group of 48 cocaine addicts who were seeking treatment were randomly divided into two groups of 24. One group was treated with a new drug called desipramine, while the other group was given lithium. The results are summarized in the table below, where we consider patients who do not relapse as successfully treated.

	No Relapse	Relapse
Desipramine	14	10
Lithium	6	18

While we observe that desipramine was more successful than lithium in this particular experiment, can we conclude that the improvement is statistically significant. (i.e., Would we expect to see such a large difference if the drugs were equally effective and it was just the random assignment process that happened to get so many more successful cases in the desipramine group?) We will address this question through simulation, first using a physical demonstration based on shuffling cards, then with a computer simulation that allows us to see the differences for many random assignments of the addicts to the treatment groups.

Physical Simulation

Take a deck of 54 playing cards (including two jokers) and remove six of the black cards (spades or clubs). The remaining deck should match the subjects in the cocaine experiment with all the red cards and the jokers representing patients who relapsed and the 20 black cards representing patients who were treated successfully. If we shuffle the deck and deal out two piles of 24 cards each, we will simulate the assignment of addicts to the two treatment groups when the success does not depend on which drug they take. Do so and fill in the two-way table with the "success" (black cards) and "relapse" (red/jokers) counts for each group.

	No Relapse	Relapse
Desipramine		
Lithium		

Note that once you know one number in the table, you can fill in the rest, as you know there are 24 in each treatment group and 20 will not relapse while 28 will relapse (that is why we sometimes say there is just one degree of freedom in the 2x2 table). To keep things simple then, we can just keep track of one count, such as the number of "no relapse" in the desipramine group.

Shuffle all the cards again, deal 24 for the desipramine group, and count the number of black cards.

Number of "no relapse" in desipramine group = _____

Pool the results for your class (counting # of black cards in each random group of 24 cards assigned to the "desipramine" group) in a dotplot. How often was the number of black cards as large as (or larger than) the 14 cases observed in the actual experiment?

The *p*-value of the original data is the proportion, assuming both drugs are equally effective, of random assignments that have 14 or more "no relapse" cases going to the desipramine group. Estimate this proportion using the data in your class dotplot.

Computer Simulation

To get a more accurate estimate of the proportion of random assignments that put 14 or more no relapse cases into the desipramine group, we'll turn to a computer simulation.

Start with a data set (provided online) consisting of two columns and 48 rows. The first column (Treatment) has the value "desipramine" in the first 24 rows and "lithium" in the remaining 24 rows. The second column (Result) has the values "no relapse" and "relapse" to match the data in the original 2x2 table from the cocaine experiment.

Have the computer permute the values in the "Result" column to represent a new random assignment of subjects to the treatment groups where the outcome does not depend on which drug was taken. Count the number of "no relapse" cases in the desipramine treatment group and have the result stored somewhere. Automate this process to repeat 1,000 times[11].

Look at a histogram or dotplot of the distribution of counts for the 1,000 simulations. Does it seem unusual to have as many as 14 "no relapse" cases in the desipramine group?

Count the number of simulations that have 14 or more successes in the desipramine group (either from the graph, if feasible, or by sorting the simulated counts column) and divide by 1,000 to get another approximation of the *p*-value for the original data.

Does it seem reasonable that the larger number (14) of successful cases appeared in the desipramine group by chance, or would it be more appropriate to conclude that desipramine probably works better than lithium at treating cocaine addiction?

[11] *Some technology alternatives:* The most difficult step here is to automate the simulations to record the counts for many random assignments. Some packages, such as Fathom, have easy-to-use tools designed for exactly such purposes. Others, such as Minitab, allow a bit of programming through macros that can be built in advance and repeated in a loop. A somewhat less enlightening simulation could be accomplished with a stat package that allows generation of random data from a hypergeometric distribution, although students would then lose the connection to the physical randomizations. Finally, an ambitious instructor could construct (or possibly find on the web) an applet to perform the required simulations and collect the results.

