Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report
DRAFT February 2016

Committee:

Robert Carver (Stonehill College), Michelle Everson (The Ohio State University), John Gabrosek (Grand Valley State University), Ginger Holmes Rowell (Middle Tennessee State University), Nicholas Horton (Amherst College), Robin Lock (St. Lawrence University), Megan Mocko (University of Florida), Allan Rossman (Cal Poly – San Luis Obispo), Paul Velleman (Cornell University), Jeffrey Witmer (Oberlin College), and Beverly Wood (Embry-Riddle Aeronautical University)

Contents

APPENDIX A: Evolution of Introductory Statistics and Emergence of Statistics Education Resources

Transformation 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 to prospective scientists who were still completing their degrees. 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 statistics still had to lean heavily on probability for its legitimacy.

During the late 1960s and early 1970s, John Tukey's ideas of exploratory data analysis launched the "data revolution" in the beginning statistics curriculum, freeing certain kinds of data analysis from ties to probability-based models. Analysis of data began to acquire status as an independent intellectual activity that did not require hours chained to a bulky mechanical calculator. Computers later expanded the types of analysis that could be completed by learners.

Two influential books appeared in 1978: *Statistics*, by David Freedman, Robert Pisani, and Roger Purves, and *Statistics: Concepts and Controversies*, by David S. Moore. These textbooks were distinctive in focusing almost exclusively on statistical concepts rather than statistical methods. They were aimed at a broad audience of consumers of statistical information rather than for students who need to learn to conduct statistical analyses. Then in the 1980s more and more introductory textbooks on statistical methods included a focus on concepts and real data[12].

The evolution of content has been paralleled by other trends. One of these is a striking and sustained growth in enrollments. Statistics from three groups of students illustrate the growth:

- At two-year colleges, according to the Conference Board of the Mathematical Sciences[13], statistics enrollments grew from 27% the size of calculus enrollments in 1970 to 74% in 2000 and exceeded calculus by 2010.
- Also from the CBMS survey, enrollments in elementary statistics courses at four-year institutions were up 56% in math departments and 50% in statistics departments from 2005 to 2010.
- The Advanced Placement exam in statistics was first offered in 1997 when 7,500 students took it, more than in the first offering of an AP exam in any subject at that time. More than four times as many students were taking the exam and by 2015, there were over 195,000 students testing[14].

[12] Cobb, G.W. (1987), "Introductory Textbooks: A Framework for Evaluation," *Journal of the American Statistical Association*, 82(397), 321-339, DOI:10.1080/01621459.1987.10478436
[13] http://www.ams.org/profession/data/cbms-survey/cbms
[14] http://research.collegeboard.org/programs/ap/data/participation/ap-2015

The democratization of introductory statistics has broadened and diversified the backgrounds, interests, and motivations of those who take the course. Statistics is no longer reserved for future scientists in narrow fields but is now 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 of today's beginning statistics courses can no longer assume that students are quantitatively skilled and adequately motivated by their career plans.

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 teachers can effectively help them learn.

Influential Documents on the Teaching of Statistics

As part of the Curriculum Action Project of the Mathematics Association of America (MAA), George Cobb coordinated a focus group about important issues in statistics education. The 1992 report was published in the MAA volume *Heeding the Call for Change*[15]. It included the following recommendations for teaching introductory courses:

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. The importance of data production.
- The omnipresence of variability.
- The quantification and explanation of variability.

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, automate calculations and graphics.

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.

[15] Cobb, G.W. (1992), "Teaching Statistics," in *Heeding the Call for Change: Suggestions for Curricular Action* (MAA Notes No. 22), 3–43. Washington, D.C.: The Mathematical Association of America.

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). Cobb's focus group evolved into the joint ASA/MAA Committee on Undergraduate Statistics. A growing body of statistics educators were implementing the recommendations and actively sharing their experiences with peers, often through projects and workshops funded by the National Science Foundation (NSF).

In the late 1990s, Joan Garfield led an NSF-funded survey[16] to explore the impact of this educational reform movement. 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 were included. The responses were encouraging: many reported increased use of technology, diversification of assessment methods, and successful implementation of active learning strategies.

The American Statistical Association funded a strategic initiative to create a set of Guidelines for Assessment and Instruction in Statistics Education (GAISE) at the outset of the 21st century. This was a two-part project that resulted in the publication[17] of *A Pre-K–12 Curriculum Framework*[18] and the original 2005 *College Report* that expanded upon the recommendations from the Cobb Report to address technology and assessment. These two reports have had a profound effect on the practice of teaching statistics and on the training of statistics educators at all levels.

Since the GAISE publications, the widespread adoption of the *Common Core State Standards*[19] has both strengthened the status of statistics as an academic necessity and challenged the content of a first collegiate course in statistics. The arrival of students from high school already exposed to topics formerly taught only in a college course (e.g., probability, exploratory data analysis, measures of center and spread, basic ideas of inference) is a shift as profound as the 1970s arrival of students without strong quantitative skills. New technology allowed the restructuring of the 20th century curriculum to include focus on concepts rather than computation as an adaptation to the new type of student. Forty years later, the 21st century curriculum has the opportunity to build on a broader foundation of prior knowledge that leaves room to delve deeper and farther than ever before possible.

[16] Garfield, J. (2000), *An evaluation of the impact of statistics reform: Final Report.* National Science Foundation (REC-9732404).
[17] amstat.org/education/gaise
[18] Christine Franklin (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-k–12 curriculum framework.* American Statistical Association: Alexandria, VA
[19] http://www.corestandards.org/

The Emergence of Statistics Education Research and Resources

Even before the publication of the original GAISE College Report, distinctions between mathematics education and statistics education were being made[20]. The connection between the disciplines, however, remains important and interesting to both mathematicians and statisticians. The American Statistical Association (ASA) maintains joint committees with the Mathematical Association of America (MAA), the American Mathematical Association of Two-Year Colleges (AMATYC), and the National Council of Teachers of Mathematics (NCTM).

The Statistical Education Section is one of the oldest sections within the ASA, founded in 1948, originally focused on the education of professional statisticians. The current mission statement of the Section includes advising the Association on educational elements in communication with non-statistical audiences, promoting reach and practice in statistical education; supporting the dissemination of development/funding opportunities, teaching resources, and research findings in statistical education; and, improving the pipeline from K-12 through colleges to statistics professionals[21].

In 2014 the ASA and MAA jointly endorsed a set of guidelines[22] for those teaching an introductory statistics course. These guidelines stipulate that instructors of statistics ideally meet the following qualifications:

- Experience with data and appropriate use of technology to support data analyses
- Deep knowledge of statistics and appreciation for the differences between statistical thinking and mathematical thinking
- Understanding the ways statisticians work with real data and approach problems and experiencing the joys of making discoveries using statistical reasoning
- Mentoring by an experienced statistics instructor for an instructor unfamiliar with the data-driven techniques used in modern introductory statistics courses

These guidelines recommend minimum qualifications for teaching introductory statistics as consisting of:

- Two statistical methods courses, including content knowledge of data-collection methods, study design, and statistical inference
- Experience with data analysis beyond material taught in the introductory class (e.g., advanced courses, projects, consulting, or research)

In 2000, the MAA founded a special interest group (SIGMAA) on Statistics Education. Their purpose is also four-fold: facilitate the exchange of ideas about teaching statistics, the

[20] Ben-Zvi, D., & Garfield, J. (2008). "Introducing the Emerging Discipline of Statistics Education," *School Science & Mathematics*, 108(8), 355-361. doi:10.1111/j.1949-8594.2008.tb17850.x

[21] http://community.amstat.org/statisticaleducationsection/home

[22] http://magazine.amstat.org/blog/2014/04/01/asamaaguidelines/

undergraduate statistics curriculum, and other issues related to providing effective/engaging encounters for students; foster increased understanding of statistics through publication; promote the discipline of statistics among students; and, work cooperatively with other organizations to encourage effective teaching and learning[23].

The AMATYC Committee on Statistics was founded in 2010 to provide a forum for the exchange of ideas, the sharing of resources, and the discussion of issues of interest to the statistics community. The Committee pays particular attention to activities that provide professional development and foster the use of the GAISE College report in the community college setting. It also serves as a liaison with faculty at four-year institutions and with other professional organizations for the purpose of resource sharing (see the AMATYC Statistics Resources Page[24]).

A 2006 charter established the Consortium for the Advancement of Undergraduate Statistics Education CAUSE which had grown out of a 2002 strategic initiative within the ASA. The mission of CAUSE is to support and advance undergraduate statistics education through resources, professional development, outreach and research. CAUSEweb.org serves as a repository for all of those areas. CAUSE also coordinates the US Conference on Teaching Statistics (USCOTS) which has been held in the spring of odd-numbered years since 2005. Since 2012, the electronic Conference on Teaching Statistics (eCOTS) has provided a virtual conference experience on even-numbered years.

The oldest conference for statistics educators, however, is sponsored by the International Association of Statistics Educators (IASE)[25], a section of the International Statistical Institute. The International Conference on Teaching Statistics (ICOTS) has been held every four years since 1982 at various global locations. The IASE also supports the *Statistics Education Research Journal* (SERJ), a peer-reviewed e-journal in publication since 2002.

Other refereed journals of interest to statistics educators include *Teaching Statistics*[26], the *Journal of Statistics Education*[27], and *Technology Innovations in Statistics Education*[28].

[23] http://sigmaa.maa.org/stat-ed/art1.html
[24] http://www.amatyc.org/?page=StatsResources
[25] http://iase-web.org/
[26] http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9639
[27] http://www.amstat.org/publications/jse/
[28] http://escholarship.org/uc/uclastat_cts_tise

APPENDIX B: Multivariable Thinking

The 2014 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 (ASA, 2014). (possible footnote: See also Wild's "On locating statistics in the world of finding out", http://arxiv.org/abs/1507.05982.) An introductory statistics course cannot cover these topics in depth, but it is important to expose students to them even in their first course (Meng, 2011). 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 describe simple examples where a third factor clouds the association between two other variables. Simple approaches (such as stratification) can help to discern the true associations. Stratification requires no advanced methods, nor even any inference, though some instructors may incorporate other related concepts and approaches such as multiple regression. These examples can help to introduce students to techniques for assessing relationships between more than two variables.

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 and examples of observational (or "found" data) that arise more commonly than results from randomized comparisons.

Smoking in Whickham

A follow-up study of 1,314 people in Whickham, England characterized smoking status at baseline, then mortality after 10 years (Appleton et al, 1996). The summary data are provided in the following table:

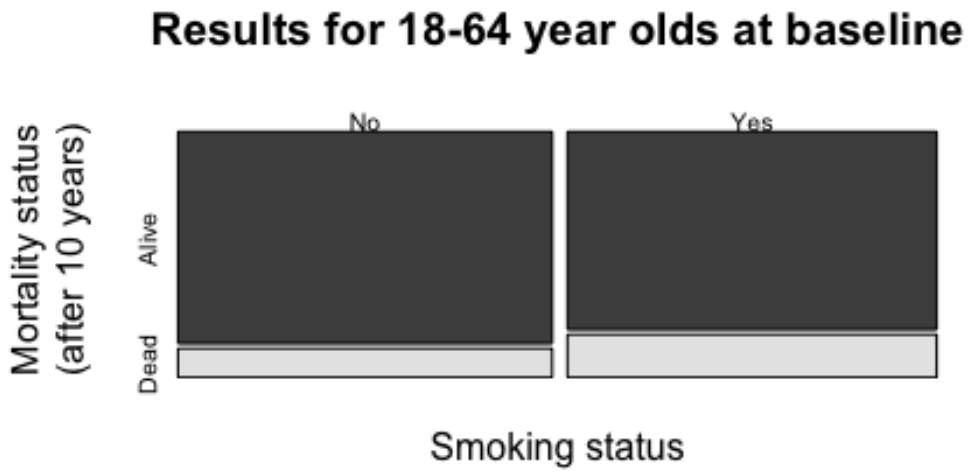
```
SMOKER   Alive        Dead
No       502 (68.6%)  230 (31.4%)
Yes      443 (76.1%)  139 (23.9%)
```

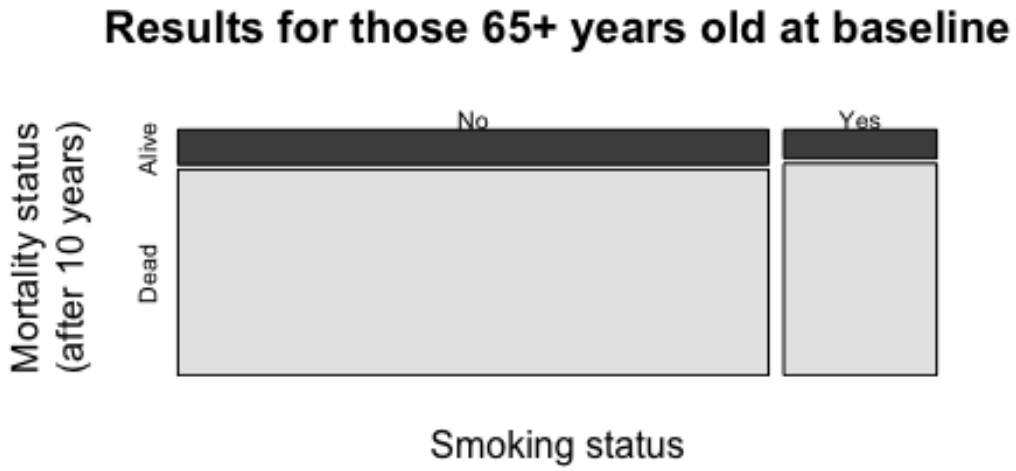
We see that the risk of dying is lower for smokers than for non-smokers, since 31.4% of the non-smokers died, but only 23.9% of the smokers did not survive over the ten year period. A graphical representation using a mosaicplot (also known as an *Eikosogram*) represents the cell probabilities as a function of area.

We note that the majority of subjects have survived, but that the number of the smokers who are still alive is greater 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 (older vs. younger). The following table and figure display the relationship between smoking and mortality over a 10-year period for two groups: those age 18-64 and subjects that were 65 or older at baseline.

Baseline age	SMOKER	Alive	Dead
18-64	No	474 (87.9%)	65 (12.1%)
18-64	Yes	437 (82.1%)	95 (17.9%)
65+	No	28 (14.5%)	165 (85.5%)
65+	Yes	6 (12.0%)	44 (88.0%)





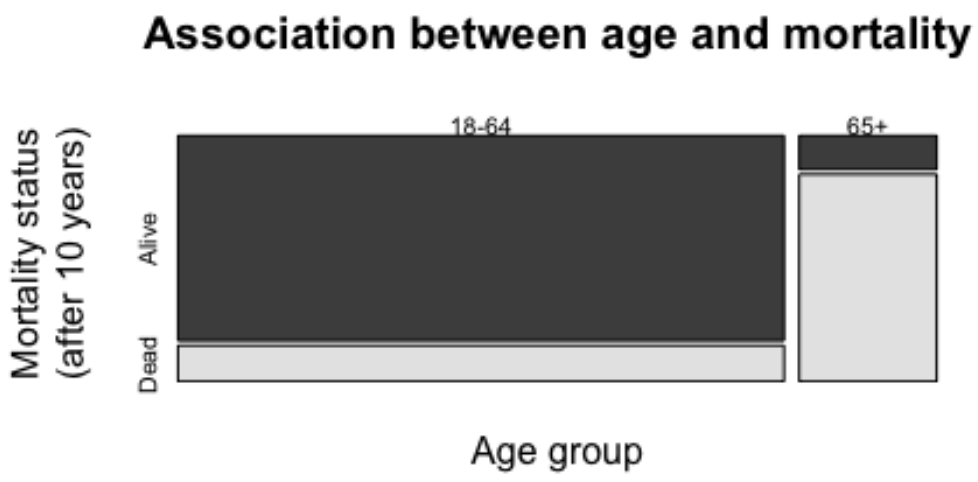
We see that mortality rates are low for the younger group, but the mortality rate is slightly higher for smokers than non-smokers (17.9% for smokers vs 12.1% for the non-smokers).

(possible foonote: Smoking is "bad" within both of the subgroups of age, while smoking is "good" overall.)

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.

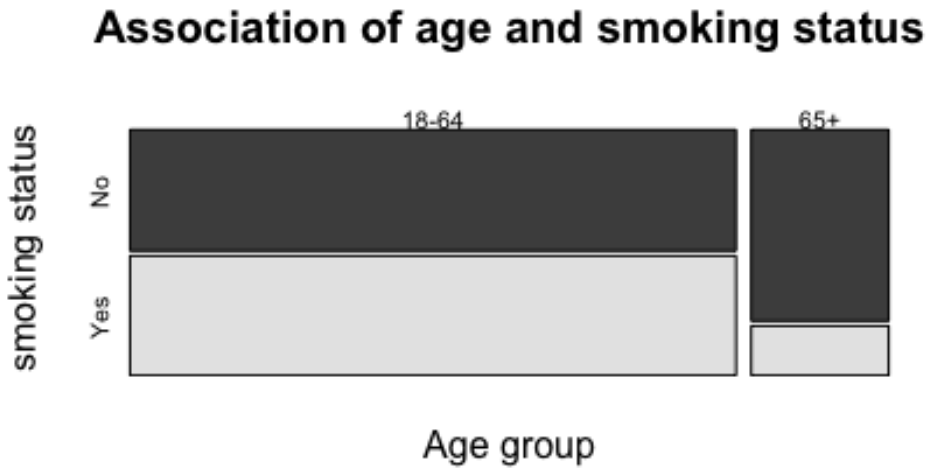
This example represents a classic example of *Simpson's paradox* (Simpson, 1951; Norton and Divine, 2015). For all of the subjects, smoking appears to be "protective", but within each age group smokers have a higher probability of dying than non-smokers.

How can this be happening? The following figure and table 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 the following figure and table.



```
Age group  Non-smoker    Smoker
18-64      539 (50.3%)   532 (49.7%)
65+        193 (79.4%)    50 (20.6%)
```

Smoking is associated with age, with younger subjects more likely to have been smokers at baseline.

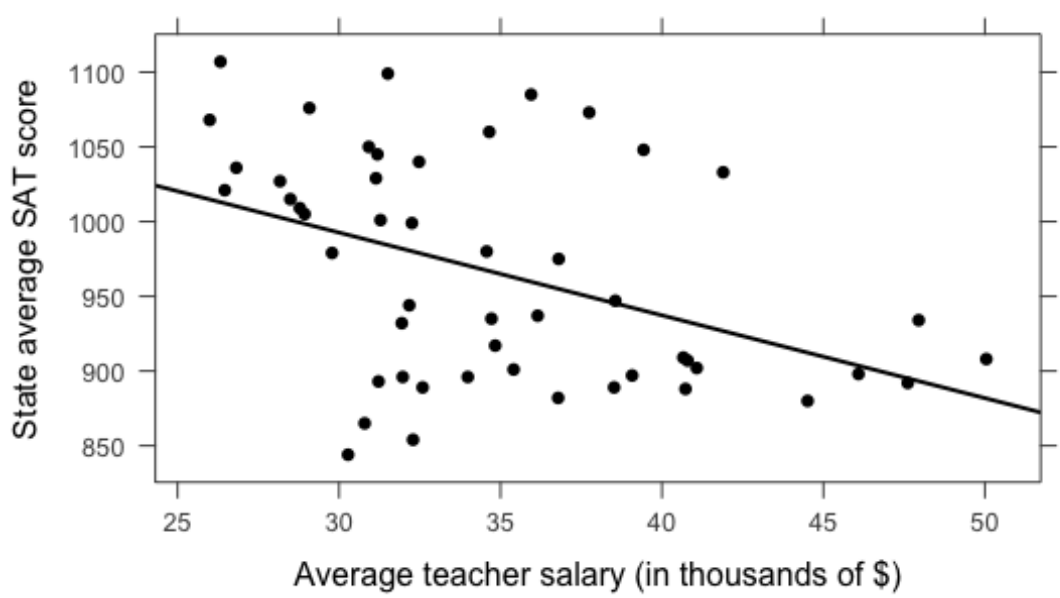
What should we conclude? After controlling for age, smokers have a higher rate of mortality than non-smokers in this study. This other factor is important when considering the association between smoking and mortality.

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 (Guber, 1999; Horton, 2015). These high stakes high school exams are sometimes used as a proxy for educational quality.

The following figure displays the (unconditional) association between these variables. There is a statistically significant negative relationship (β_1 hat =-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.