Appendix E: Multivariable Thinking

Introduction

The new ASA guidelines for undergraduate programs in statistics recommend that students obtain a clear understanding of principles of statistical design and tools to assess and account for the possible impact of other measured and unmeasured confounding variables[12]. An introductory statistics course cannot cover these topics in depth, but it is important to expose students to them even in their first course[13].

Perhaps the best place to start is to consider how a third variable can change our understanding of the relationship between two variables.

In this appendix we offer two simple examples where "other factors" may arise where appropriate analyses can be undertaken through stratification. Such an approach requires no advanced methods, nor even any inference (though some instructors may incorporate other related concepts and approaches). These examples can help to introduce students to these concepts.

What is especially concerning in a first (and often only) statistics course is that the simple examples we use to introduce statistics concepts can give students the false impression that all data come from well-conducted randomized trials with no dropout, full adherence, and sufficient blinding and that regression models always include all the appropriate predictors. We hope that introductory statistics courses can open students' eyes to the complexity of real data and the need to take that complexity into account.

Including one or more multivariable examples early in an introductory statistics course may help to prepare students to deal with more than one or two variables at a time.

[12] American Statistical Association. Curriculum guidelines for undergraduate programs in statistical science. Technical report, 2014. URL http://www.amstat.org/education/curriculumguidelines.cfm, lastaccessedAugust16,2015

See also Wild's "On locating statistics in the world of finding out", http://arxiv.org/abs/1507.05982.

[13] X L Meng. Statistics: Your chance for happiness (or misery). *The Harvard Undergraduate Research Journal*, 2(1), 2011. URL http://thurj.org/as/2011/01/1259

The *Journal of Statistics Education* Datasets and Stories department features a number of interesting multivariable examples that feature confounding.

Smoking in Whickham

A follow-up study of 1,314 people in Whickham, England characterized smoking status at baseline, then mortality after 10 years[14]. The

[14] DR Appleton, JM French, and MPJ Vanderpump. Ignoring a covariate: an example of Simpson's paradox. *The American Statistician*, 50(4):340–341, 1996

data are provided in Table 1.

SMOKER	Alive	Dead
No	502 (68.6%)	230 (31.4%)
Yes	443 (76.1%)	139 (23.9%)

Table 1: Distribution of smoking status and mortality status (after ten years) from the Whickham study

We see that the risk of dying is lower for smokers than for non-smokers, since 31.4% of the non-smokers died, but only 23.9% of the smokers did not survive over the ten year period.

A graphical representation using a mosaicplot (also known as an *Eikosogram*) represents the cell probabilities as a function of area.

Figure 1: Mosaicplot representing the association between smoking status and mortality in the Whickham study

We note that the majority of subjects have survived, but that the area for the smokers who are still alive is larger than we would expect if there were no association between these variables. What could explain this result?

Let's consider stratification by age of the participants. Table 2 and Figure 2 displays the relationship between smoking and mortality over a 10–year period for those age 18–44, those 45-64, and subjects that were 65 or older at baseline.

Baseline age	SMOKER	Alive	Dead
18-44	No	327 (96.5%)	12 (3.5%)
18-44	Yes	270 (94.7%)	15 (5.3%)
45-64	No	147 (73.5%)	53 (26.5%)
45-64	Yes	167 (67.6%)	80 (32.4%)
65+	No	28 (14.5%)	165 (85.5%)
65+	Yes	6 (12.0%)	44 (88.0%)

Table 2: Distribution of smoking status and mortality status (after 10 years) stratified by age (at baseline) from the Whickham study

We see that mortality rates are low for the youngest group, but the mortality rate is slightly higher for smokers than non-smokers (5.3%