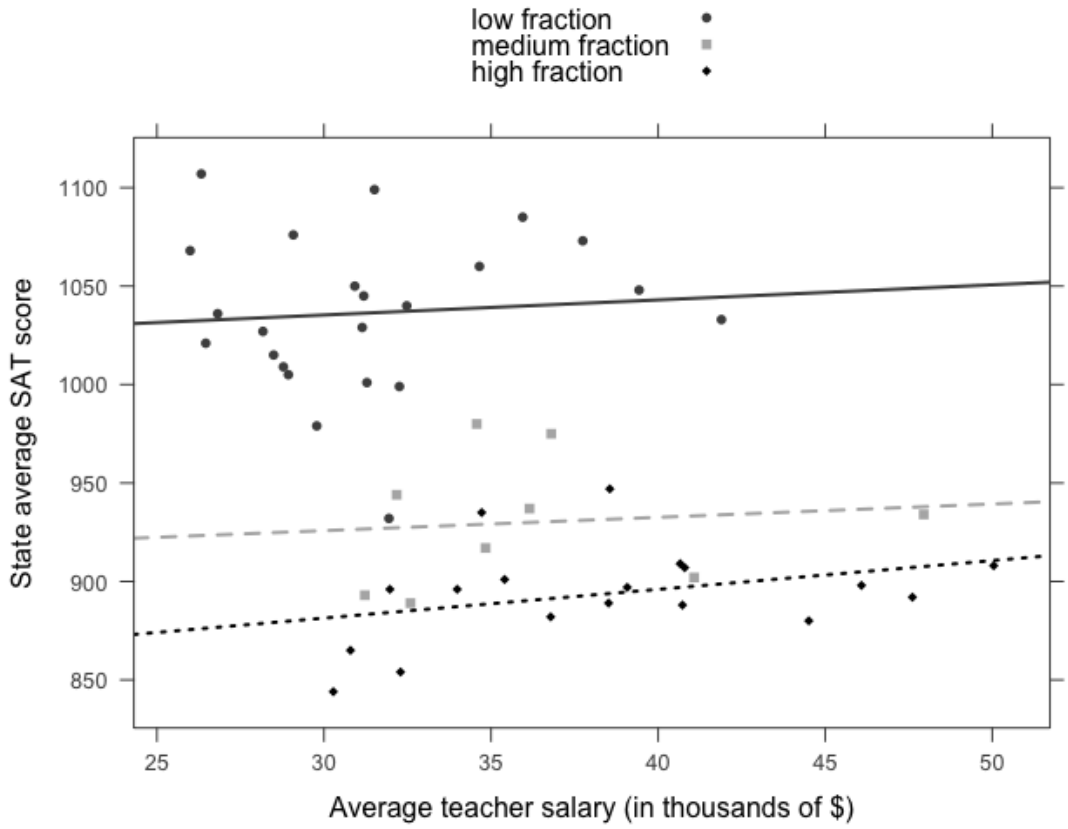
But the real story is hidden behind one of the "other factors" that we warn students about but do not generally teach how to address! The proportion of students taking the SAT varies

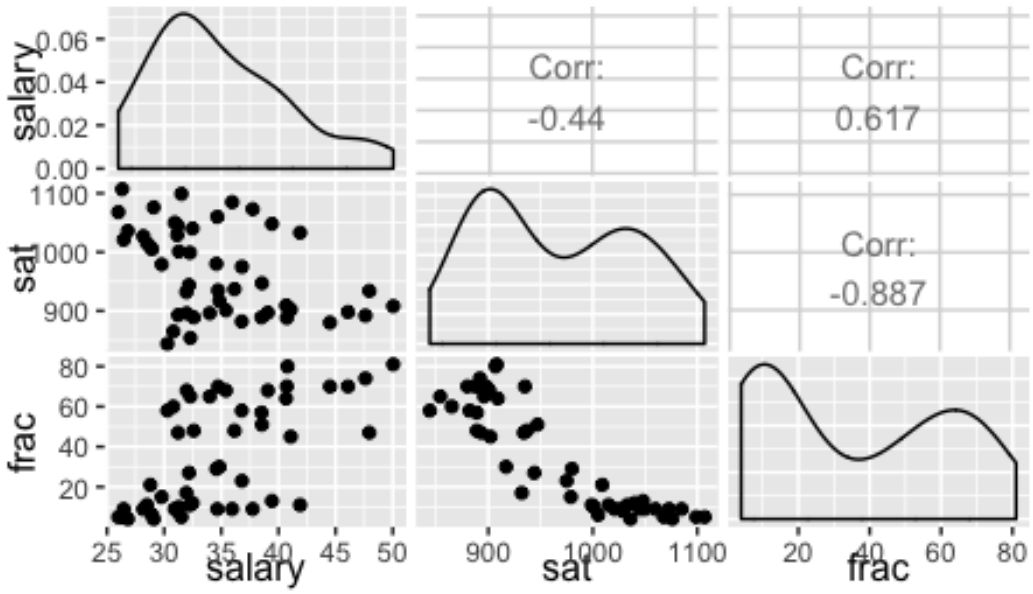
dramatically between states, as do teacher salaries. In the midwest and plains states, where teacher salaries tend to be lower, relatively few high school students take the SAT exam. Those that do are typically the top students who are planning to attend college out of state, while many others take the alternative standardized ACT test that is required for their state. 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 the next figure 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.





Further light is shed via a matrix of scatterplots (see the above figure): 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.

Recall that 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.

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

Multiple regression

The most common multivariable model is a multiple regression. Regression can be introduced as soon as students have seen scatterplots and thought about the patterns we look for in them. When students have access to a statistics program on a computer, they can find regression analyses for themselves. But even without computer access, they can learn about typical regression output tables. The point is to show students a model involving three (or more) variables and discuss some of the subtleties of such models. Here is one example.

Scottish hill races are scheduled throughout the year and throughout the country of Scotland (http://www.scottishhillracing.co.uk). The official site gives the current records (in seconds) for men and women in these races along with facts about each race including the distance covered (in km) and the total amount of hill climbing (in meters). Naturally, both the distance and the climb affect the record times. So a simple regression to predict time from either one would miss an important aspect of the races.

For example, the simple regression of time versus climb for women's records looks like this:

```
Response variable is:   Women's Record
R squared = 85.2%     R squared (adjusted) = 84.9%
s = 1126  with  70-2 = 68  degrees of freedom

Variable        Coefficient     SE(Coeff)     t-ratio    P-value
Intercept       320.528         222.2          1.44        0.1537
Climb             1.755           0.088       19.8       < 0.0001
```

We see that the time is greater, on average, by 1.76 seconds per meter of climb. The R^2 value of 85.2% assures us that the fit of the model is good with 85.2% of the variance in women's records accounted for by a regression on the climb.

But surely that isn't all there is to these races. Longer races should take more time to run. And although an R^2 of 0.852 is good, it does leave almost 15% of the variance unaccounted for.

It is straightforward for students to learn that multiple regression models work the same way as simple regression models, but include two or more predictors. Statistics programs fit multiple regressions in the same way as simple ones. Here is the regression with both `Climb` and `Distance` as predictors:

```
Response variable is:   Women's Record
R squared = 97.5%     R squared (adjusted) = 97.4%
s = 468.0  with  70 - 3 = 67  degrees of freedom
Variable      Coefficient  SE(Coeff)     t-ratio   P-value
Intercept     -497.656     102.8         -4.84     < 0.0001
Distance       387.628      21.45        18.1      < 0.0001
Climb            0.852       0.0621      13.7      < 0.0001
```

This regression model shows both the distance and the climb as predictors and has an R^2 of 0.975: a substantial improvement. More interesting, the coefficient of `Climb` has changed from 1.76 to 0.85. That's because in a multiple regression, we interpret each coefficient as the effect of its variable on y after allowing for the effects of the other predictors.

Closing thoughts

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

Multivariable models are necessary when we want to model many aspects of the world more realistically. The real world is complex and can't be described well by one or two variables. 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.

Simple examples are valuable for introducing concepts, but when we don't show students realistic models they are left with the impression that statistics is trivial and not really useful. This report recommends that students be introduced to multivariable thinking, preferably early in the introductory course and not as an afterthought at the end of the course.

References

Appleton, D. R., and French, J. M. and Vanderpump, M.P. (1996), "Ignoring a covariate: an example of Simpson's paradox", *The American Statistician*, 50(4):340-341.

American Statistical Association (2014), American Statistical Association Undergraduate Guidelines Workgroup. 2014. "2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science". Alexandria, VA: Author. Available at http://www.amstat.org/education/curriculumguidelines.cfm.

Guber, D. L. (1999), "Getting what you pay for: the debate over equity in public school expenditures", *Journal of Statistics Education*, 7(2).

Horton, N.J. (2015), "Challenges and opportunities for statistics and statistical education: Looking back, looking forward, *The American Statistician*, 69(2):138–145.

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

Norton, H. J. and Divine, G. (2015), "Simpson's paradox, and how to avoid it", *Significance*, 40-43.

Simpson, E. H. (1951), "The interpretation of interaction in contingency tables", *Journal of the Royal Statistical Society, Series B*, 13:238-241.

APPENDIX C: Activities, Projects, and Datasets

The GAISE College Report emphasizes the importance of students being actively engaged in their own learning. Activities, projects, and interesting datasets can help instructors engage students. In this appendix, we begin with a description of desirable characteristics of class activities. We provide examples of activities that illustrate a simple two quantitative variable data collection, a randomization test for the difference in two means, experimental design in a matched pairs study, and multivariable thinking. We conclude with examples of datasets and websites that house data.

Desirable Characteristics of Class Activities

In this appendix we focus on activities to be conducted in the classroom. Many of the desirable characteristics described are applicable to unsupervised activities conducted outside the traditional classroom setting.

Structure and timing…

- Learning Goals – An activity should have clear and attainable learning goals. The activity should build upon what students already know and lead students to discover or explore a statistical concept. Ideally, activities completed early in the course become scaffolding for concepts explored later in the course.

- Self-Contained and Complete – An activity should include all the important statistical concepts, necessary materials, and information from past class activities to complete the activity in a timely fashion.

- Beginning and Ending an Activity – The activity should begin with an overview and end with a summary of what is being done and why. This should include connections that build upon and extend statistical conceptual and methodological knowledge and application, how the statistical analysis helps to answer questions specific to the context of the activity, and what students are expected to learn from the activity.

Choosing data…

- Relevance – The activity should involve data about topics that interest students. Using real data makes data relevant to a wide variety of student majors. If real data are not used, then the activity should mimic a real-world situation. It should not seem like "busy work" to students. For example, if you use coins or cards to conduct a binomial experiment, explain real-world binomial experiments they could represent.

  Note: Student interests vary such that a dataset that might be interesting and relevant for one student may not be as interesting or relevant for another student. It is important to use a variety of datasets that speak to students from diverse backgrounds, majors, and interests. One way to gauge student

interest is to give the class an option of what dataset to work with in an activity. The choice could be made by student vote.

- Contextual Background – Students should read and be asked questions about the background that informs the context of the data. For example, if the data involves the number of friends a person has on Facebook, then students should read a brief background on some aspect of Facebook creation and or usage.

If the activity involves collecting/generating data…

- Design Decisions and Data Collection – Activities can include those that require class input into design and data collection and those that are more prescriptive. It is desirable that the class be involved in some of the decisions about how to conduct the activity when it is time-realistic and advances class learning objectives.

  Note: When students are involved in construction and implementation of design and data collection decisions, it is important that they invoke good design and data collection principles taught in the class. For instance, when designing an experiment, students should consider principles of good experimental design including randomization, replication, controlling outside factors, etc., rather than "intuitively" deciding how to conduct the experiment.

- Human Subjects Review – Most classroom data collection activities are exempt from the need for review by an Institutional Review Board (IRB). However, students should be made aware of the importance of review when collecting data, especially data on human subjects. Many students will work with an IRB in research methods courses in their own disciplines.

Working in groups…

- Team Work – Students can learn effectively from each other. While many students are drawn to working in teams, whether formal or informal, some students may resist working with peers. Because working effectively in teams is a highly valued skill in government, industry, and academia that can be practiced in the classroom, instructors should consider requiring some degree of team work in activities and projects.

  Note: Appendix F on Learning Environments includes a discussion of the use of and value of cooperative groups in the classroom.

- The Role of Groups in Design Decisions – It is sometimes better to have students work in teams to discuss how to design an 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 decisions. It depends on how difficult the issues to be discussed are and whether each team will need to carry out data collection in exactly the same way.

Sharing activities…

- Resources for Activities – There are numerous sources for activities including the Journal of Statistics Education (http://www.amstat.org/publications/jse/), STatistics Education Web (http://www.amstat.org/education/stew/), and Teaching Statistics (http://www.teachingstatistics.co.uk/).

- Sharing Activities with Other Instructors – For an activity to be easily usable and modifiable, the following characteristics are desirable: (1) quick data collection with low cost in time and resources, (2) available in a file format (such as Word) that makes modification by the instructor easy, and (3) includes a sample answer key for instructors.

Final thoughts…

- In our experience, students enjoy seeing their own data amongst their classmates' data. Activities that collect non-sensitive data from students, either inside class or outside class, perhaps through an online survey, provide this opportunity.

- The activity should be substantive, compelling, and, when possible, fun!

This is a list of desirable characteristics of class activities. This does not imply that an activity that does not meet every characteristic on this list is a poor activity. These characteristics are items to consider when creating, adapting, or using an activity.

Example Activities/Datasets

In-Class Data Collection/Analysis Activities

EXAMPLE ACTIVITY 1: LEG LENGTH AND STRIDE LENGTH

Exploring bivariate relationships is an important part of the introductory statistics course. In this activity, students investigate whether there is a relationship between the length of a person's leg and the number of steps required to walk a specified distance.

Materials Needed:

- Tape Measures

Procedure:

- Split class into groups of 3-4
- Have group measure each student's leg length from outside hip bone to floor
- Simultaneously, have each student walk a specified route. Same route for each student. Instruct students to silently count the number of steps they take as they walk the route.
- Have students enter data into a simple data collection form

Name	Sex (M or F)	Leg Length (inches) – measure right leg	Count of Steps

- Compile class data and use to illustrate scatterplots, correlation, and regression. It is likely that the relationship will be weak and negative.

Instructor Notes:

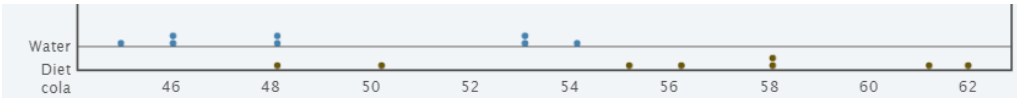
- The weak relationship lends itself to a discussion of what other variables might impact step count. Students usually identify that different people have different gaits (though they are unlikely to use the term gait). Gait analysis can be used to assess deviations from normal, especially if a person's baseline gait has been analyzed prior to an injury.
- A short distance (no more than a few hundred steps) is sufficient; data collection takes about 5 minutes. Data entry can be done on the spot or by the teacher between class sessions.

Example Activity 2: Randomization Test for a Difference in Means - Cola and Calcium

Setting: A study by Larson et al. (2010) examined the effect of diet cola consumption on calcium levels in women. A sample of 16 healthy women aged 18-40 were randomly assigned (eight to each group) to drink 24 ounces of either diet cola or water. Their urine was collected for three hours after ingestion of the beverage and calcium excretion (in mg) was measured. The researchers were investigating whether diet cola leaches calcium out of the system, which would increase the amount of calcium in the urine for diet cola drinkers. Low calcium levels are associated with increased risk of osteoporosis (Kahn and Laflamme, 2015).

Drink	Calcium
Diet Cola	48
Diet Cola	50
Diet Cola	55
Diet Cola	56
Diet Cola	58
Diet Cola	58
Diet Cola	61
Diet Cola	62
Water	45
Water	46
Water	46
Water	48
Water	48
Water	53
Water	53
Water	54

Results:



Diet Cola mean $\bar{x}_D = 56.0$ Water mean: $\bar{x}_W = 49.125$

The difference in means is: $\bar{x}_D - \bar{x}_W = 56.0 - 49.125 = 6.875$.

Key Question: Does this difference (6.875) provide *convincing* evidence that the mean amount of calcium excreted after drinking diet cola is higher than after water OR could this difference just be due to *random chance* (in assigning volunteers to the two groups)?

Approach: Simulate new samples generated by random chance and see how often we get a difference as large as (or larger than) what was observed in the original sample (6.875). We will do this first using a physical simulation (by hand), then switch to computer technology to automate the process.

Physical Simulation

1. Start with a sheet of paper that has the 16 calcium amounts from the experiment (such as the table above) and cut/tear the paper so that the numbers are separated from the diet cola/water groups and each value is on its own slip of paper. [Instructor alternative: Put the 16 calcium amounts on individual cards.]

2. Shuffle the slips/cards with calcium amounts and "deal" them randomly into two groups with 8 going to the diet cola group and 8 going to the water group.

3. Find the mean for each group and the difference in the two means.

   $\bar{x}_D =$ _____ (Diet cola) $\bar{x}_W =$ _____ (Water) $\bar{x}_D - \bar{x}_W =$ _____ (Difference)

   Is this difference bigger than the 6.875 from the original sample? ____

4. Look at some of the other simulated differences from your classmates How many of them are bigger than 6.875? [Instructor note: Perhaps draw a class dotplot of differences].

Simulation via technology

[*Instructor note:* Specific instructions below will depend on your technology. See the technology notes below for several options including StatKey (http://lock5stat.com/statkey), a Rossman/Chance applet (http://www.rossmanchance.com/applets), or the R package.]

5. Use technology to simulate the process you just did by hand – scrambling the 16 calcium values and reassigning them to diet cola & water groups.

   Which group got the smallest amount (45)? Diet Cola Water

   Which group got the largest amount (62)? Diet Cola Water

6. Put the difference in means for your simulated sample in the table below, then repeat to do four more simulations and record the difference in means for each simulation.

Simulated
$\bar{x}_D - \bar{x}_W$

7. Now use the technology to generate a thousand or more simulations. Look at a dotplot (or histogram) of the differences in means for all of these simulations. This is called a *randomization distribution* of the differences and shows what we might expect to see if there really is no difference in calcium excretion between the two groups.

   Where is your randomization distribution centered? _____

   Why does this make sense?

   Does it look like the difference from the actual sample (6.875) is in an unusual place in your randomization distribution?

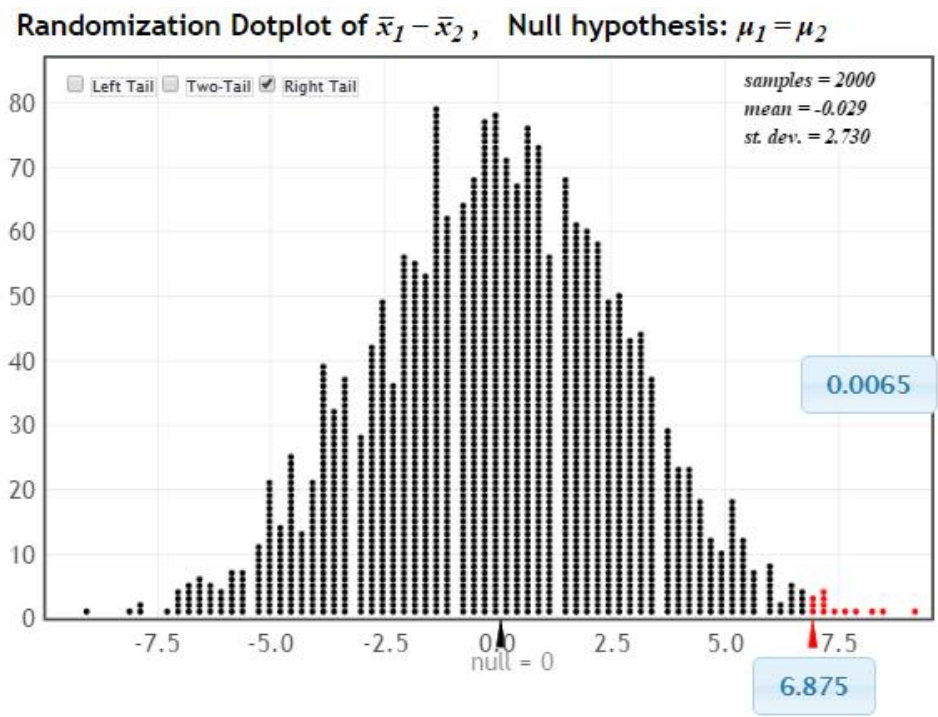
8. To quantify the last question, we will estimate a **p-value** as the proportion of all those random chance samples that have a difference in means as large as (or larger than) the original difference of 6.875. Use technology to estimate the p-value for your randomization distribution.

   What proportion of your randomization differences are 6.875 or larger?

   p-value = _____

9. Interpretation: What does this p-value tell you about the "significance" of the difference in the original sample? Does the difference look unusually large (indicating strong evidence that mean calcium excretion tends to be higher after drinking diet cola) or does the difference look more typical of what you would expect to see by random chance alone?

Instructor note: Here is a typical example (from StatKey) of a randomization distribution students might produce in this activity:



Technology notes for the Randomization Activity:

We provide three different technology options (each freely available) for doing the randomizations needed for parts 5-9 of the activity above.