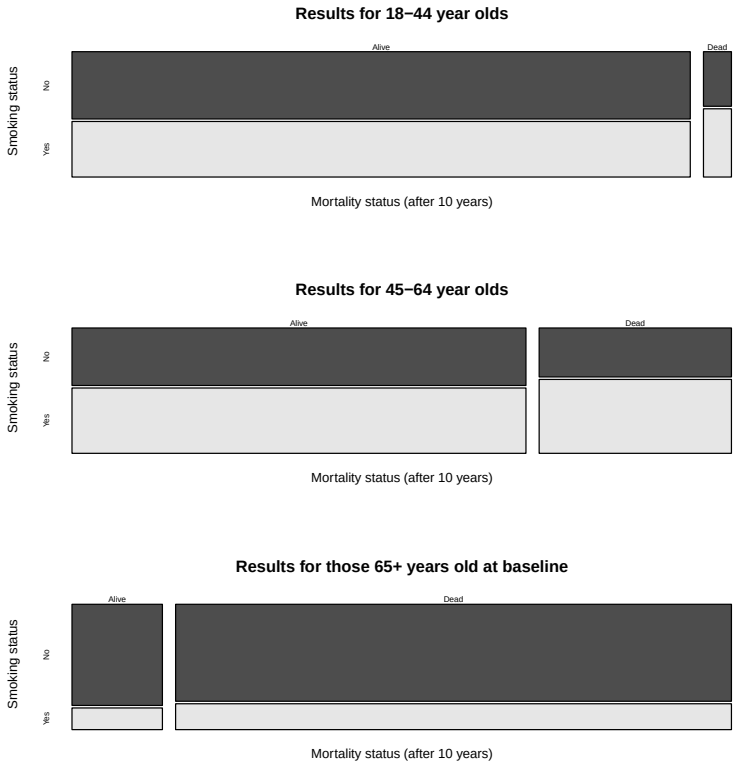


Figure 2: Mosaicplot representing the association between mortality and smoking status (stratified by baseline age) in the Whickham study

for smokers vs 3.5% for non-smokers). (Given such low baserates, the information value of the first of the three stratified mosaicplots is low.)

Similar results are seen for subjects who are between 45–64 years old at baseline: smokers have a higher probability of mortality than non-smokers. Almost all of the participants who were 65 or older at baseline died during the followup period, but the probability of dying was also slightly higher for smokers than non-smokers.

This example represents a classic example of *Simpson's paradox*[15], where overall smoking appears to be "protective", but within each age group smokers have a higher probability of dying than non-smokers.

Smoking is "bad" within each of the subgroups of age, while smoking is "good" overall.

[15] EH Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13:238–241, 1951

How can this be happening? Figure 3 and Table 3 help us to disentangle these relationships.

Not surprisingly, we see that mortality rates are highest for the oldest subjects.

We also observe that there is an association between age group and smoking status, as displayed in Figure 4 and Table 4.

Smoking is also associated with age, with those between the ages

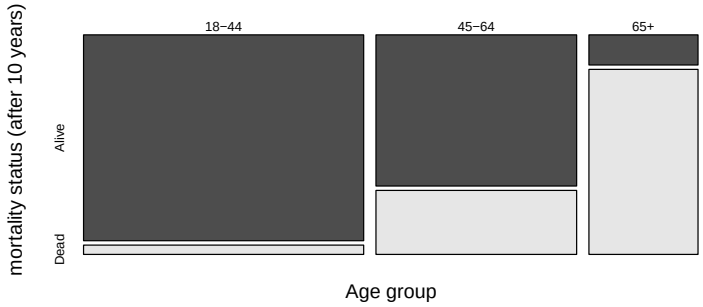


Figure 3: Mosaicplot representing the association between mortality and age in the Whickham study

Age group	Alive	Dead
18-44	597 (95.7%)	27 (4.3%)
45-64	314 (70.2%)	133 (29.8%)
65+	34 (14.0%)	209 (86.0%)

Table 3: Distribution of age group and mortality status (after 10 years) from the Whickham study

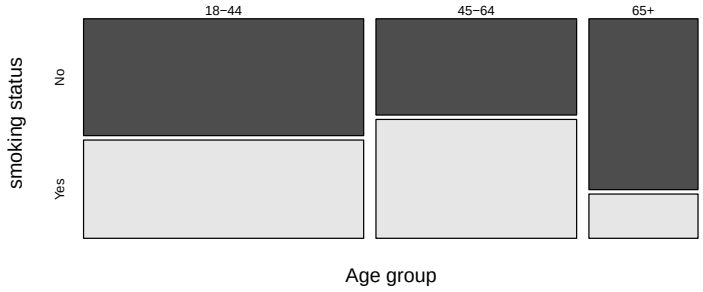


Figure 4: Mosaicplot representing the association between smoking status and age in the Whickham study

Age group	Non-smoker	Smoker
18-44	339 (54.3%)	285 (45.7%)
45-64	200 (44.7%)	247 (55.3%)
65+	193 (79.4%)	50 (20.6%)

Table 4: Distribution of age group and smoking status (at baseline) from the Whickham study

of 45 to 64 most likely to have been smokers at baseline.