<u>StatKey</u> (available at *http://lock5stat.com/statkey*)

- From the main StatKey page choose the Randomization "Test for Difference in Means".
- This dataset is already included in StatKey, so click on the drop down menu (labeled "Leniency and Smiles, just below the StatKey icon) to bring up a list of datasets and choose "Cola and Calcium excretion".
- Check that the data and summary statistics shown in the "Original Sample" graph match the data for this activity. Note: For data not already in StatKey, you can use the "Edit Data" button to copy/paste or enter your own data.
- Click on "Generate 1 Sample" to do a single randomization (Step 5 in the activity). The randomized data is displayed and summarized in the bottom right and the difference in means is plotted in the main dotplot at the left. Repeat this for several more randomizations (Step 6).
- Click on the "Generate 1000 Samples" a few times to get a better picture of the randomization distribution (Step 7).
- To find what proportion of the randomizations gave differences as large as the original difference (Step 9) choose the "Right Tail" option, click on the blue box that appears on the horizontal axis, and change the endpoint to 6.875 (the difference in the original sample). The p-value is shown in the box about the right tail.

<u>RossmanChance Applet</u> (available at *http://www.rossmanchance.com/applets*)

- Under "Statistical Inference", choose the option for "two means" (under "Randomization test for quantitative response").

- You need to copy/paste or enter the data from the table in the activity above to replace the default data in the applet. Also, the applet wants the group identifiers to be single words so delete the spaces to change "Diet Cola" to "DietCola" for the first 8 cases.
- Click on "Use Data" and check that the plot and summary statistics match the original sample.
- Click the box next to "Show Shuffle Options" to bring up the controls for the randomizations.
- Leave the number of Shuffles at 1 and click on 'Shuffle Responses" to generate one randomization (Step 5). The shuffled difference is shown below the summary statistics and plotted to the right. Repeat for Step 6.
- Change the number of shuffles to a larger number (like 3000) and "Shuffle Responses" again to generate a histogram of the randomization distribution (Step 7).
- To find what proportion of the randomizations gave differences as large as the original difference (Step 9), fill in that value (6.875) in the box after "Count Samples" leaving the "Greater than ≥" alone. Click on the "Count" button to see the count and proportion.

R (downloadable from *http://www.r-project.org*)
Here is an R script for creating the randomization differences and seeing what proportion are as extreme as the 6.875 difference in the original sample. It uses the nice do( ) function from the *mosaic* package to generate repeated samples without needing a formal loop.

```
#Randomization test to compare two means

library(mosaic)        # Load the mosaic package
library(Lock5Data)     # Load a package with the ColaCalcium dataset
data("ColaCalcium")    # Load the dataset

mean(Calcium~Drink, data=ColaCalcium)    #Compare means for two groups

#Compare means when the Drink values have been randomly permuted
mean(Calcium~shuffle(Drink), data=ColaCalcium)

#collect such simulated means for both groups for 2000 simulations
manymeans=do(2000) * mean(Calcium~shuffle(Drink), data=ColaCalcium)

head(manymeans)        #See some of what the do( ) function collects

#Find the difference in means for each simulation
randomdiffs=manymeans$Diet.cola - manymeans$Water

dotPlot(randomdiffs,width=0.1)    #get a plot of the random differences

#find the proportion of simulations with differences as large as 6.875
sum(randomdiffs>=6.875)/2000
```

EXAMPLE ACTIVITY 3: COMPARING MANUAL DEXTERITY UNDER TWO CONDITIONS

(Adapted from Project 12.2, Utts and Heckard, 2007)

Included in this activity description is:

- An Overview of the Activity
- Suggestions for Design and Analysis
- Project Team Form

Overview of Activity

These instructions are for the teacher. Instructions for students are on the "Project Team Form." (below)

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

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

- *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 second hand*

Instructor Note: A variation is to have students do the task both with and without wearing a "surgical" or "food-service" glove instead of with the dominant and non-dominant hand. In that case you will need *N pairs of gloves*.

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 non-dominant hand. (Or, if you are using 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. (To add some context, you can state that each color bean represents an automotive part of a slightly different size – for example, a front door bolt and a back door bolt.) A comparison will be done after using dominant and non-dominant 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.)

Design and Analysis

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

1. What are the treatments? What are the experimental units?
2. 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.
   a. Blocking or creating matched-pairs
   b. Randomization of treatments to experimental units, or randomization of order of treatments
   c. Blinding or double blinding
   d. Control group
   e. Placebo
   f. Learning effect or getting tired
3. What is the parameter of interest?
4. What type of analysis is appropriate – hypothesis test, confidence interval or both? What numerical and graphical analyses are appropriate?

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 non-dominant 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?

Note: Example 4 below deals with multivariable thinking in data analysis. Study design is an example of multivariable thinking where different variables are controlled so that the relationship between variables of interest can be isolated. For the bean sorting experiment, the matched pairs design controls for student-to-student variability and randomizing the order of dominant/non-dominant hand controls for the learning effect.

Step 2: Divide into teams and carry out the experiment.

The Project 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 plot of the differences. Discuss whether the necessary conditions for any inferential analysis are met. Were there any outliers? If so, can they be explained? Compute the mean and standard deviation for the differences.

Step 4: Inference

Have each team find a confidence interval for the mean difference and conduct the hypothesis test.

Step 5: Reconvene the class and discuss conclusions

Instructor Notes on Design:

- On Step 1
  a. Blocking or creating matched-pairs - Each student should be used as a matched pair, doing the task once with each hand.
  b. Randomization of treatments to experimental units, or randomization of order of treatments - Randomize the order of which hand to use for each student.
  c. Blinding or double blinding - Obviously the student knows which hand is being used, but the time-keeper doesn't need to know.
  d. Control group - Not relevant for this experiment.
  e. Placebo - Not relevant for this experiment.
  f. 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 non-dominant hand. Time students to see how long it takes to separate the beans. The person timing can be blinded to the condition by not watching.

Instructor Notes on Analysis:

- *What is the parameter of interest?*
  Define the random variable of interest for each person to be a "manual dexterity difference" of

  d = *number of extra seconds required with non-dominant hand*
  = *time with non-dominant hand − time with dominant hand.*

  Define μ_d = population mean manual dexterity difference.

- *What are the null and alternative hypotheses?*
  $H_0 : \mu_d = 0$ and Ha: $\mu_d > 0$ (faster with dominant hand)
  To carry out the test, compute $t = \frac{\bar{d} - 0}{s_d / \sqrt{n}}$ then compare to the t-table or use technology to find the p-value.

- *Is a confidence interval appropriate?*
  Yes, a confidence interval will provide information about how much faster workers can accomplish the task with their dominant hands. The formula for the confidence interval is: $\bar{d} \pm t * \frac{s_d}{\sqrt{n}}$, where t^* is from the t-table with df = n -1, and s_d is the standard deviation of the difference scores.

Project Team Form

TEAM MEMBERS:

1. _____ 4. _____
2. _____ 5. _____
3. _____ 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 or cell phone timer. Times each person for the task.
   Recorder – records the results in the table below.
2. Choose who will go first. The R*andomizer* tells the person which hand to use first. Each person should complete the task once before moving to the 2nd 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 below.
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.

NAME:	Time for *non-dominant* hand.	Time for *dominant* hand.	*d* = difference = non-dominant − dominant hand

RESULTS FOR THE CLASS:

Record the data here:

Parameter to be tested and estimated is:

Confidence interval:
Hypothesis test – hypotheses and results:

EXAMPLE ACTIVITY 4: MULTIVARIABLE THINKING IN THE ANALYSIS OF DIAMOND PRICING

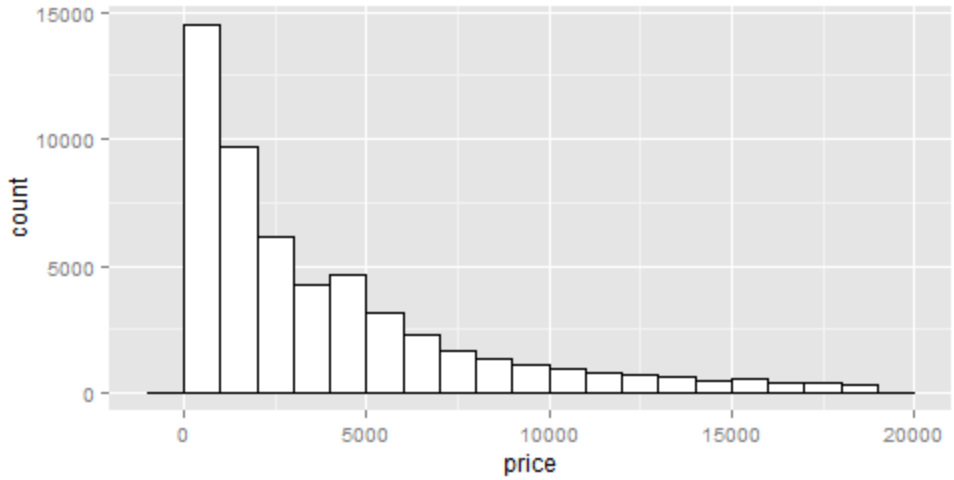
Goal: Provide students with experience investigating a dataset where the relationship between variables is conditional upon other variables. In this write-up, we use the R program and the ggplot2 package to analyze the data. We place any graphs that students should create in the body of the report. The same analysis could be done in SAS, SPSS or any other statistical software program.

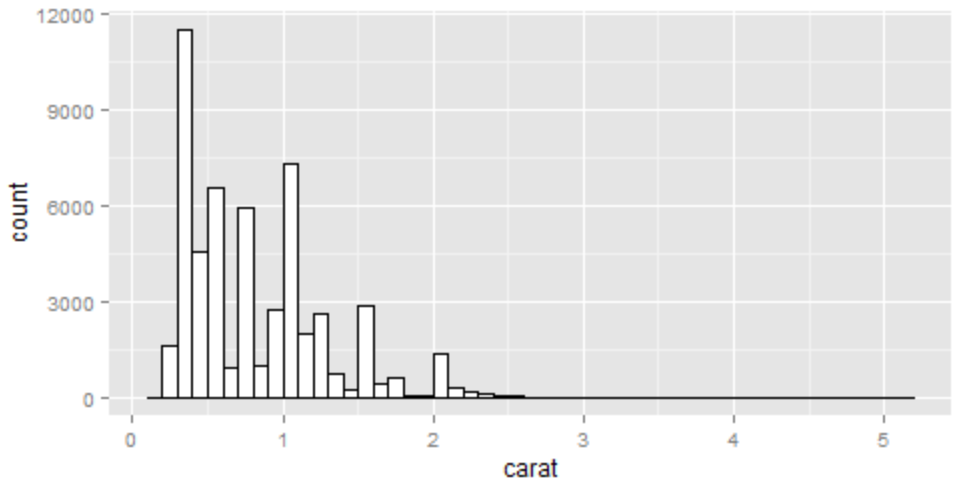
Data: This activity uses the diamonds dataset that is part of the ggplot2 R package (*http://www.inside-r.org/packages/cran/ggplot2/docs/diamonds*). The dataset includes information on 53,940 diamonds. There are ten variables measured on each diamond including price (in U.S. Dollars), cut (quality of the cut), clarity (a measurement of how clear the diamond is), color, and carat (weight from 0.2-5.01 carats). For a full description of the dataset open R and then enter code: help(diamonds).

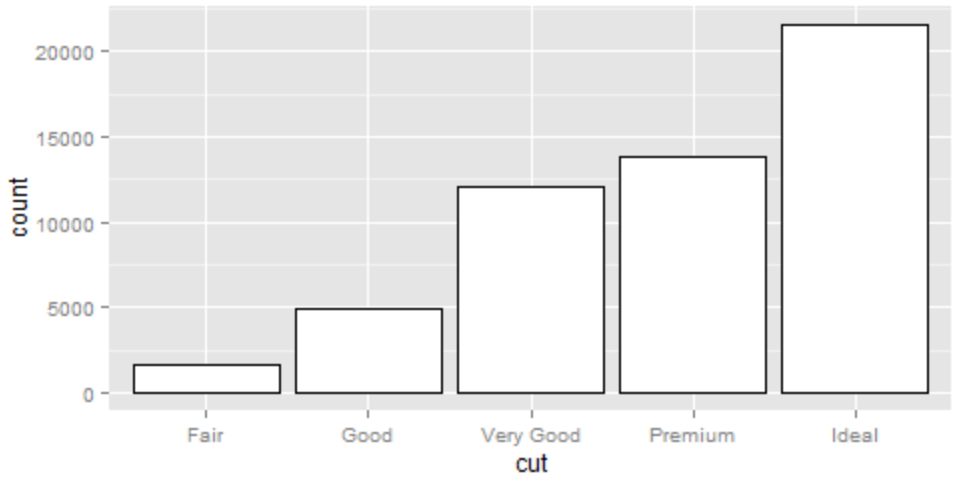
The Story: Diamond prices depend on the four C's of a diamond; cut, clarity, color, carat. It is pretty obvious that bigger diamonds cost more, or is it? In this activity you investigate the relationship between cut and price of diamonds using a dataset that includes 53,940 diamonds.

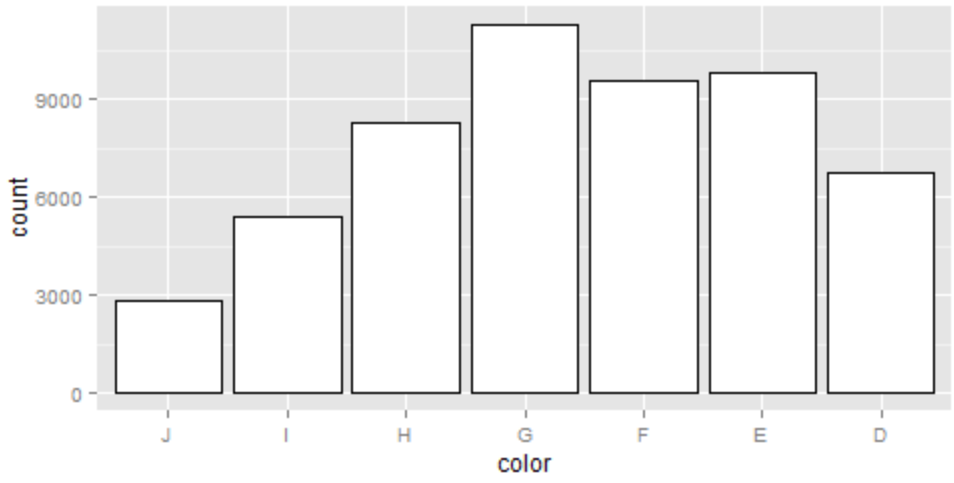
Part 1: Univariate Analysis

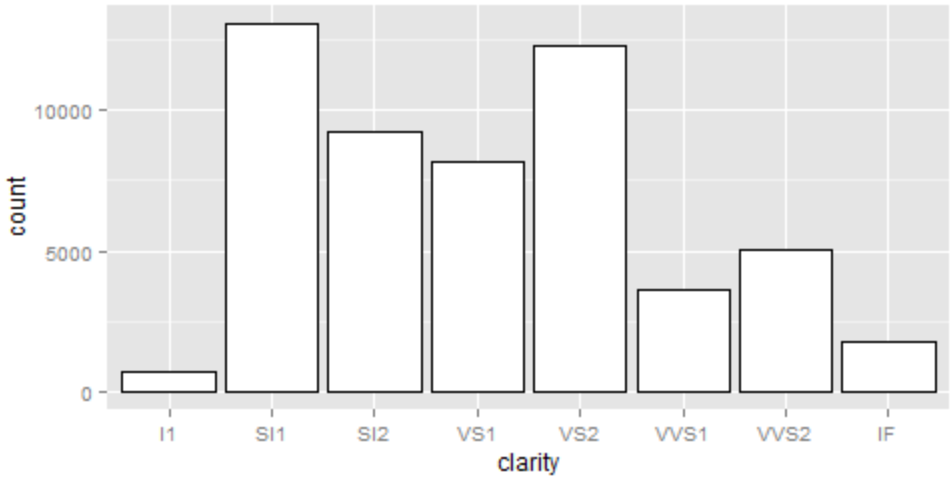
1. Make an appropriate graph for each of the five variables; price, cut, clarity, color, and carat. For the categorical variables, be sure that the categories are placed in a logical order from worst to best.




12000
9000
6000
3000
0
count
0
1
2
3
4
5
carat


20000
15000
10000
5000
0
count
Fair
Good
Very Good
Premium
Ideal
cut


9000
6000
3000
0
count
J
I
H
G
F
E
D
color

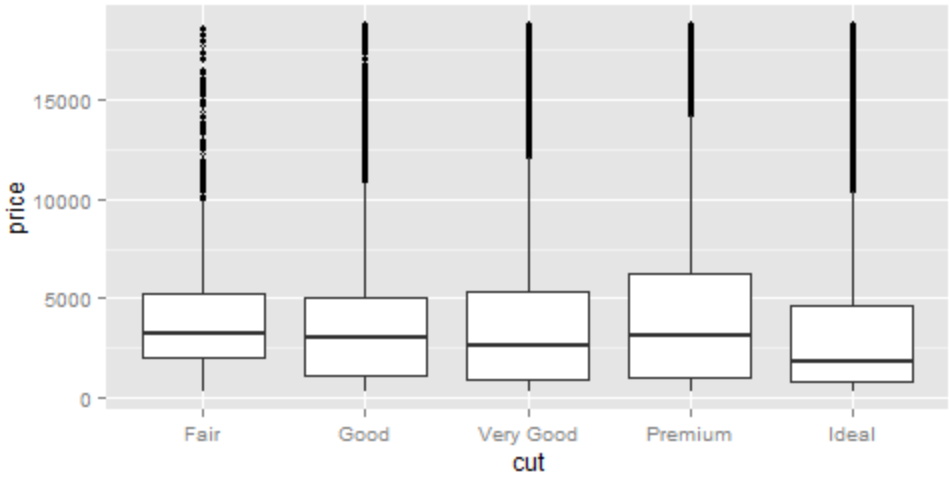


2. Describe any interesting features of the distribution of each variable.

   Students should point out that: (1) price is unimodal, peak from $0-$1000 and very skewed right; (2) carat is very choppy with a peak around 0.1-0.2 carats and then a smaller peak at around 1 carat and is skewed right; (3) most diamonds are of at least very good cut; (4) color has large variability with numerous diamonds of a lower color quality (left of G) and numerous diamonds of a higher color quality; and, (5) relatively few diamonds are of very high clarity (far right side of graph).

Part 2: Bivariate Analysis – Your goal is to investigate the relationship between a diamond's cut and the price of the diamond.

3. Make an appropriate graph to investigate the relationship between price and cut. Describe what you see. Is there anything surprising?



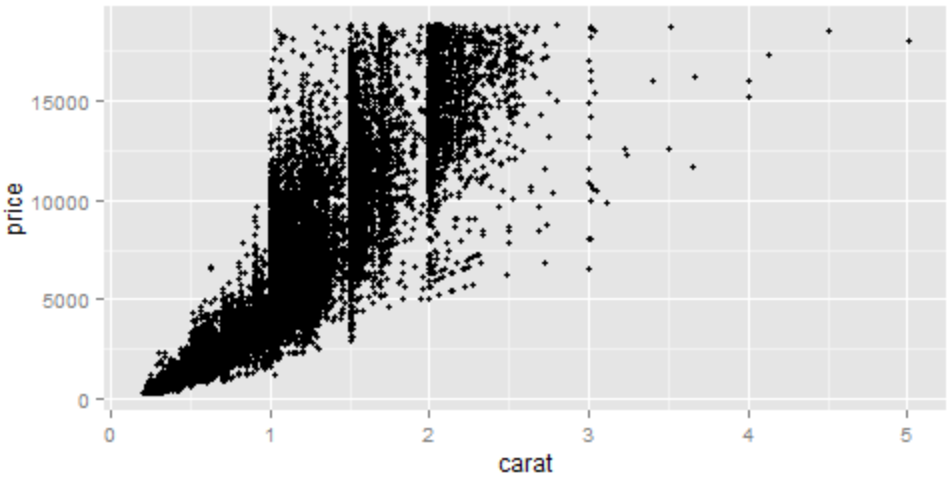
Students should point out that the median price for ideal cut diamonds is less than any of the other cuts of diamonds. This does not make sense because ideal cut is the highest quality cut possible.

Part 3: Multivariable Thinking

4. You should have noticed that ideal cut diamonds tend to have lower prices than any other cut. The median price for ideal cut diamonds is $1810, while fair cut diamonds have median price $3282. Brainstorm with a partner some ideas on why this might be true.

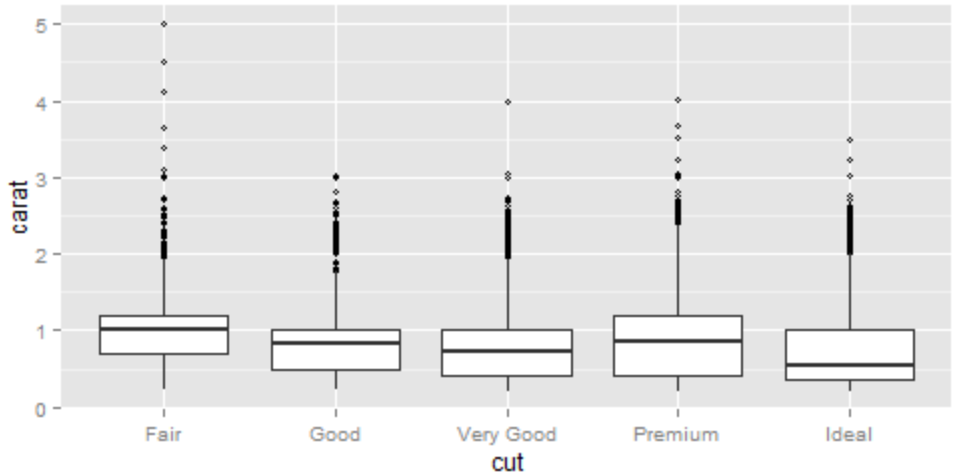
5. Now that you have brainstormed some ideas, let us see if the data can help us.

   a. First, make a scatterplot of price against carat. Describe what you see.



As expected there is a positive relationship between carat and price, with higher carat generally associated with higher price.

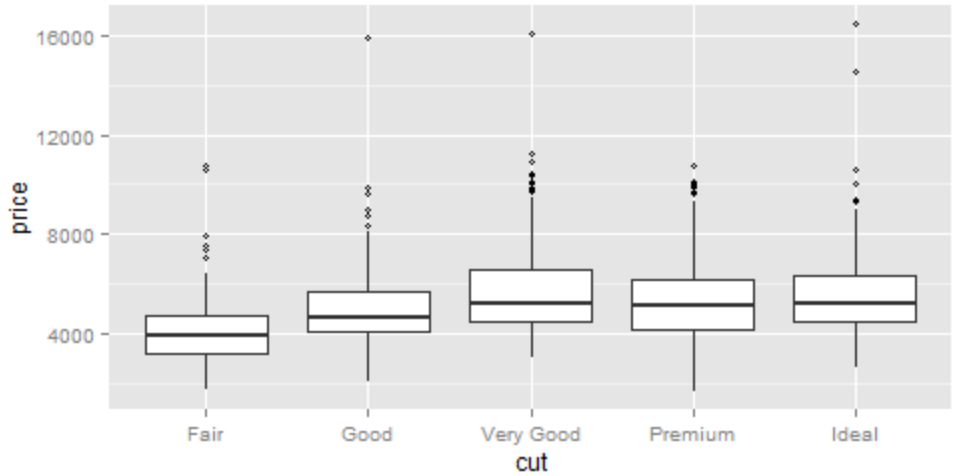
   b. Make an appropriate graph to investigate the relationship between carat and cut. Describe how cut is related to carat.



Fair cut diamonds tend to be much larger than ideal cut diamonds. Basically, it is very difficult to find a large diamond that can be cut as perfectly as necessary for an ideal cut designation.

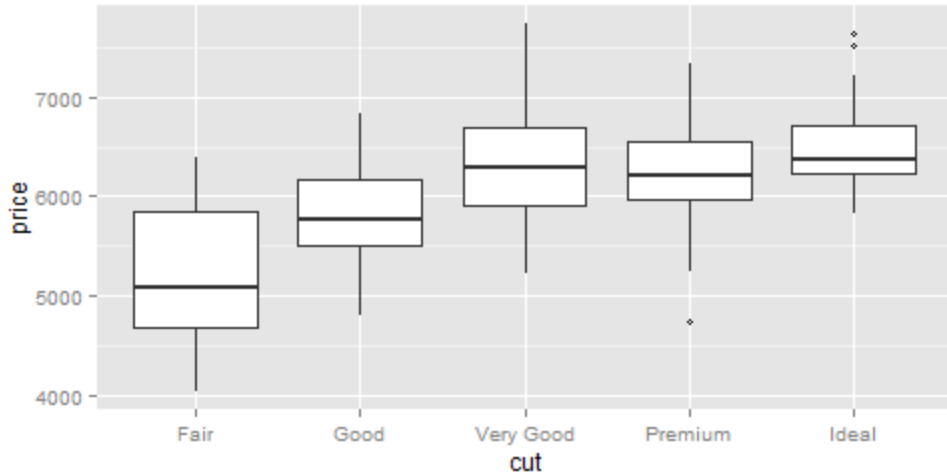
6. Now, let us look at only diamonds of size 1 carat.

   a. Below is a plot of price broken down by cut for these diamonds. What do you see?



For 1 carat diamonds, the price tends to increase as the cut quality improves. But, we have not accounted for the other variables color and clarity.

   b. Now let us take our 1 carat diamonds and only look at those of color = G or H and clarity = VS1 or VS2. There are 22 Fair, 53 Good, 64 Very Good, 82 Premium, and 27 Ideal diamonds meeting these conditions. What do you see in the plot?



When you control for carat, color, and clarity, then, as expected, fair cut diamonds are priced much less than ideal cut diamonds. The price of a diamond now seems to follow the cut.

7. What does this activity tell you about investigating the relationship between two variables?

Examples of Naked, Realistic and Real data

One of the core recommendations of this report and its predecessor is to "*Use real data.*" The next few small examples illustrate a continuum along a spectrum of "reality," starting with data having no context at all and progressing to data from an actual study designed to address a question of interest in a particular field. The task at hand (fit a least squares line) is the same in each case, and to help illustrate the distinctions, we have kept the number of data cases small in each situation. In practice, electronic access to data and technology for doing graphics and analysis frees us from restrictions of using such small datasets.

Naked data (not recommended)

Find the least squares line for the data below. Use it to predict Y when X=5.

X	1	2	3	4	6	8
Y	3	4	6	7	14	20

Critique: Made-up data with no context (not recommended). The exercise is purely computational with no possibility of meaningful interpretation.

Realistic data (better, but not ideal)

The data below show the number of customers in each of six tables at a restaurant and the size of the tip left at each table at the end of the meal. Use the data to find a least squares line for predicting the size of the tip from the number of diners at the table. Use your result to predict the size of the tip at a table that has five diners.

Diners	1	2	3	4	6	8
Tip	$3	$4	$6	$7	$14	$20

Critique: A context has been added which makes the exercise more appealing and shows students a practical use of statistics. The actual data values are made-up.

Real data (recommended)

The data below show the quiz scores (out of 20) and the grades on the midterm exam (out of 100) for a sample of eight students who took this course last semester. Use these data to find a least squares line for predicting the midterm score from the quiz score.

Assuming that the quiz and midterm are of equal difficulty this semester and the same linear relationship applies this year, what is the predicted score on the midterm for a student who got a score of 17 on the quiz?

Quiz	20	15	13	18	18	20	14	16
Midterm	92	72	72	95	88	98	65	77

Critique: Data are from a real situation that should be of interest to students taking the course, and the question asked is relevant for the situation.

Real Data, from a Real Study (even better)

In a study of honeybees, Seeley (2010) observed that scout bees do a "waggle dance" to help communicate the distance to a new nest site to bees back in the original nest. The table below shows the distance to the new site (in meters) and duration of the dance (in seconds) recorded for seven different scout honeybees. Use the data to find a least squares regression line and predict the distance to a new nest when a honeybee dances for 1.5 seconds.

Duration (seconds)	0.40	0.45	0.95	1.30	2.00	3.10	4.10
Distance (meters)	200	250	500	950	1950	3500	4300

Critique: Data come from a real study (with reference) to address a real research question.

Data (with Background and Stories) Available on the Web

EXAMPLE 1: AMES, IOWA REAL ESTATE DATA

The paper and dataset by DeCock (2011) describes sale of residential properties in Ames, Iowa from 2006 to 2010. The dataset contains 2930 observations of home sales and 80 variables. The data lends itself to a variety of analyses that can be done at the introductory statistics level (regressing y = sales price on x = square footage is one example) and at a more advanced modelling level.

While the paper does not contain a complete, ready-to-go activity, the author describes in detail potential uses of the dataset. He provides helpful hints to employ and potential pitfalls to avoid when using the dataset. It is quite easy to construct a simple activity that utilizes the data to illustrate concepts in regression.

The dataset meets many of the desirable characteristics listed previously, including:

- Data Relevance - The dataset is real data that represents actual real estate sales.
- Contextual Background - Understanding contextual background is important to understanding the data. Students need to be made aware of what different variables mean (the paper includes a documentation file with detailed variable descriptions) to be able to realistically model sales prices.
- Team Work - The richness of the dataset lends itself to use in a semester-long project that can best be done in teams working together. This is especially true if teams are tasked with developing a best regression model from the more than 70 potential predictor variables.

Teacher Hints:

1. The dataset is rich enough for many uses. In a regression modeling course students could be given the dataset and asked to find an appropriate model to predict sales price. In an introductory statistics course Sales Price can be summarized using basic numerical and graphical techniques. Pairs of variables can be used to discuss correlation, two-way tables, etc.
2. The introductory statistics instructor may want to work with a smaller subset of the variables so as not to overwhelm the students.
3. Instructors might use the Ames dataset and article as a template for collecting (or having students collect) similar data from the local area.

Note: Appendix D on Technology includes a further discussion of the Ames, Iowa real estate dataset.

EXAMPLE 2: U.S. ROAD LOCATION DATA

The paper by Stoudt et al. (2014) describes a lesson to randomly sample points in the continental United States, determine whether or not each point is within one mile of a road, and use the sample data to infer the proportion of the continental United States that is within one mile of a road. The paper requires use of the R programming language and Internet access.

As with all papers published on the STEW website, the paper includes a complete, ready-to-go activity.

The dataset meets many of the desirable characteristics listed previously, including:

- Data Relevance - The dataset is real data collected in real-time by the students to answer a question of importance to biologists, natural resource managers, and others concerned with providing habitat for plants and animals.
- Design and Data Collection – Students collect data to answer an important question. Using simple tools available on the Internet students are able to quickly and accurately collect data to address the question.
- Team Work – The data collection allows for students to collect roughly 20 data points in a class period. Students see that by pooling their data collection results they are rewarded with a more precise interval estimate of the proportion of the U.S. within one mile of a road.

Teacher Hints:

1. Students use the latitude, longitude coordinates of a point to do two things: (i) determine if the point falls within the continental U.S. and (ii) assuming the point is within the continental U.S., determine whether or not the point is within one mile of a road.
2. Students will generate points that do not fall in the continental U.S. (points in the Pacific Ocean, Atlantic Ocean, and Mexico are common). Because of this students will have unequal sample sizes unless instructed to continue generating points until 20 are within the continental U.S.
3. Students can edit their data file in Excel and then read back into R for the analysis if desired.

Websites with Data

There are numerous websites that have freely available data. Data formats vary, but it is usually simple to convert one of the datasets found at these sites to work with software available to the instructor. The complexity of the data and the amount of processing (i.e. data cleaning) to get the data ready for classroom use varies greatly. The list below provides a few places where an instructor can get data.

- Journal of Statistics Education (JSE) - http://www.amstat.org/publications/jse/ The Data Archive link includes many datasets. Each dataset includes the data (often in several formats) and a documentation file explaining variable names. Most of the datasets

include an accompanying JSE paper that describes the story and how to use the dataset in the classroom.

- Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) - https://www.causeweb.org/ The CAUSE website includes links to hundreds of locations for data. On the Home page in the upper right corner type "Datasets" in the Search field.
- New York City Open Data - https://data.cityofnewyork.us/ More than 1000 datasets on various aspects of life in the Big Apple. You can search a specific term or click to view all available datasets.
- Winner data - http://www.stat.ufl.edu/~winner/datasets.html Larry Winner from the Department of Statistics at the University of Florida has amassed hundreds of datasets. Each dataset includes a description.
- UMass-Amherst - http://www.umass.edu/statdata/statdata/ Datasets categorized by statistical topic.
- OzDASL - http://www.statsci.org/data/multiple.html Datasets categorized by statistical topic.
- Revolution Analytics R Data - http://mran.revolutionanalytics.com/documents/data/ - Links to various data sources.
- Kaggle - https://www.kaggle.com/ Website that hosts data analysis competitions. Many datasets here are quite large and very messy. Establishing a free account is necessary for access to the data.

Note: Appendix D on Technology includes other sources of data available on the web.

An additional source of data is from published papers. You can contact the author and journal asking for permission to use a dataset in teaching. Many authors and journals will grant permission for educational purposes and provide you with the dataset.

References

DeCock, D. (2011), "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project," *Journal of Statistics Education, 19(3)*. Available at http://www.amstat.org/publications/jse/v19n3/decock.pdf

Kahn, A. and Laflamme, M. R. (2015), "Calcium Deficiency Disease," *Heathline*. Available at www.healthline.com/health/calcium-deficiency-disease. Accessed Jan. 5, 2016.

Larson, N. S., Amin, R., Olsen, C., and Poth, M. A. (2010), "Effect of Diet Cola on Urine Calcium Excretion," *Endocrine Reviews – Endo 2010 Abstracts, 31(3),* S1070.

Seeley, T. (2010), *Honeybee Democracy*, Princeton University Press, Princeton, NJ, p. 128.

Stoudt, S., Cao, Y., Udwin, D., and Horton, N. J. (2014), "What Percent of the Continental US is Within One Mile of a Road?" *STatistics Education Web*. Available at http://www.amstat.org/education/stew/pdfs/PercentWithinMileofRoad.pdf

Utts, J. and Heckard, R. (2007), *Mind On Statistics – Instructor's Resource Manual*, Duxbury Press, Belmont, CA.

APPENDIX D: Examples of Using Technology

This appendix introduces different forms of technology that can be used to fulfill the GAISE recommendation to "Use of technology for developing conceptual understanding and analyzing data." Additionally, these forms of technology can help us to achieve some of the other GAISE College Report recommendations, such as stressing conceptual understanding, gathering and using real data, and fostering active learning.

Because technology changes so quickly and access to forms of technology varies from instructor to instructor, we will be highlighting how certain methods can be used to meet the recommendation; specific guidelines for particular brands or forms of technology will be avoided. A special effort has been made to keep the material current and relevant in light of constant advancements in technology.

When considering the use of technology in the classroom, the instructor should first start with the learning goal or learning objective and then carefully consider what forms of technology could be used to best meet that learning goal or objective. Next, the students in the classroom must be considered. What form of technology would students learn most quickly (if needed)? How much training would be necessary to allow students to seamlessly engage with the technology? It is important to pick technology that does not become an additional burden for students or that hinders them further from meeting learning goals.

In this appendix, we will focus on the following:

1. Interactive Applets
2. Statistics Software
3. Accessing Real Data online (Observational, Experimental, Survey)
4. Using Games and Other Virtual Environments
5. Real Time Response Systems

Using Interactive Applets

Interactive applets can be used to emphasize important statistical concepts without being encumbered by lots of calculations. [29] It's important to note, however, that free applets vary widely in terms of support, maintenance, and compatibility with evolving technology platforms. We recommend that instructors test applets on classroom systems each time they plan to use them in the classroom.

Applets are available that focus on a wide array of topics. To list just a few, there are applets that are available for using randomization and bootstrap techniques to conduct inference, for

[29] There are many places on the web that house statistical applets, and an internet search on a statistical concept can often generate a few openly available applets. Often times, applets are also available with certain textbooks.

discussing sampling distributions of the sample proportion and the sample mean, and for demonstrating the concept of "confidence" with confidence intervals. Additional applets can show the visualizing the effect of outliers on the simple linear regression equation as well as the effect of outliers on the values of measures of center and variability, as well as applets to simulate probabilities taken over the long run.

Applets can also be used in many ways, for example as the focus of a class demonstration, a homework assignment, a computer lab activity, a class project, a quiz or even as part of an exam. Additionally, applets can used by a single student at a time, as a team/partner activity or as a whole class.

Best Practices and Ideas found in Statistics Education Literature

- Applets work well with the query first method. This means that the students try to answer the conceptual questions first on their own and then again after using the applet.
    - To see more information, see the following article:
      delMas, R., Garfield, J., and Chance, B. (1999), "A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning", *Journal of Statistical Education* [online], 7, 3. Available at http://www.amstat.org/publications/JSE/secure/v7n3/delmas.cfm .

- When demonstrating with an applet that uses the concept of repeated sampling for randomization tests or bootstrapping techniques, first sample one at a time, and then stop to explain what is happening. You may need to take another sample and explain the process again. After the students appear to understand, you can then increase the number of samples to 1,000 or a higher value.
    - For an example of this process, see the following article:
      Lock Morgan, Lock, R. Lock, P.F. Lock, E. and Lock D. (2014), "StatKey: Online Tools for Bootstrap Intervals and Randomization Tests", *iCots9 Proceedings* [online]. Available at http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_9B2_LOCKMORGAN.pdf

- Pick applets that make it easier to focus on the concepts for introductory students and experience the entire investigative process. For example, the simulation should be similar to by hand method that the students could use to illustrate a concept, for example by using cards or coins. Additionally, the simulation should allow for easy transition to multiple types of inference, eg, from inference about difference of two independent proportions to difference in two independent means.
    - To see more about this and how simulation based inference is changing the modern curriculum, see the following articles:
      Rossman, A. and Chance, B. (2014), "Using Simulation-based inference for learning introductory-statistics", *WIREs Computational Stat* [online], Available at Doi:10.1002/wics1302 .

Tintle, et al. (2015), "Combating anti-statistical thinking using simulation-based methods through the undergraduate curriculum", *The American Statistician* [online], 69, 4. Available at http://dx.doi.org/10.1080/00031305.2015.1081619 .

Future Direction of Applets and Interactive Visualizations

In the past, the statistics education community has mostly relied on a handful of people and organizations to provide applets to help students build conceptual understanding. However, it is becoming easier and easier to design one's own statistical applets and other interactive visualizations, and soon, instructors will be able to use readily available open software to create their own public interactive visualizations. For example, Shiny, a software program, allows the user to create interactive visualizations (although some level of familiarity with R is needed).

- To see examples of some interactive visualizations: http://shiny.rstudio.com/gallery/
- For more information about writing code for these visualizations: http://www.r-bloggers.com/interactive-visualizations-with-r-a-minireview/

Even instructors who might not have the time or the desire to create their own visualizations might find the list of example visualizations under the Shiny gallery a way to bridge some of the gap between what students may traditionally see in an introductory course and the real data they may see outside of the classroom (e.g., movie reviews, airline data, bus route data, etc.).

- For further discussion of data for the modern student, see the following article: Gould, R., (2010), "Statistics and the Modern Student", *International Statistical Review*, [online], 78, 2, 297–315. Available at doi:10.1111/j.1751-5823.2010.00117.x

EXAMPLE 1: USING STATISTICAL APPLETS TO PERFORM RANDOMIZATION AND BOOTSTRAPPING TECHNIQUES

Since the first GAISE College Report, more and more introductory courses have been incorporating randomization and bootstrapping techniques into the curriculum, and one way in which these techniques can be incorporated into a course is by using statistical applets[30]. To see an example of this type of activity, please see the Appendix C of this report.

EXAMPLE 2: CREATING A STORY BOARD, VIDEO OR CARTOON ABOUT FINDINGS FROM AN APPLET

The following is a sample handout that can be used or modified in conjunction with an activity involving the use of an applet.

Student Handout

In class, we have been talking about confidence intervals for the population mean. What does the term "confidence" mean? A link to an applet about confidence intervals has been provided by your instructor. With your teammates, explore the applet and determine what it is trying to demonstrate.
You should be able to answer these three questions:

[30] Two free websites that can be used to do this are the Rossman/Chance website (http://www.rossmanchance.com/applets/) as well as the Lock 5 website (http://www.lock5stat.com/statkey/index.html). Some statistical software packages do this as well such as Statcrunch, R and JMP.

1.) If you were to take 100 different random samples and construct 100 95% confidence intervals for the population mean, would each of the intervals be exactly the same -- having exactly the same upper and lower bound? Explain your reasoning: why would or wouldn't they be the same?

2.) For these intervals, what do we know about the population mean in relation to those 100 confidence intervals?

3.) What does it mean to be 95% confident?

Now that you understand the simulation, do one of the following activities:

- Create a script for a two minute educational video that explains what is happening in the applet. The audience of the educational video should be people who have not
- Imagine that you have been given the opportunity to create a cartoon about statistics for the college newspaper. Create a cartoon demonstrating the concept of "confidence."
- Create a quick two minute video using a free online recording program that explains what is being demonstrated in the applet.

Inspiration:

Fewster, R. (2014), "Teaching statistics to Real People: adventures in social stochastics", *iCOTS 9 Proceedings [online]*. Available at http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_PL3_FEWSTER.pdf .

Turner, S. and Dabney, (2014), A. "A Story Based Simulation for Teaching Sampling Distributions." *Teaching Statistics [online]*, 37, 1. Available at http://onlinelibrary.wiley.com/doi/10.1111/test.12067/pdf .