After controlling for age, smokers have a higher rate of mortality than non-smokers in this study.

Simple methods such as stratification can allow students to think beyond two dimensions and reveal effects of confounding variables. Introducing this thought process early on helps students easily transition to analyses involving multiple explanatory variables.

SAT scores and teacher salaries

Consider an example where statewide data from the mid-1990's are used to assess the association between average teacher salary in the state and average SAT (Scholastic Aptitude Test) scores for students [16] [17]. These high stakes high school exams are sometimes used as a proxy for educational quality. Figure 5 displays the (unconditional) association between these variables. There is a statistically significant negative relationship ($\hat{\beta_1}$=-5.54 points, p = 0.001). The model predicts that a state with an average salary that is one thousand dollars higher than another would have SAT scores that are on average 5.54 points lower.

[16] DL Guber. Getting what you pay for: the debate over equity in public school expenditures. *Journal of Statistics Education*, 7(2), 1999

[17] NJ Horton. Challenges and opportunities for statistics and statistical education: Looking back, looking forward. *The American Statistician*, 69(2): 138–145, 2015

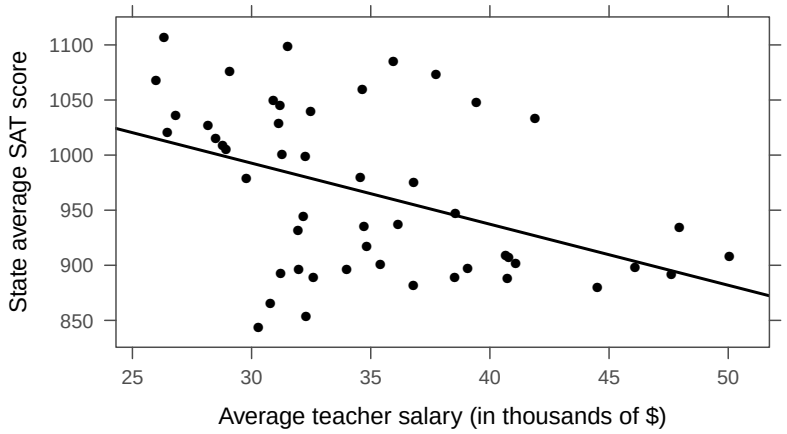


Figure 5: Association of state average teacher salary with average SAT score

But the real story is hidden behind one of the "other factors" that we warn students about but do not generally teach how to address! The proportion of students taking the SAT varies dramatically between states, as do teacher salaries. In the midwest and plains states, where teacher salaries tend to be lower, relatively few high school students take the SAT exam. These are typically the top students who are planning to attend college out of state, while many others take

the alternative standardized ACT test. For each of the three groups of states defined by the fraction taking the SAT, the association is non-negative. The net result is that the fraction taking the SAT is a confounding factor.

This problem is a continuous example of Simpson's paradox. Statistical thinking with an appreciation of Simpson's paradox would alert a student to *look for* the hidden confounding variables. To tackle this problem, students need to know that multivariable modeling exists (but not all aspects of how it can be utilized).

Within an introductory statistics course, the use of stratification by a potential confounder is easy to implement. By splitting states up into groups based on the fraction of students taking the SAT it is possible to account for this confounder and use bivariate methods to assess the relationship for each of the groups.

The scatterplot in Figure 6 displays a grouping of states with 0-22% of students ("low fraction", top line), 23-49% of students ("medium fraction", middle line), and 50-81% ("high fraction", bottom line). The story is clear: there is a positive or flat relationship between teacher salary and SAT score for each of these groups, but when we average over them, we observe a negative relationship.