Teaching Note:

It's important to think carefully about how long students should spend on this task. You may want to give them a timeline of one class period so they can focus more on the statistical concepts and less, for example, on perfecting a two minute recording.

Example 3: Exploring Misconceptions about Sampling Distributions by using an Applet

Here is example student handout that can be used or modified in conjunction with an applet that demonstrates the sampling distribution.

Student Handout

Before using the applet, answer the following questions. For these questions, write down what you think is the best answer. Please write these down in pen, because as you complete this activity you might find out that these ideas have been confirmed or are incorrect. If they are incorrect, it is important to see why they are incorrect and to identify them correctly later on. Seeing mistakes and misconceptions is important so that you remember it later on.

PART ONE

Sketch the graph of each of these.

Normal	**Sampling Distribution of the sample mean with n = 10, where the original population was Normal**	**Sampling Distribution of the sample mean with n = 100 where the original population was Normal**

1.) Describe the centers of the distributions. Are they the same? Different? In what way?
2.) Describe the variability of the distributions. Are they the same? Different? In what way?
3.) Describe the shape of the distributions. Are they the same? Different? In what way?
4.) Very briefly, explain your thinking: *why* do you anticipate these specific descriptions?

Sketch the graph of each of these.

Right Skewed	**Sampling Distribution of the sample mean with n = 10, where the original population was Right Skewed**	**Sampling Distribution of the sample mean with n = 100 where the original population was Right Skewed**

1.) Describe the centers of the distributions. Are they the same? Different? In what way?
2.) Describe the variability of the distributions. Are they the same? Different? In what way?
3.) Describe the shape of the distributions. Are they the same? Different? In what way?
4.) Very briefly, explain your thinking: *why* do you anticipate these specific descriptions?

Sketch the graph of each of these.

Uniform	**Sampling Distribution of the sample mean with n = 10, where the original population was Uniform**	**Sampling Distribution of the sample mean with n = 100 where the original population was Uniform**

1.) Describe the centers of the distributions. Are they the same? Different? In what way?
2.) Describe the variability of the distributions. Are they the same? Different? In what way?
3.) Describe the shape of the distributions. Are they the same? Different? In what way?
4.) Very briefly, explain your thinking: *why* do you anticipate these specific descriptions?

PART TWO

Now go to the interactive applet website[31] given to you by your instructor. Complete the table below.

	Shape	Mean	Standard deviation
Normal Distribution			
Sampling Distribution of the			

[31] There are many options for applets that may work. Some textbook publishers include applets with their textbook. Applets can also be found within some statistical packages like StatCrunch or even openly available on the internet. Since the resources of the instructors will vary and websites change rapidly, a website address is not listed. When choosing an applet, make sure that you pick an applet that allows the students to easily change the sample size and the population distribution. It should also allow you to show the results one sample at a time, a few samples at a time and then many samples at a time.

sample mean with n = 10 where the original population was Normal.			
Sampling Distribution of the sample mean with n = 100 where the original population was Normal			

1.) Describe the centers of the distributions. Are they the same? Different? In what way?
2.) Describe the variability of the distributions. Are they the same? Different? In what way?
3.) Describe the shape of the distributions. Are they the same? Different? In what way?

Complete the table below.

	Shape	Mean	Standard deviation
Right Skewed Distribution			
Sampling Distribution of the sample mean with n = 10 where the original population was Right Skewed			

Sampling Distribution of the sample mean with n = 100 where the original population was Right Skewed			

1.) Describe the centers of the distributions. Are they the same? Different? In what way?
2.) Describe the variability of the distributions. Are they the same? Different? In what way?
3.) Describe the shape of the distributions. Are they the same? Different? In what way?

Complete the table below.

	Shape	Mean	Standard deviation
Uniform Distribution			
Sampling Distribution of the sample mean with n = 10 where the original population was Uniform.			

Sampling Distribution of the sample mean with n = 100 where the original population was Uniform.			

1.) Describe the centers of the distributions. Are they the same? Different? In what way?
2.) Describe the variability of the distributions. Are they the same? Different? In what way?
3.) Describe the shape of the distributions. Are they the same? Different? In what way?

REFLECTIONS

Where did you see that you were initially correct?
Where did you see that you were initially incorrect?
What did you learn during this experience about what happens to the sampling distribution of the sample mean as n increases?

Inspiration:

delMas, R., Garfield, J., and Chance, B. (1999) "A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning", *Journal of Statistical Education* [online], 7, 3. Available at http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm .

Teaching Notes:

- It is important that students understand they don't have to be correct on the first attempt. They should do their best to think about the issues involved, but they are not required to be correct. In fact, if they discover a misconception, it gives them something to write about in the reflection part of the experience. During the lab activity, the instructor should walk around the room and help students discover their own misconceptions. Perhaps, even a few students might share their misconceptions, so that other students might realize that they also had that misconception.

- It is helpful if you give out this activity in two parts. First, give out part one and have students complete it in pen; then, give out part two. It is also helpful, when possible, if part one and part two can be copied onto different colored sheets of paper.

Teaching Concepts and Analyzing Data with Statistical Software

It is self-evident that statistical software can eliminate or reduce the computational burdens in a statistics course, especially when using large real data sets. One of the key advantages of incorporating a package like SPSS, Minitab, JMP, R, any of the many Excel add-ins, or on-line tools into the course is that they can relieve both students and teachers of the drudgery of computational tasks. Software can considerably reduce the amount of class and homework time devoted to calculation, and can free up cognitive and time resources for other ends. Most introductory-level textbooks now include examples and/or instruction in the use of software and provide datasets to accompany the text. Statistical software allows us to easily show an example data set with thousands of observations and explore a potential multivariate relationship within that data set. Other examples that involve having students perform common analytical tasks and explore data using statistical software can be found in Appendix C.

Using Statistical Software to Teach Concepts

Educators who view a statistical package only as a computational engine may want to consider the considerable potential of these packages for helping students build deep understanding of fundamental abstract **concepts**. The statistics education literature contains numerous articles that both advocate for, and demonstrate the efficacy of, using software to improve student learning (Chance, *et al.* 2007; West, 2009). Software simplifies and expedites the process of constructing and modifying graphs and also allows for replicating operations. For example, in the past, we might have been reluctant to ask students to make multiple histograms of the same data to illustrate the effects of bin width. With software, the task becomes easy. As such, software affords instructors the chance to create in-class demonstrations and homework assignments that guide learners to the "aha! moment" – that moment when a concept is no longer a technical textbook definition but an insight the student genuinely owns. One common example is the concept of a **distribution**. By using software to make dotplots, boxplots and histograms to visualize how individual data points vary along a number line, students gradually construct the idea that observations spread out across a certain range, while also concentrating in certain regions.

For a more subtle example, consider **sampling variability** among simple random samples. This is a concept that is particularly elusive for many students, even for those with considerable prior exposure. Students may read lucid explanations, hear a clear lecture, view or interact with a Central Limit Theorem applet, and yet still not really be able to write or speak confidently about sampling variability or sampling error. Students often have trouble reconciling the images of

"all possible samples" with the knowledge that we typically draw a single random sample. Software may provide an additional avenue to build understanding.

Example Activity

To use this activity in class, students need to have access to computers with the school's favored statistical package available. In classes where this is not feasible, the activity can be modified to be a homework assignment to be completed before class. As suggested elsewhere in this report, if technology is not available to support a homework assignment, instructors might still demonstrate software in class or share images of relevant software output.

Select a large dataset (say, $N > 1000$) with at least one continuous variable. Have each individual student open the data set within the software, and tell the class that, for the purposes of this demonstration, the dataset will serve as a population. The purpose of this demonstration is to explore the extent to which different random samples reliably produce "good" estimates of various population parameters. For example, we might consider the mean, median, and standard deviation.

The instructor uses her/his computer to find the parameters of the population. For dramatic effect, one might even write them on the board and then conceal them, as a metaphor for the true but unknown parameter.

Then, in the first stage, each student uses the software to select a random sample of, perhaps $n = 50$ rows of data, and saves this subset as the student's personal, single random sample. Indicate to the class that each student is a separate, independent investigator gleaning information from a large population. In a small class, students might take multiple samples to achieve the desired result.

The instructor would then ask each student to use the software to find the sample mean and construct both a 95% and a 90% confidence interval for the mean of the population.

Once the students have constructed their confidence intervals, they would be reminded of the value of the actual population mean. The instructor could then say, "First look at your 95% confidence interval, and see if it contains/brackets the actual value of *mu*. Remember, ordinarily we don't know mu, and our only knowledge about the population would come from our one sample. If your only knowledge of this population had been your sample, how many of you would have 'missed' *mu*?"

Students could be asked to stand in place and count off. Naturally, this should be roughly 5% of the class.

Once the students are seated, the instructor could conduct a very brief discussion to inquire why their results were "wrong," leading to the conclusion that sampling error is inherent in the practice of random sampling, rather than any kind of mistake by the investigator.

While these students remain standing, the instructor could ask the same question again, this time referencing the 90% interval, and ask unlucky students to stand. The instructor could point out to the class that (a) the same students are standing again, but (b) they are now joined by an approximately equal number.

Depending on time and the type of software, one might also construct confidence intervals for the median and standard deviation.

If students do not have software in class, one might simply have them take the random samples and construct CIs for homework, bringing their results to class and/or submitting the results online prior to class. At that point, the instructor could create a graph summarizing the distribution of sample means and sample 95% and 90% confidence intervals.

Inspiration:

Chance, B. *et al*, (2007), "The Role of Technology in Improving Student Learning of Statistics." *Technology Innovation in Statistics Education* 1 (1). Available at http://escholarship.org/uc/item/8sd2t4rr.

Diez, D.M., Barr, C.d. and Cetinkaya-Rundel, M. (2012), *OpenIntro Stats* 2ns Ed.. Available at https://www.openintro.org/stat/textbook.php?stat_book=os.

Scheaffer, R., Gnanadesikan, M., Watkins, A., and Witmer, J. (1996), *Activity-Based Statistics: Instructor Resources*, New York: Springer Verlag.

West, W. (2009), "Social Data Analysis with StatCrunch: Potential Benefits to Statistical Education". *Technology Innovations in Statistics Education*, 3(1). Available at http://escholarship.org/uc/item/67j8j18s.

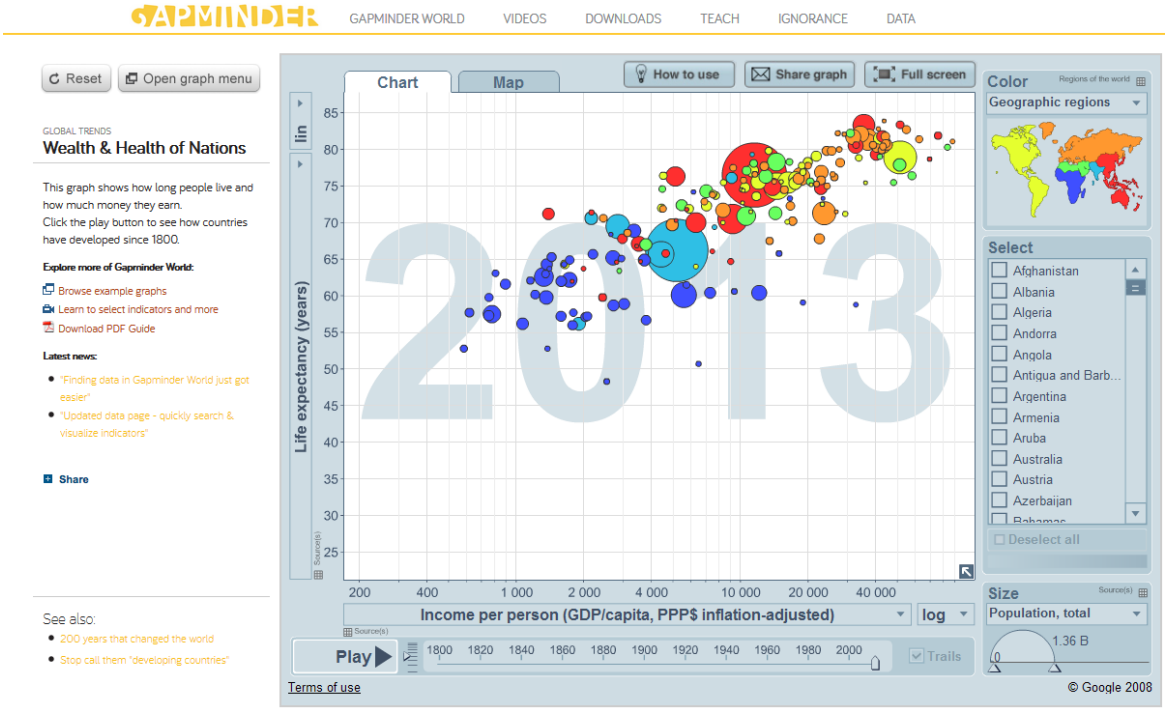
Using Software to Create a Wider Variety of Visualizations

Software not only facilitates the creation and manipulation of traditional statistical graphs, it also has introduced new methods of visualizing large data sets in ways that can stimulate insight and curiosity among undergraduates (and their teachers). Through the use of interactive controls and visual primitives like color, shape, and size, a user can quickly learn to create, modify, or manipulate multidimensional graphics. Such graphing tools are engaging and fun, and they invite the kind of exploration that lies at the heart of statistical thinking.

One innovative graph type is the *bubble chart*, which might be thought of as a super-charged scatterplot. In a bubble plot, there are two quantitative variables on the X and Y axes. Additionally, by replacing dots with "bubbles" of varying sizes and colors, one can represent two additional categorical or quantitative variables. Finally, one can include a time dimension and animate the scatterplot.

As a first illustration, visit the http://www.gapminder.org/world to see a vivid interactive five-dimensional graph that is interactive and animated (see image below). At the same site, one can

download the data (originally from the World Bank's World Development Indicators) and the software needed to create the graph.

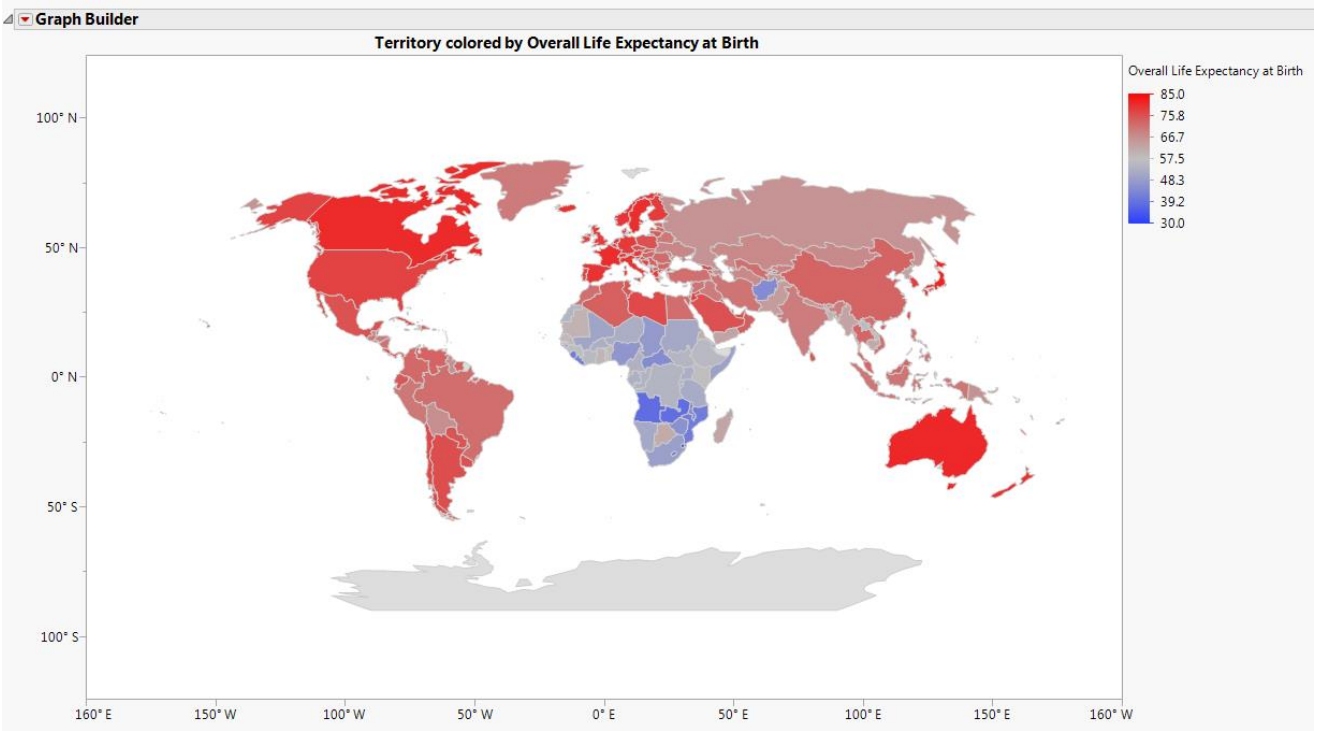


In the default graph as shown, we have data from more than 200 countries covering the period from 1800 through 2013. The vertical axis is Life Expectancy at Birth (in years) and the horizontal axis is the log of Income per Person (GDP per capita, in inflation-adjusted purchasing-power parity in US dollars). Bubble areas correspond to the population of each country annually, and the bubble colors indicate the region of the world for each country.

By pressing the Play button, one sets the graph into motion, tracing the changes in the variables starting in 1800 and progressing through 2013. As the animation continues, patterns and deviations from those patterns appear quite vividly. For example, in the years from roughly 1913 through 1919, life expectancy in much of the world plummets and then rebounds; this time period corresponds to World War I and the Spanish flu epidemic. In our classroom experience, students recognize the dramatic shift in the numbers and raise questions about it. In other words, students engage in statistical thinking as a consequence of viewing this particular visualization. Moreover, one does not need extensive instruction in how to interpret such a graph.

Mapping is another increasingly common visualization that is intuitive and insight-provoking, though statistics textbooks have been slow to add maps to the canon of basic statistical graphing. The Gapminder site includes a world map, as do some other software packages. For this illustration, we'll look at the mapping feature that is standard in JMP, along with a dataset that ships with JMP. This example does not provide complete step by step instructions, but merely illustrates the capability of the software.

JMP's "World Demographics" data set contains observations of 32 variables for 238 countries of the world in 2009. After opening the data table, the user invokes JMP's Graph Builder platform which presents a list of available variables and a blank "canvas" for graph creation. To make a map, one selects a geographic identifier variable and another variable to determine a color gradient. Below is a world map showing the 2009 life expectancy at birth, by country.



Student users can create this graph in a few steps with a point-and-click interface, and can explore different variables in similar fashion. Here again, both the construction and the interpretation of the visualization can occur with minimal instruction—in contrast to, for example, a histogram or box-and-whiskers plot of the same data. Given their visual impact, simple maps like this provide an excellent opportunity to tell a story from data, a key goal of undergraduate statistics education.

Inspiration:

JMP Software, Sample Data Files.

Rosling, Hans. Gapminder World, accessible at www.gapminder.org/world. Accessed 9 June 2015.

Using Software for Reproducibility and Better, Clearer Student Assignments

One recent trend in the scientific community is an emerging consensus on the value and importance of reproducibility in scientific publications (see, for example, the editorial in *Nature*,

2014). In 2014, at a gathering convened by the US National Institutes of Health and the journals *Nature* and *Science*, attendees adopted a set of guidelines calling for, among other provisions, publication of statistical procedures, method of randomization and other detailed information to allow for others to reproduce the published work.

In a statistics classroom, instructors may seek reproducibility as well; in addition to asking students to report final results of an analysis, we may also wish to see the code or dialog choices that generated the output, as well as reading the conclusions that student authors drew. Here again, technology can simplify and expedite the process.

One freely available tool is R Markdown. For this example, we'll demonstrate the ease of use and the instructional benefits of using R Markdown as implemented in RStudio. In this example, we use the Old Faithful dataset to illustrate how the R Markdown environment integrates a student's code with output and with whatever commentary or responses a student might add. Full instructions are beyond the scope of this example; we include it to indicate how this particular technology can overcome a common obstacle in the preparation of coherent complete lab reports.

In R Markdown, a user can combine text with "chunks" of R code, and by "knitting" the R Markdown document, produce a fully-editable Word document (or HTML or pdf object) for submission. The rendered document will contain the student's code, output, and written work. Presumably, the student would be advised to develop and test the code in R prior to transferring chunks to the R Markdown document.

Below is a screen shot of an R Markdown document, followed by the Word document created by it.

RMarkdown code.Rmd

Knit Word | Run | Chunks

```
---
title: "Markdown sample for GAISE Technology Appendix"
author: "A Student"
date: "Month xx, 20xx"
output: word_document
---
NOTE: The text below this is automatically generated when the user creates a new R Markdown
document.  The user can add text simply. This example uses the Old Faithful dataset that ships
with R.

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF,
and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as
well as the output of any embedded R code chunks within the document. You can embed an R code
chunk like this:

```{r}
str(faithful)
summary(faithful)
```

You can also embed plots, for example:
```{r, echo=FALSE}
hist(faithful$eruptions)
plot(faithful)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R
code that generated the plot.

A.Student notices that the histogram of duration is bimodal, and that there is a positive
association between eruption waiting time and the duration of eruptions. Also there are two
clusters of points in the scatterplot.
```

27:1 (Top Level) R Markdown

Markdown sample for GAISE Technology Appendix

A Student

Month xx, 20xx

NOTE: The text below this is automatically generated when the user creates a new R Markdown document. The user can add text simply. This example uses the Old Faithful dataset that ships with R.

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

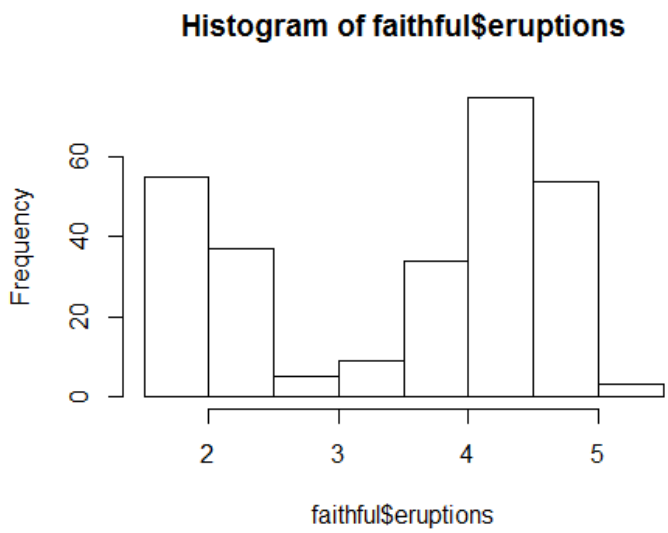
str(faithful)

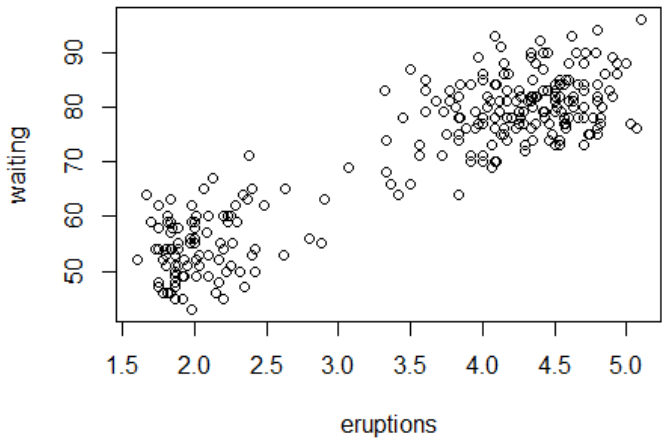
```
## 'data.frame':    272 obs. of  2 variables:
##  $ eruptions: num  3.6 1.8 3.33 2.28 4.53 ...
##  $ waiting  : num  79 54 74 62 85 55 88 85 51 85 ...
```