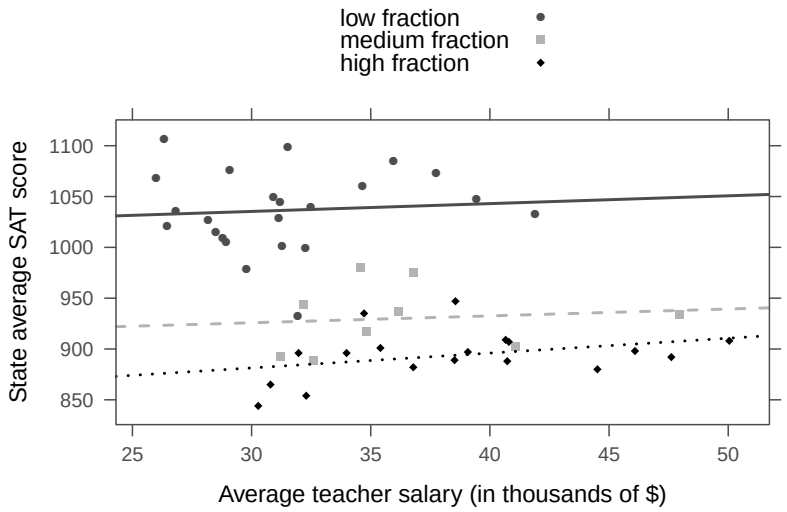


Figure 6: Association after accounting for the fraction of students taking the SAT in that state

Further light is shed via a matrix of scatterplots (see Figure 7): we see that the fraction of students taking the SAT is negatively associated with the average statewide SAT scores and positively associated with statewide teacher salary.

It's important to have students look for possible confounding factors when the relationship isn't what they expect, but it is also important when the relationship is what is expected. It's not always

Another natural approach to understanding of confounding and multivariate thinking is multiple regression. While this is not a traditional topic included in introductory statistics, an increasing number of textbooks and courses are incorporating the basic principles (often purely as a descriptive summarization of the data). In a multiple regression model that controls for the fraction of students taking the SAT variable, the sign of the slope parameter for teacher salary flips from negative to positive.

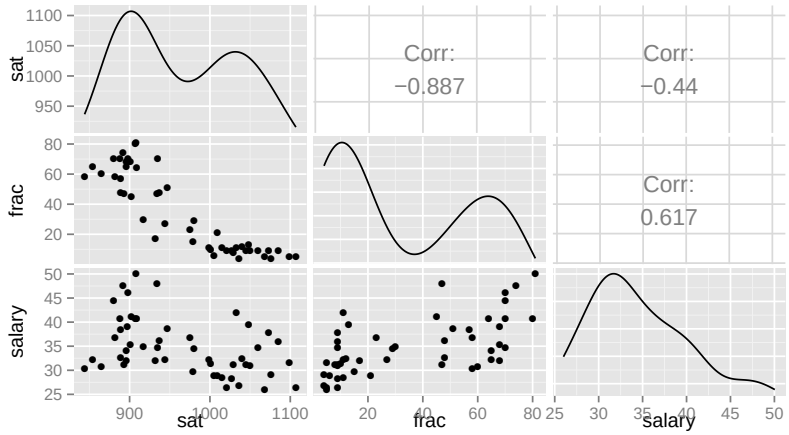


Figure 7: Matrix of scatterplots for the state-level SAT data

possible to stratify by factors (particularly if important confounders are not collected).

Multivariable thinking is critical to make sense of the observational data around us. This type of thinking might be introduced in stages:

1. learn to identify observational studies,
2. explain why randomized assignment to treatment improves the situation,
3. learn to be wary of cause-and-effect conclusions from observational studies,
4. learn to consider potential confounding factors and explain why they might be confounding factors,
5. use simple approaches (such as stratification) to address confounding

If students do not have exposure to simple tools for disentangling complex relationships, they may dismiss statistics as an old-school discipline only suitable for small sample inference of randomized studies.