```
summary(faithful)
```

```
##    eruptions        waiting
##  Min.   :1.600   Min.   :43.0
##  1st Qu.:2.163   1st Qu.:58.0
##  Median :4.000   Median :76.0
##  Mean   :3.488   Mean   :70.9
##  3rd Qu.:4.454   3rd Qu.:82.0
##  Max.   :5.100   Max.   :96.0
```

You can also embed plots, for example:





Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.
A.Student notices that the histogram of duration is bimodal, and that there is a positive association between eruption waiting time and the duration of eruptions. Also there are two clusters of points in the scatterplot.

Inspiration:

Baumer, Ben; Cetinkaya-Rundel, Mine; Bray, Andrew; Loi, Linda; & Horton, Nicholas J. (2014). "R Markdown: Integrating a Reproducible Analysis Tool into Introductory Statistics." *Technology Innovations in Statistics Education*, 8(1).Available at http://escholarship.org/uc/item/90b2f5xh.

Editorial (2014), "Journals unite for reproducibility". *Nature*, 515 (7525) 05 November.

Accessing Real Raw Data online

The third GAISE recommendation is to "Integrate real data with a context and a purpose." Appendix C suggests several ways to generate or acquire real data. This section augments those methods by highlighting web-based avenues to obtain discipline-specific observational, experimental, and survey data. Instructors seeking real data are advised to begin with visits to curated data repositories such as:

- CAUSEweb, the site of the Consortium for the Advancement of Undergraduate Statistics Education.
- Data and Story Library (DASL), maintained by Department of Statistics at Carnegie Mellon University
- Journal of Statistics Education -- click on the JSE Data Archive
- Nationmaster – portal to international economic, demographic and social data
- University of Michigan -- Documents Center "Statistical Resources on the Web"
- UCLA's Datasets sites, maintained by the Statistics Department of UCLA.

Accessing Observational Data

Massive volumes of real-time, automatically generated and/or captured data are now available publicly and for free across disciplines, making it possible to find and use raw data attuned to the needs of courses and the interests of students. This short section briefly describes how one might locate and extract such data, using two examples that illustrate both ends of the "ease of use" spectrum. Because on-line observational data can allow instructors flexibility in choosing and creating assignments, the classroom illustrations shown in this section are framed as templates that instructors should tailor to their students and courses.

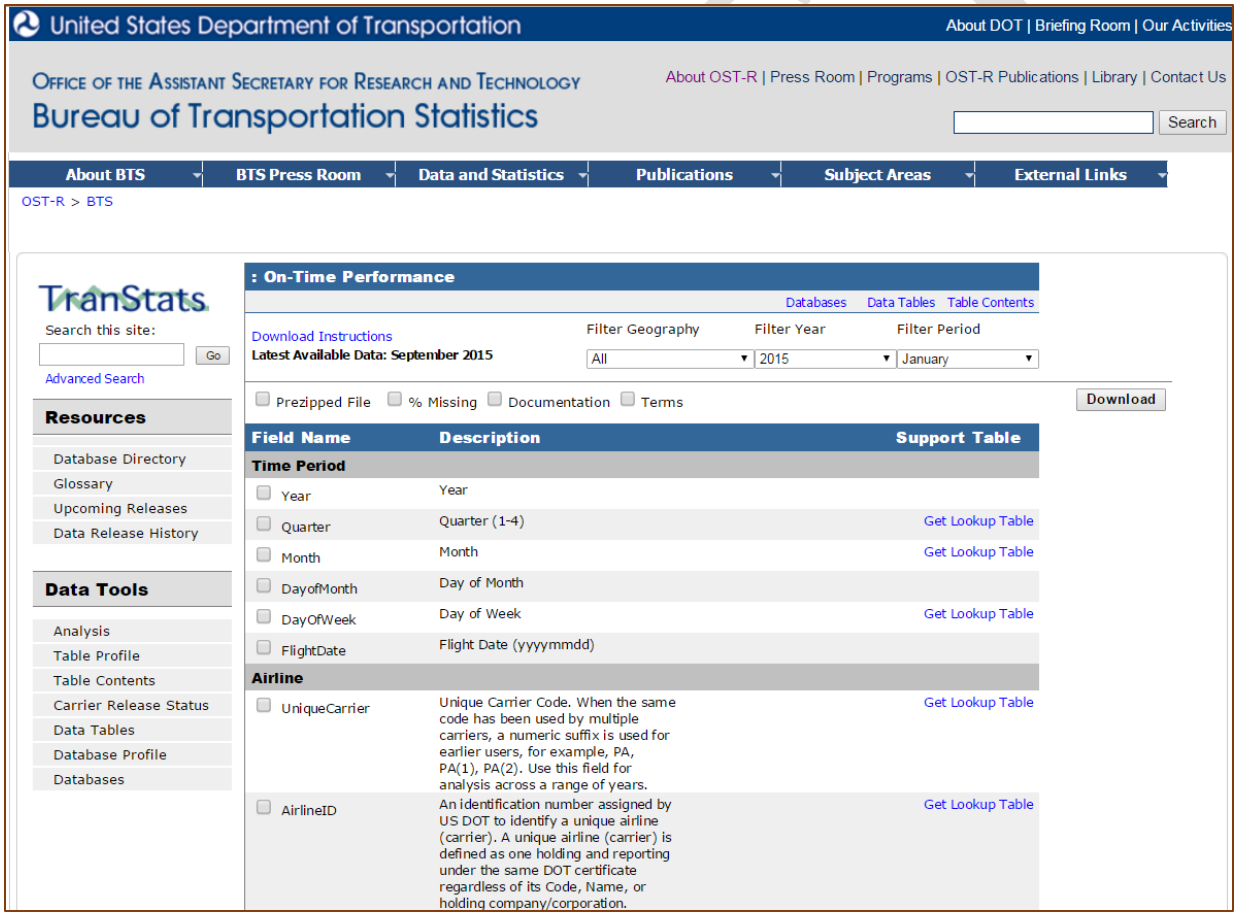
USER-FRIENDLY ACCESS ILLUSTRATION:

Many federal agencies in the U.S. provide user interfaces to build custom queries for transactional databases. The example here comes from the Bureau of Transportation Statistics

and the On-Time Performance database. According to the BTS website, the database "contains on-time arrival data for non-stop domestic flights by major air carriers, and provides such additional items as departure and arrival delays, origin and destination airports, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, taxi-out and taxi-in times, air time, and non-stop distance." Users can download 109 light-level variables for a selected month and year, and can filter geographically. Hence, an instructor can obtain data for nearby airports and recent time periods, selecting variables of particular interest.

The query screen presents a set of drop-down filter selectors and variable checkboxes to specify which variable fields a user wants to download. After making the selections, the site generates and delivers a zipped comma-separated-values (csv) file of raw data.

In late 2015, the search screen looked like this:



Prototype Assignment—Student Prompt

Have you ever been on a flight that left the gate on time, only to be frustrated by a long wait before takeoff? How often do such delays occur? The U.S. Department of Transportation maintains a database with information about the departures and arrivals of every domestic commercial flight in the United States.

We have a dataset that can help us answer the question, and to compare our local airport with some of the busiest airports in the country. In the airline industry, the time elapsed between departing the gate and "wheels up" is known as "Taxi out" time. For this exercise, I have downloaded from the Bureau of Transportation Statistics and placed it in a file called {filename}. The file contains several variables about individual flights departing from our nearest airport (airport code XXX), as well as flights from three airports with very heavy traffic: Atlanta (ATL), Los Angeles (LAX), and Chicago (ORD) during {MONTH, YEAR}.

For this assignment, the three variables of interest are:

- DayofWeek numeric code for day (1 = Sunday, 2 = Monday, etc.)
- Origin three-letter airport ID code
- Carrier airline identification code
- DepTimeBlk standard departure time intervals from the Computerized Reservations System (CRS)
- TaxiOut taxi out time, in minutes

For the month of {MONTH}, use appropriate graphical and descriptive summaries to investigate these questions and prompts:

1. On a typical flight that month at our airport, how long did it take for flights to take off after leaving the gate?
2. How did our taxi out times compare to ATL, LAX and ORD?
3. At our airport, how did the different airlines compare in terms of taxi out times?
4. How (if at all) did taxi out times vary by the day of the week at our airport? Is the variability similar at the other large airports in the dataset?
5. How (if at all) did taxi out times vary by time of day at our airport?
6. Suppose you are planning to fly out of our airport next month. Would you be inclined to take this analysis into account in preparing for your flight? Why or why not?

COMPUTING-INTENSIVE DATA ACCESS ILLUSTRATION:

Professional sports have been among the early adopters of the managerial and strategic use of statistical analysis due in some measure to the technologies that automatically capture data at very granular level. Major League Baseball (MLB) has led the way in this regard. With the types of data available to date, one can investigate questions related to teams, players, plays, and even individual pitches.

For example, since 2007, MLB has been tracking and publishing measurements of every pitch thrown in every game of the season (see Inspiration and References below). The technology behind the data collection is called PITCHf/x, and the relevant MLB website contains XML files within a hierarchical directory tree structure organized by date. More specifically, to access PITCHf/x data for a single game, one first identifies the year, month, date and "game id" for the game of interest.

The data are freely available in the sense that there is no payment required, but in contrast to the prior example, actually accessing and preparing the data for analysis is best done by writing code to automatically scrape dates, games and innings of interest. There are some third-party sites that help with access, but the MLB site exemplifies the challenges and rewards of the "big data" era: the available data open the doors to previously unimaginable analysis, but the doors are complicated to navigate.

This illustration is inspired by a 2010 article in the *Journal of Statistics Education* by Professor Jim Albert, a leader in using sports data for instruction. We are aware that sports data examples excite some students and can alienate others. The point here is to illustrate the availability and challenges of accessing some web sources, not to advocate the use of this particular type of baseball data.

In Albert's article, we have a dataset with the following variables:

Table 4. Variables and description of pitch-by-pitch data frame **pitchdata** using the text file "pitchfx.dat".

Variable	Description	Variable	Description
pitcher	name of pitcher	start_speed	starting speed of pitch
game	game number	end_speed	speed of pitch crossing plate
id	pitcher id number	sz_top	top of strike zone
inning	inning of game	sz_bot	bottom of strike zone
num	number of batter	pfx_x	deviation in horizontal direction
batter	batter id number	pfx_z	deviation in vertical location
stand	hitting side of batter	px	pitch location in x direction
b_height	height of batter	pz	pitch location in z direction
p_throws	throwing side of pitcher	pitch_type	pitch classification
des	play description	count	current pitch count
event	result of plate appearance	new.count	new pitch count
brief_event	brief description of result	value	pitch value
des2	pitch outcome	new.count.type	PA event or new count
type	ball, strike, or in-play?	count.adv	pitcher or batter or neutral count

Prototype Assignment—Student Prompt

In Major League Baseball, successful pitchers combine athletic skill and tactical judgment to outwit opposing batters. Pitchers vary in many respects, including the variety of pitches they use (fastballs, curveballs, etc.), the way they sequence pitches, as well as the speed and movement of their different pitches.

Since the 2007 season, the MLB has used a digital recording system to measure particular parameters of every pitch thrown in every game. The technology involved is known as "PITCHf/x" and for this assignment we have a data file called {FILENAME}. This file contains

all of the games played by our local team, the {TEAM} during the {YEAR} season. In this assignment, we'll focus on the pitching performance of our ace pitcher {PITCHER}. Your tasks:

1. Because teams have several pitchers who rotate from game to game, our first task is to isolate only the data rows involving {PITCHER}. Use software to create a subset of {FILENAME} that contains only pitches thrown by PITCHER.
2. What types of pitches does {PITCHER} tend to throw, and how often does he throw each?
3. What were the outcomes of his pitches as the end of the plate appearance?
4. How do the outcomes vary by type of pitch?
5. Are particular pitches more successful at inducing batters to swing and miss?

Of the different pitch varieties, some are distinguished by their "movement," that is the extent to which they deviate from an imaginary straight line between the pitcher's hand and home plate. In the PITCHf/x data, pfx_x and pfx_z represent the movement in the horizontal and vertical directions respectively. The horizontal break is calculated from a perspective behind the point of home plate, so that negative values of pfx_x move towards a right-handed batter and away from a left-handed batter.

Pitches also vary in speed, and PITCHf/x records two speeds: when the ball leaves the pitcher's hand and when it crosses the plate.

6. Use software to create a scatter plot of the horizontal and vertical movement of {PITCHER's} 4-seam fastballs, curveballs, and change-ups. In the plot, use different symbols for the three pitch types and vary the intensity of the color by the end_speed of the pitch (speed arriving at the batter).
7. In plain English, explain what you see in this scatterplot.

Inspiration and References

Albert, Jim (2010) "Baseball Data at Season, Play-by-Play, and Pitch-by-Pitch Levels" Journal of Statistics Education Volume 18, Number 3 (2010) www.amstat.org/publications/jse/v18n3/albert.pdf.

Major League Baseball (2010), Pitch F/X data files, available at http://gd2.mlb.com/components/game/mlb.

Silver, Nate (2015) "The Best and Worst Airlines, Airports and Flights, Summer 2015 Update". *fivethirtyeight.com*. June 17, 2015. http://fivethirtyeight.com/features/the-best-and-worst-airlines-airports-and-flights-summer-2015-update/.

U.S. Department of Transportation, Bureaus of Transportation Statistics, (2015) "Database Profile" available at http://www.transtats.bts.gov/TableInfo.asp?Table_ID=236, accessed 6 December 2015.

U.S. Department of Transportation, (2015) Office of the Assistant Secretary for Research and Technology, Airline On-Time Performance Data database.

http://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time.

Accessing Experimental Data

Traditionally, raw experimental data has been proprietary and therefore difficult to obtain. More recently, we are seeing contemporaneous phenomena including the Open Science, Science Commons, and Citizen Science movements which all make use of the Internet to advance the sharing of data. As Dawson (2012) has stated,

> "Taking inspiration from the open source software and open access movements, some scientists are now sharing their lab notebooks and raw experimental data openly online. Open science is a broad concept that includes these closely related areas of open notebook science and open data. Advocates of open science believe that there should be no insider information, and all protocols and results -- even those of failed experiments -- should be made visible and open to reuse as soon as possible in open lab notebooks and data repositories."

Recently, in the United States, the National Science Foundation has, as a matter of policy, committed itself to pressing NSF-funded researchers to publish not only results and reports of research, but to publish raw data as well (NSF, 2013).

These developments offer great promise for statistics education to use web browsers and statistical software to obtain and access experimental data, including the ability to access both significant and non-significant results drawn for a wide variety of client disciplines.

The following example uses a subset of data contributed to the University of California Irvine Machine Learning Repository, and it illustrates a common practice of A-B testing in web-based environments. The data were provided by YouTube, as described in the Student Handout that follows. This example deals with a substantive domain that should be familiar to most undergraduates.

Student Handout

YouTube Comedy Slam was a video discovery experiment running on YouTube's version of labs (called TestTube) for a few months in 2011 and 2012. In this experiment, a pair of videos was shown to each user, and users were asked to vote for the video they found to be funnier. Left/right positions of the videos were randomly selected before being presented to users to eliminate position bias. Videos were selected from a large pool of weekly updated sets of videos. Users were self-selected visitors to YouTube.

We have provided you a dataset (n = 3,545 observations), drawn from the original sample of more than 1.1 million preference votes. Each line in this dataset corresponds to one vote over a pair of YouTube videos. One of the videos was a compilation of amusing footage of cats in various settings, and the other was a practical joke played on a co-worker by a friend.

Each video is represented by its YouTube video ID (see references for the URLs of each video). There are three columns in the dataset: Left, Right, and Choice. The first two columns just indicate which video ID was shown in which position, and the third column represents the user's choice.

Shetty (2012) provides a more complete description of the experimental setting:

Quantifying comedy on YouTube: why the number of o's in your LOL matter

Posted by Sanketh Shetty, YouTube Slam Team, Google Research

In a previous post, we talked about quantification of musical talent using machine learning on acoustic features for YouTube Music Slam. We wondered if we could do the same for funny videos, i.e. answer questions such as: is a video funny, how funny do viewers think it is, and why is it funny? We noticed a few audiovisual patterns across comedy videos on YouTube, such as shaky camera motion or audible laughter, which we can automatically detect. While content-based features worked well for music, identifying humor based on just such features is AI-Complete. Humor preference is subjective, perhaps even more so than musical taste.

Fortunately, at YouTube, we have more to work with. We focused on videos uploaded in the comedy category. We captured the uploader's belief in the funniness of their video via features based on title, description and tags. Viewers' reactions, in the form of comments, further validate a video's comedic value. To this end we computed more text features based on words associated with amusement in comments. These included (a) sounds associated with laughter such as hahaha, with culture-dependent variants such as hehehe, jajaja, kekeke, (b) web acronyms such as lol, lmao, rofl, (c) funny and synonyms of funny, and (d) emoticons such as :), ;-), xP. We then trained classifiers to identify funny videos and then tell us why they are funny by categorizing them into genres such as "funny pets", "spoofs or parodies", "standup", "pranks", and "funny commercials".

Next we needed an algorithm to rank these funny videos by comedic potential, e.g. is "Charlie bit my finger" funnier than "David after dentist"? Raw viewcount on its own is insufficient as a ranking metric since it is biased by video age and exposure. We noticed that viewers emphasize their reaction to funny videos in several ways: e.g. capitalization (LOL), elongation (loooooool), repetition (lolololol), exclamation (lollll!!!!), and combinations thereof. If a user uses an "loooooool" vs an "loool", does it mean they were more amused? We designed features to quantify the degree of emphasis on words associated with amusement in viewer comments. We then trained a passive-aggressive ranking algorithm using human-annotated pairwise ground truth and a combination of text and

audiovisual features. Similar to Music Slam, we used this ranker to populate candidates for human voting for our Comedy Slam.

So far, more than 75,000 people have cast more than 700,000 votes, making comedy our most popular slam category. Give it a try!

Further reading:

"Opinion Mining and Sentiment Analysis," by Bo Pang and Lillian Lee.

"A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews," by Oren Tsur, Dmitry Davidov, and Ari Rappoport.

"That's What She Said: Double Entendre Identification," by Chloe Kiddon and

---

1. After reading the background material and exploring the data set, briefly describe this experimental design. To what extent will it be reasonable to generalize from any conclusions drawn? Explain your thinking.
2. What is the response (dependent) variable in this experiment?
3. What is/are the explanatory (independent) variable(s) in this experiment?
4. Discuss how you intend to use your statistical software to prepare the data for analysis and then analyze the data from this experiment.
5. Run an appropriate analysis of the data and report on your conclusions.

Teaching Notes:

- Technology is present in three forms in this exercise. First, the experiment is inherently technologically-based and exemplifies a widely-adopted practice in the design of web interfaces. Second, the data were obtained on-line and required some manipulation to create a student-friendly dataset. Finally, students should be expected to perform the analysis. Depending on the statistical software available to them, instructors may wish to reformat the data.
- The questions listed above should be tailored depending upon course emphasis.

Suggestions to Instructors for further investigation

This example is based on a dataset available at the Machine Learning Repository at the University of California at Irvine (see references below). Instructors who want to find other experimental datasets to create an original assignment might consult other sites such as:

- There are compendia of web resources like http://www.statsci.org/datasets.html or CAUSEWeb, which both curate datasets as well as links to other data compilations.
- Those who teach students in health-related disciplines should visit US government sites such as http://clinicaltrials.gov/ for a clearer sense of current trends in making shared data available.

Inspiration and References

Dawson, D. (2012) "Open Science and Crowd Science: Selected Sites and Resources." Issues in Science and Technology Librarianship, Spring, available at http://www.istl.org/12-spring/internet2.html#online-tools. Accessed 10 October 2014.

"Hilarious Cats" (n.d.) Available at https://www.youtube.com/watch?v=7zCIRPQ8qWc. Accessed 10 October 2014.

National Science Foundation, (2013). "National Science Foundation Collaborates with Federal Partners to Plan for Comprehensive Public Access to Research Results", Press Release 13-030, February 22, 2013. Available at http://www.nsf.gov/news/news_summ.jsp?cntn_id=127043, accessed 10 October 2014.

"O susto do gordinho", https://www.youtube.com/watch?v=W9y6nwBwwyQ. Accessed 10 October 2014.

Shetty, Sanketh. (2012) "Quanitfying Comedy on YouTube: Why the number of p's in your LOL matter" Google Research Blog, February 9. Available at http://googleresearch.blogspot.com/2012/02/quantifying-comedy-on-youtube-why.html, accessed 12 October, 2014.

UC Irvine Machine Learning Repository (2012), available at https://archive.ics.uci.edu/ml/datasets/YouTube+Comedy+Slam+Preference+Data#. Accessed 8 October 2014.

Accessing Real Survey Data

In their quest to use real data from client disciplines, instructors have ready access to an enormous variety of large-scale survey data collected by reputable agencies. Typically, such datasets are accessible via user-friendly web interfaces which include codebooks and background information, and many such sites permit selection by variable. This example uses the General Social Survey (GSS). Its website (NORC, 2014b) describes it as an annual "full-probability, personal interview survey designed to monitor changes in both social characteristics and attitudes" in the United States. The GSS is a project of the National Opinion Research Center (NORC) at the University of Chicago, and has been administered consistently since 1972.

The NORC website (2014a) provides this brief overview of the survey: "The GSS contains a standard 'core' of demographic, behavioral, and attitudinal questions, plus topics of special interest. Many of the core questions have remained unchanged since 1972 to facilitate time-trend studies as well as replication of earlier findings. The GSS takes the pulse of America, and is a unique and valuable resource. It has tracked the opinions of Americans over the last four decades." Subject areas cover a wide range of social and cultural issues, including attitudes about government, religion, the workplace, equality, and popular culture. Annually, the sample size is approximately 2,000 respondents, aged 18 years and above.

Most variables are categorical, using Likert-type scales. Downloads typically include respondent identifiers, basic demographics (e.g., age, region, gender), as well as interview dates and sampling weights. The GSS covers a wide variety of topics, so it is suitable for many

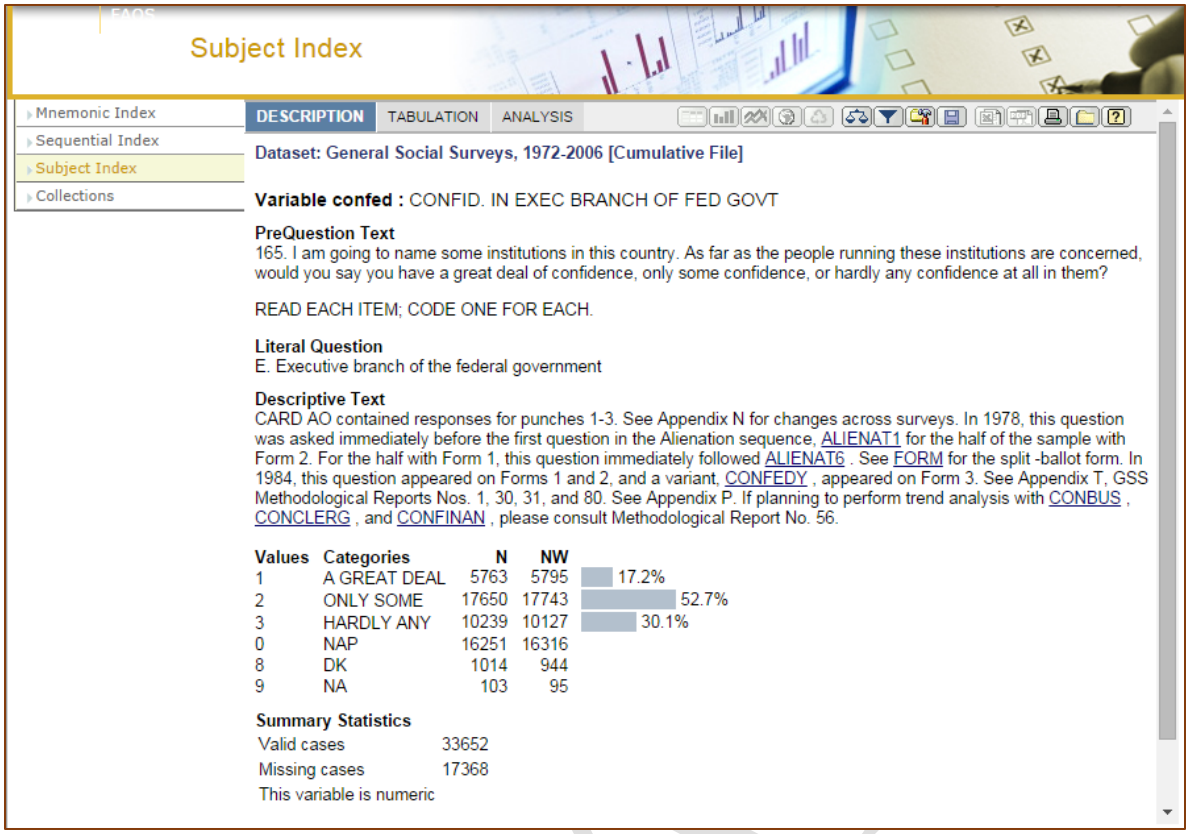
introductory courses, but should be of particular value for any instructor teaching an introductory statistics course with a social science focus.

Due to the changing design and content of websites, this discussion minimizes references to specific URLs or menus, but a web search for NORC or the General Social Survey should suffice. At the time of this writing, one should navigate to the main NORC site: http://www3.norc.org/GSS+Website/

- From the NORC site, users choose between SPSS or STATA formats. These are available freely to any user.
- Alternatively, the NORC download site also provides links to the GSS Data Explorer, to the Roper Center at the University of Connecticut and to ICPSR (Inter-university Consortium for Political and Social Research at the University of Michigan) including ASCII, SAS, delimited, and R. At the latter two sites, membership and/or fees may be required.

Users can download multi-year cumulative files or annual complete datasets. Alternatively, one is able to browse variables by subject, by variable name, or in other ways. Hence, instructors seeking a dataset suitable for a particular course have considerable flexibility of access.

Questions at the core of the GSS appear annually in the instrument, while others may be included just once or periodically. For example, a nearly-annual question (NORC, 2014c) asks "I am going to name some institutions in this country. As far as the people running these institutions are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?... Executive branch of the federal government." The NORC site provides detailed documentation for each variable, as well as summary statistics for the cumulative period 1972–2006. The figure below is a screen capture of the Subject Index page for this particular question, showing the question text and a descriptive summary of responses aggregated over the cumulative time period.



Subject Index

Mnemonic Index
Sequential Index
Subject Index
Collections

DESCRIPTION | TABULATION | ANALYSIS

Dataset: General Social Surveys, 1972-2006 [Cumulative File]

Variable confed : CONFID. IN EXEC BRANCH OF FED GOVT

PreQuestion Text
165. I am going to name some institutions in this country. As far as the people running these institutions are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?

READ EACH ITEM; CODE ONE FOR EACH.

Literal Question
E. Executive branch of the federal government

Descriptive Text
CARD AO contained responses for punches 1-3. See Appendix N for changes across surveys. In 1978, this question was asked immediately before the first question in the Alienation sequence, ALIENAT1 for the half of the sample with Form 2. For the half with Form 1, this question immediately followed ALIENAT6 . See FORM for the split -ballot form. In 1984, this question appeared on Forms 1 and 2, and a variant, CONFEDY , appeared on Form 3. See Appendix T, GSS Methodological Reports Nos. 1, 30, 31, and 80. See Appendix P. If planning to perform trend analysis with CONBUS , CONCLERG , and CONFINAN , please consult Methodological Report No. 56.

Values	Categories	N	NW	
1	A GREAT DEAL	5763	5795	17.2%
2	ONLY SOME	17650	17743	52.7%
3	HARDLY ANY	10239	10127	30.1%
0	NAP	16251	16316	
8	DK	1014	944	
9	NA	103	95	

Summary Statistics

Valid cases	33652
Missing cases	17368

This variable is numeric

Given the wide-ranging scope of the GSS, instructors can make use of the data for a variety of descriptive and inferential assignments and examples, and highlight important concepts in survey construction and/or interpretation of raw data. Here are some possible activities related to this particular question. Note that the frequency table provided above reports both the raw number of responses (N) and the *weighted* number of responses (NW). The reported percentages use the weighted counts divided by the number of valid cases (33,652). The difference between weighted and unweighted counts is probably beyond the scope of most introductory courses, but instructors should preface these questions by noting that survey researchers typically use weighting as a way to compensate for the fact that some demographic groups are over- or under-represented by a particular sampling method.

1. Trained employees of NORC administer the General Social Survey. The "PreQuestion Text" and "Literal Question" and the instruction "READ EACH ITEM; CODE ONE FOR EACH" are worded quite precisely. Why do you think the General Social Survey administrators are so particular about the wording of questions? What difference would it make if an interviewer asked the question using different wording?

2. Before analyzing the responses to this question, look closely at the **Categories** listed. The first three are straightforward enough to interpret, but "NAP," "DK," and "NA" are ambiguous. Visit the NORC website and search for these three abbreviations.

Report briefly on what you find.

3. Below the heading "Summary Statistics," notice the phrase "This variable is numeric." Is it? What do you think the General Social Survey folks mean by this?

4. In the United States, the Federal government consists of three branches: executive, legislative, and judicial. In the aggregate from 1972 through 2006, approximately 17.2% of respondents expressed "a great deal" of confidence in the executive branch. How does that compare to confidence in the other two branches? Visit the NORC site (or use data provided by your instructor) to locate the corresponding variables for the legislative and judicial branches. Write a few short sentences comparing respondents' confidence in the three branches.

5. Looking at these aggregated responses over a period of more than 30 years may raise some questions in your mind. Write down one or two questions about confidence in the institutions of government that you would like to investigate further using GSS data.

Inspiration:

National Opinion Research Center (2014a), "General Social Survey". Available at http://www3.norc.org/GSS+Website/, accessed September 19.

National Opinion Research Center (2014b), "National Data Program for Social Sciences". Available at http://www3.norc.org/GSS+Website/About+GSS/National+Data+Program+for+Social+Sciences/, accessed September 19.

National Opinion Research Center (2014c), "Browse GSS Variables Subject Index". Available at http://www3.norc.org/GSS+Website/Browse+GSS+Variables/Subject+Index/, accessed Sept. 19.

Using Games and Other Virtual Environments

Computer gaming has become a large source of entertainment for many people, including college students. In the past few years, energy has been spent to design online games for use in statistics classrooms. The hope is that by using computer games in the statistics classroom, a higher level of engagement can occur. There are multiple ways that we can use games in the classroom.

Real Data: Student can have personnel experiences with games. Students can put together jigsaw puzzles, complete crossword puzzles or play some other quick online game. The students can then analyze the completion times or other variables from these games.[32]

Gathering Virtual Data: Sometimes data from virtual reality environments can also be used to engage students. For example, students can experience trying to collect data from an endangered species[33] or conduct a health survey across an entire Island. [34]

Experimental Design: Another way to incorporate the use of games into the classroom is by having students think about what factors affect the time to win a game or the points earned in a game. For example, does gender, the amount of hints, color of the pieces, and/or seeing a preview of a game affect the chance of winning the game?

Statistical Concept: With some games, players can only advance or move ahead in the game if they have mastered a statistical concept. By learning and applying a statistical concept, you are able to win. For example, perhaps by analyzing a set of past attempts, a more successful method of getting to the goal can be discovered. Or, perhaps studying the conditional probabilities of moves or studying a scatterplot may increase the number of points a player has in the game.[35]

Inspiration:

Erikson, T. (2013), "Designing Games for Understanding in a Data Analysis Environment", *Technology Innovations in Statistics Education* [online], 7, 2. Available at http://escholarship.org/uc/item/31t469kg .

Erikson, T. and Triggs, M. (2014), "An Early Look at Rich Learning Analytics: Statistics Students Playing "Markov" ", *iCOTS 9 Proceedings* [online]. Available at http://icots.info/9/proceedings/pdfs/ICOTS9_9G1_ERICKSON.pdf .

Kuiper, S. and Sturdivant, R. (2014) "Using Fun and Games to Engage Real-World Learning", *eCOTS 14 Workshop* [online]. Available at https://www.causeweb.org/ecots/ecots14/34/ .

Sturdivant, R. Jackson, J., and Cummiskey, K. (2013), "TigerStat: An Immersive 3-D Game for Statistics Classes", *CAUSE webinar* [online], Available at https://www.causeweb.org/webinar/activity/2013-04/ .

[32] Games and written lab activities can be found at this website http://web.grinnell.edu/individuals/kuipers/stat2labs/. Other statistical games and puzzles can be found online through a google search.
[33] For example, Shondra Kuiper has designed TigerStat that can be used to gather information about tigers: http://statgames.tietronix.com/TigerSTAT/.
[34] Michael Bulmer has also designed the Islands for data collection across a virtual population http://islands.smp.uq.edu.au/login.php
[35] Some of these games can be found at http://play.ccssgames.com/

Bulmer, M. and Haladyn, J.K. (2011), "Life on an Island: a Simulated Population to Support Student Projects in Statistics", *Technology Innovations in Statistics Education* [online], 5**,** 1. Available at http://escholarship.org/uc/item/2q0740hv .

Baglin, J., Bedford, A. and Bulmer, M. (2013), "Students' Experiences and Perceptions of Using a Virtual Environment for Project-Based Assessment in an Online Introductory Statistics Course", *Technology Innovations in Statistics Education* [online], 7**,** 2. Available at http://escholarship.org/uc/item/137120mt .

Student Handout

Your instructor will present you with a game.[36] Take a few minutes to familiarize yourself with the game.

Part One: As you familiarize yourself with the game, what are you interested in discovering? List a few of these questions here:

1.)

2.)

3.)

Part Two: Pick one of your above questions and think about how you would gather evidence to answer it. Here are some issues that you should consider.

1. What type of data would you need to collect?
2. Would these variables be categorical or quantitative?
3. What type of lurking variables would you need to be concerned about? Why?
4. How would you include random sampling and/or random allocation?
5. What statistical method have we discussed in class could be used to analyze that data?

Part Three: Write up a study protocol. Be very detailed. Write the instructions for how an experimenter would go about conducting this experiment.

Teaching Note:

[36] There are many online games where data can be collected. You can even use a traditional puzzle or board puzzle.

After this activity has been completed, one possible next step is to have the class pick one of the study ideas. The class can then collect data and analyze it. It might be a good idea for the students to complete the game outside of class time. Some students may take longer to complete certain puzzles and you don't want to accidently embarrass someone that takes twice as long as other students. If you decide to have students complete the puzzle outside of class, you can then also discuss what lurking variables this might introduce. If you decide to complete the puzzle in class, another option would be to give the students 1 minute to complete as much of the puzzle as they can and then measure percentage of completion.

Real Time Response

While not limited to statistics classrooms, real time responses systems (clickers) can be an asset in achieving the GAISE recommendations. The GAISE recommendations encourage instructors to use technology for computation and emphasizing concepts. Real time responses allow us to explore concepts, as well as gather data to analyze in the classroom.

Initially, real time response systems[37] or audience response systems started as devices similar to TV remotes. The device, commonly known as a clicker, could only be used to respond in class to a teacher's multiple choice question posted on a projection screen. Over the past decade or more, these devices have evolved. Devices that only allow for multiple choice entry are still available, and some also allow for numeric entry. Moreover, now there are systems that allow the students to enter information from their computers, tablets or even phones. With these devices, questions sometimes even appear directly on the device and the students are no longer limited to just multiple choice. The students can enter numerical values, equations, and even draw on the screen. These responses can then be combined and shared with the class. There is a vast array of options in how anonymous the responses from the students can be. The student's responses could be made anonymous from both other students and even from the instructor (in some cases). Having various options for anonymity allows for flexibility in teaching methods.

For example, some systems post a summary of responses provided by the class – the percentage that answered A, B, C, etc. This allows the students to be able to judge where they stand in the class. Since this data is summarized and not individual responses, it provides a layer of protection for a shy student who might not ordinarily participate.

Real Time Response Systems allow us to receive responses synchronously in a face-to-face class from all students. However, these systems can also be used for completion of assigned outside of

[37] Some of these devices include: specific devices: "clickers" such as iClicker or H~ITT; web based services such as Learning Catalytics, TopHat, b.socrative.com, Google Forms, Survey Monkey, Qualtrics, Poll Anywhere, Course Management Software

class. The questions can be assigned for homework and recorded on a real time responses system outside of class.

Some of the new systems even help enable team-based answering. For example, the system first asks a question or several questions which are answered by the individual student. Then, the students form teams. This team could be created by the instructor, self-selected by the students or created by the real time response system based on their individual answers. Then, the team discusses the questions, and answers them again as a team. Some systems even show what each member of the team initially answered, enabling better team discussion.

The real time class responses can be used in multiple ways in the face-to-face classroom environment to. Some of these examples include:

- Review the previous day's assignment or ask questions about an assigned reading
- Expose misconceptions and use them as talking points
- Illustrate a concept
- Collect data to analyze
- Allow instructors to determine if they are going too fast/ too slow
- Supplement applets by helping to focus the students' explorations of the applet. For example, first ask the students to answer a series of questions, then have them play with an applet and re-answer the same series of questions.

Future Direction of Real Time Response

With the presence of cloud based computing, more opportunities will become available for students to collaborate in real time but separate spaces. For example, the students might work together to create one document[38] or annotate an existing document[39].

Best Practices and Ideas from Statistics Education Literature

- Use a small number of clicker questions with a clear defined reason for each question and its placement in the lesson.
    - From:
      McGowan, H. & Gunderson, B. (2010), "A Randomized Experiment Exploring How Certain Features of Clicker Use Effect Undergraduate Student's Engagement and Learning in Statistics," *Technology Innovations in Statistics Education* [online], 4, 1. Available at http://escholarship.org/uc/item/2503w2np .

- Use clickers to promote understanding of concepts, not just calculation, for topics such as inference and applets.
    - See the following for a compilation of ideas:

[38] For example, using Goggle Docs or etherpad.
[39] For example, using ClassroomSalon (www.classroomSalon.com) or https://oeit.mit.edu/gallery/projects/nb-pdf-annotation-tool

Kaplan, J. (2011), "Innovative Activities: How Clickers can facilitate the Use of Simulations in Large Lecture Classes", *Technology Innovations in Statistics Education*, 5, 1. Available at http://escholarship.org/uc/item/1jg0274b/.

- Ask concept questions or questions that investigate common misconceptions.
  - See the following for more information:
    Perez, C, and Lane, D. (2012), "Simulations, Audience Response Systems and the Classroom: Engaging the Modern Student", *eCOTS 12* [online]. Available at https://www.causeweb.org/ecots/ecots12/breakouts/8/

Best Practices and Ideas from Education Literature from Other Disciplines

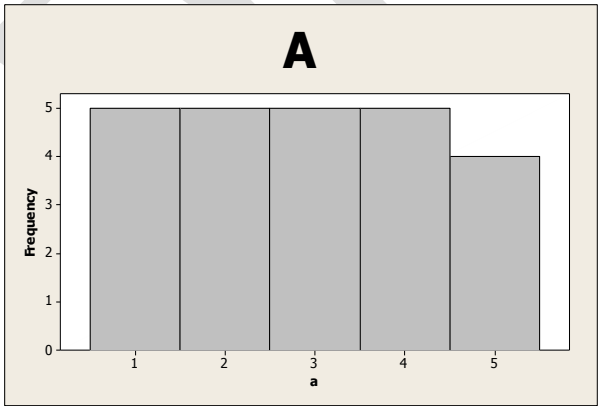
- For summary of best practices from various academic disciplines including life sciences and physics see the following:
  Caldwell, J. (2007) "Clickers in Large Classroom: Current Research and Best Practice Tips", *CBE Life Sciences Education* [online], Spring, 6, 1, 9 – 20. Available at http://ncbi.nlm.nih.gov/pmc/articles/PMC1810212/ .

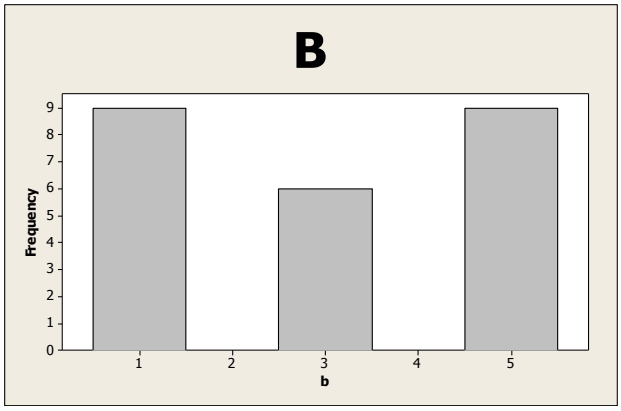
- For a summary of experiences with Peer Instruction:
  Lasry, N., Mazur, E., and Watkins, J. (2008), "Peer Instruction: From Harvard to Community Colleges", *American Journal of Physics*, 76, 1066-1069. Available at http://dx.doi.org/10.1119/1.2978182 .

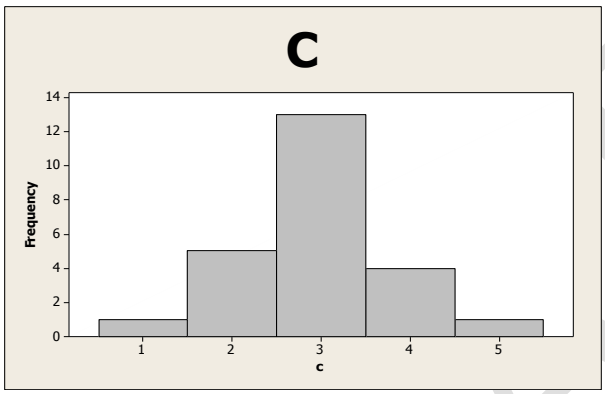
EXAMPLE ASSIGNMENT 1: USING REAL TIME RESPONSE AS AN EXAMPLE OF HOW TO BRING ATTENTION TO A MISCONCEPTION.

Teacher Resources: Slides Posing Questions

Which of the following graphs has the smallest standard deviation?







1. First have the students respond on their own using their own device.

2. After the students have responded, have the students re-answer the question after a team discussion.

Inspiration:

- Kaplan J, Gabrosek, J., Curtiss, P. and Malone, C. (2014), "Investigating student understanding of a histogram", *Journal of Statistical Education* [online], 22, 2. Available at http://www.amstat.org/publications/jse/v22n2/kaplan.pdf .
- Cooper, L.L. and Shore, F.S. (2010), "The Effects of Data and Graph Type on Concepts and Visualizations of Variability, "*Journal of Statistical Education* [online], 18, 2, Available at http://www.amstat.org/publications/jse/v16n2/cooper.html .

Teaching Notes:

- A known misconception is that students seem to assume that flatter histograms mean less variation.
- The correct answer is C. In histogram C, the average distance of points from the mean is less than in the other histograms. In histograms A and B, more of the points are a further distance from the mean.

EXAMPLE ASSIGNMENT 2: USING A REAL TIME RESPONSE AS A WAY TO GUIDE A STUDENT'S EXPERIENCE WITH AN APPLET.

Teacher Resources: Slides Posing Questions

1. Do outliers affect the value of the standard deviation?
   a.) No
   b.) Yes

2. Suppose that your data set has a point that is much lower than the rest. What type of effect (if any) would this have on the value of the standard deviation?
   a.) It would make it larger.
   b.) It would make is smaller.
   c.) It would stay the same.
   d.) Unable to be determined.

3. Suppose that your data set has a point that is much higher than the rest. What type of effect (if any) would this have on the value of the standard deviation?
   a.) It would make it larger.
   b.) It would make is smaller.
   c.) It would stay the same.
   d.) Unable to be determined.

Inspiration:

Perez, C, and Lane, D. (2012) "Simulations, Audience Response Systems and the Classroom: Engaging the Modern Student", eCOTS 12 [online]. Available at https://www.causeweb.org/ecots/ecots12/breakouts/8/ .

Teaching Notes:

- Correct Answers: 1.b, 2. a 3. a
- In order to focus students' attention on an applet, it is advised that the students be aware of the questions they are investigating before being exposed to the applet. The real time response systems help with this by focusing students' thoughts before their experience using the applet and then having them re-evaluate their answers at the end of the experience.
- Sometimes, students think that adding a lower point to a data set will cause the standard deviation to get smaller, similar to what would happen with the mean.
- You could also ask the students to explore this concept with their calculators or statistical software if you do not have internet access to an applet.

EXAMPLE ASSIGNMENT 3: ILLUSTRATING THE SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

Teacher Resources: Slides Posing Questions[40]

When you flipped the coin 1 time, what proportion of heads did you get? _____

When you flipped the coin 5 times, what proportion of heads did you get? _____

When you flipped the coin 25 times, what proportion of heads did you get? _____

Student Handout

Flip a fair coin one time, five times and then twenty-five times. Enter your responses below and enter the responses in the survey mechanism. After all students have entered their data, you will be given the data from all students to graph. Make a graph of the results for n = 1, n = 5, and n = 25.

	Your Response	Sketch of results
n = 1		
n = 5		
n = 25		

[40] This set of questions is designed for response systems that allow for quantitative responses; however, with a little work, an instructor can create multiple responses that would work. After the responses have been received, the instructor will need to download the results and distribute them to the students. The instructor can also demonstrate making the histograms.

Based on your observations above . . .

- What happens to the center as *n* increases?
- What happens to the standard deviation as *n* increases?
- What happens to the shape as *n* increases?
- In your own words why does increasing the number of flips lead to these results?

Inspiration:

Taylor, L. and Doehler, K. (2014), "Using Online Surveys to Promote and Assess Learning" *Teaching Statistics* [online], 36, 2, 34- 40. Available at http://onlinelibrary.wiley.com/doi/10.1111/test.12045/full .

Teaching Notes:

- The objective of this activity is to explore the concepts of the sampling distribution of the sample proportion.
- Have the students flip a coin and record the proportion of heads for their series of trials. Once they have completed the flips, they should enter the data into the "Real Time" response system.
- The students should notice that as *n* increases, the bell shaped distribution starts to appear, the amount of variation decreases, but the center stays at 0.5[41].
  Kaplan, J. (2011), "Innovative Activities: How Clickers can facilitate the Use of Simulations in Large Lecture Classes", *Technology Innovations in Statistics Education* [online], 5, 1. Available at http://escholarship.org/uc/item/1jg0274b/.

[41] This activity focuses on flipping a fair coin because coins are easily accessible to most students and teachers; however, you could alter the activity so that the probability of success was something other than 0.5.

APPENDIX E: 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[42], 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.

Examples of Exemplary Assessment Items

We begin by providing examples of exemplary assessment items with commentary about the items. We regard these as exemplary because they reflect the GAISE recommendations of setting problems in realistic, meaningful contexts; they are data-based; and they go beyond calculation to probe deeper understanding of concepts.

ITEM 1

Scientists use metal bands to tag penguins. Do the bands harm the birds?

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 proportions 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. Do bands hurt the penguins? Be sure to address the issue of causation as well as the issue of significance. Also justify your conclusion.

[42] 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.

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.[43]

ITEM 6

Suppose that Nancy, who is statistically savvy, wants to compare the average costs of textbooks for students at her college between the fall and spring semesters of last year. Let μ_F and μ_S represent the two population means. 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: $\mu_F = \mu_S$ vs. H_A: $\mu_F \neq \mu_S$ and tell Nancy whether you reject H_0."

2. Linda says, "Report a 95% confidence interval for $\mu_F - \mu_S$."

3. Steve says, "Conduct a test of H_0: $\mu_F = \mu_S$ vs. H_A: $\mu_F \neq \mu_S$ and report to Nancy the p-value from the test."

4. Gloria says, "Compare $\bar{y}_2$ to $\bar{y}_2$. If $\bar{y}_1 > \bar{y}_2$, then test H_0: $\mu_F = \mu_S$ vs. H_A: $\mu_F > \mu_S$ using alpha =0.05 and tell Nancy whether you reject H_0. If $\bar{y}_1 < \bar{y}_2$, then test H_0: $\mu_F = \mu_S$ vs. H_A: $\mu_F < \mu_S$ using alpha = 0.05 and tell Nancy whether you reject H_0."

[43] 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).

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 Needing Improvement and Commentary

We next give some examples of assessment items with problems and commentary about the nature of the difficulty. We recommend that questions such as these should either be improved as discussed in the following section or dropped from use.

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.[44]

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.[45]

1. State the null and alternative hypotheses.
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?[46]

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.

[44] 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.
[45] 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.
[46] 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 this study to someone without background knowledge of statistics.[47]

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 value of the standard deviation of a data set depends on the center of the distribution. True or False

CHANGED TO:

Does the size of the standard deviation of a data set depend on the center of the distribution?
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 indicates a stronger association 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.

[47] 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.

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?[48]

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.

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

ITEM 14 (REVISITED)[49]

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

[49] 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.

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.

Descriptive Statistics

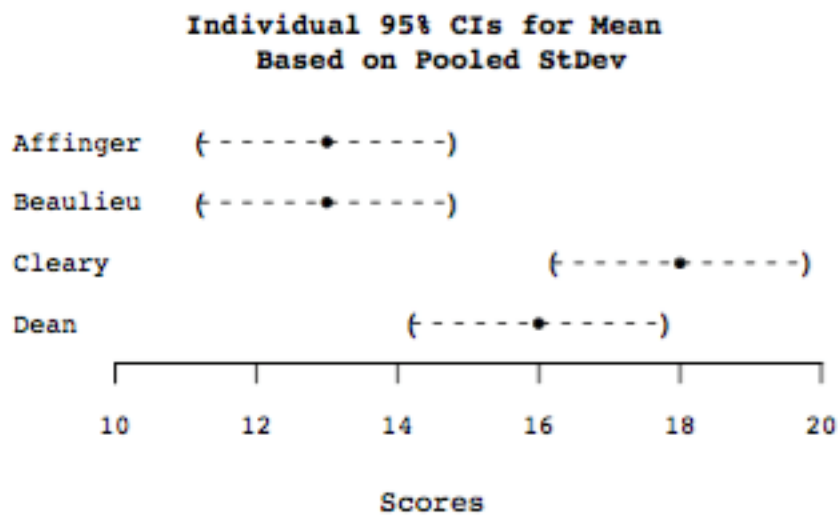
Variable	N	Mean	Median	TrMean	StDev	SEMean
Score	24.00	15.00	15.00	15.00	2.92	0.60

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



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).

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 a cause-and-effect conclusion be drawn 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 (one-sided) alternative, was 0.06. For each of the following, say whether the statement is true or false and why.

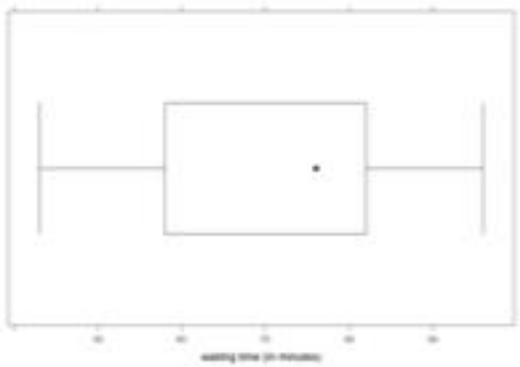
1. There is a 6% chance that the two population distributions really are the same.
2. If the two population distributions really are the same, then a difference between the two samples as extreme as the difference that these researchers observed would only happen 6% of the time.

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. 5%
2. 10%
3. 20%
4. 30%
5. 50%
6. 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 a sample of 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.

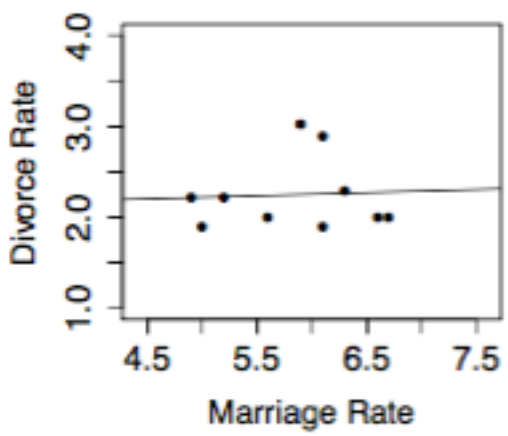
ITEM 20

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 and marriage rate (both as number per year per 1000 adults) 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

To compare a quantitative response variable across four groups, I selected random samples from each of the four groups and constructed parallel dotplots to compare the distributions across the four groups. 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. 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

Students 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 standard error of the sample slope is 0.81. Also, $s_{y|x} = 3.1$.

1. Find r^2 and interpret r^2 in the context of this problem.

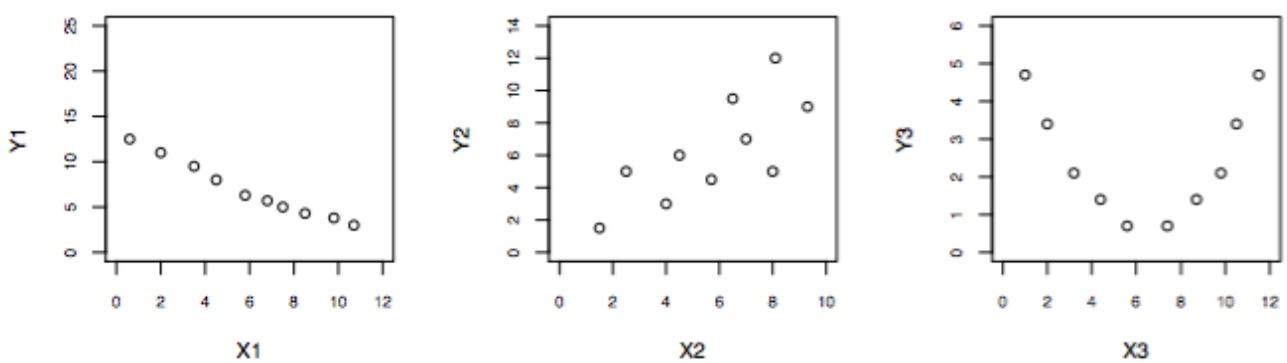
2. Suppose a cereal has 2.63 grams of fiber/ounce and costs 17.3 cents/ounce. What is the residual for this cereal?

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 situation. 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:

Confidence interval (for a mean, p, …)	Normal distribution
Determining sample size	Correlation
Test for a mean	Simple linear regression
Test for a 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

A. 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?

B. 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?

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

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. 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)

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, indicate (by letter) the appropriate statistical inference procedure for investigating the question.[50]

1. 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.

2. 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.

3. 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

[50] The list of methods or examples can be shortened.

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.

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

5. 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?

6. 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)?

7. 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.

8. 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 sold to homeowners? To answer this question, university researchers placed 12 radon detectors in a chamber that exposed them to 105 picocuries per liter of radon. The detector readings found are below, along with some descriptive statistics.[51]

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

Variable	N	Mean	Median	TrMean	StDev	SEMean	Minimum	Q1	Q3
readings	12	104.13	102.75	103.51	9.40	2.71	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. Explain what a Type I error associated with this situation would be.
3. Explain what a Type II error associated with this situation would be.
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

[51] 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.

the brewing method. Suppose you want to repeat the FDA study to obtain an estimate of the mean caffeine content to within 5 mg with 95% using your favorite brewing method. How many cups of coffee must you brew to be 95% confident? In problems such as this, we can estimate the standard deviation of the population to be 1/4 of the range.

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 achieve this goal. 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 indicated 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	Sample size	Sample mean	Sample SD
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

A group of physicians subjected polygraph testing 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 35

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. Write a brief description of the distribution 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.

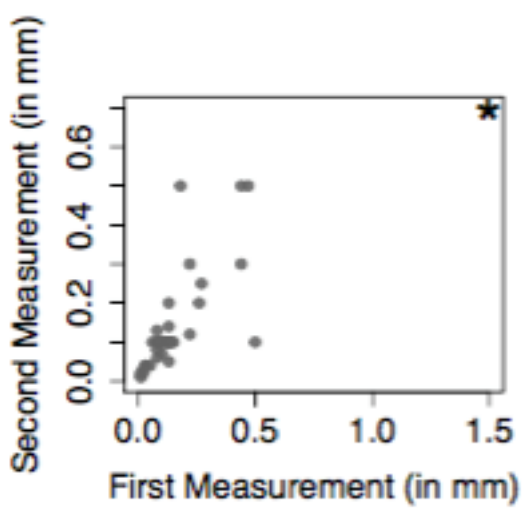
ITEM 36

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. Identify a potential confounding variable in this work.
4. Is this a matched-pair design? Defend your answer.

ITEM 37

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 relationship between the variables.
2. What effect does the starred point have on the correlation coefficient? That is, if the starred point were removed, how would be correlation coefficient change, if at all?

ITEM 38

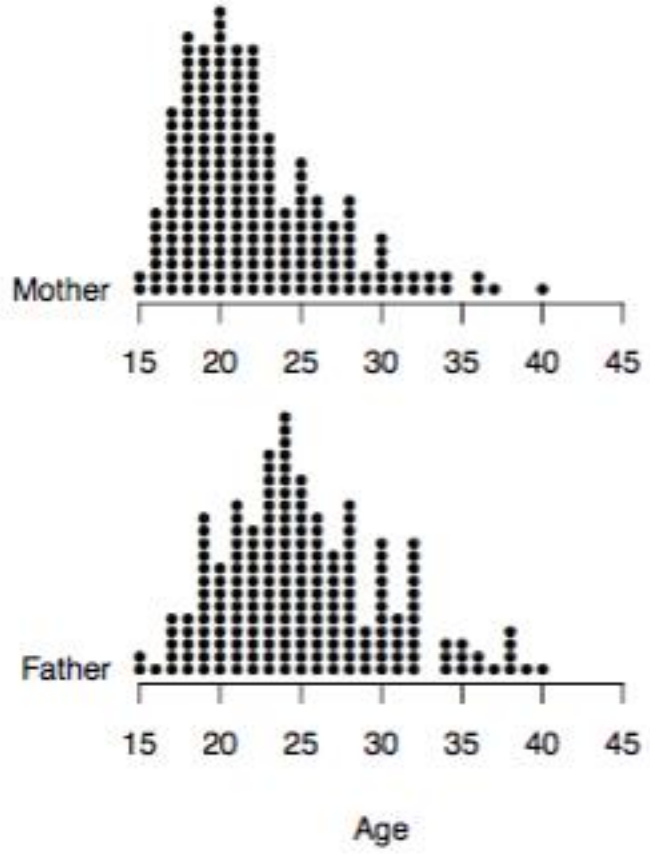
A study in the *Journal of Leisure Research* investigated the relationship between academic performance and leisure activities. Each in a sample of 159 high-school students was asked to state how many leisure activities they participated in weekly. From the list, activities that involved reading, writing, 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 39

A random sample of 200 mothers and a separate random 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 40

When conducting 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 41

When conducting 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 42

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 43

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.[52]

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

[52] 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.

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

ITEM 44

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 45

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_A: $\mu \neq 3$ 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 H0 is true. Therefore we could conclude that the evidence is against the Null Hypothesis, and H0 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 46

Explain 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 47

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.

ITEM 48

Rosiglitazone is the active ingredient in the controversial type~2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below.

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

Determine if each of the following statements is true or false. If false, explain why. *Be careful:* The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

1. Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.

2. The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was (2,593 / 67,593 = 0.038) 3.8% for patients on this treatment, while it was only (5,386 / 159,978 = 0.034) 3.4% for patients on pioglitazone.

3. The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.

4. Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.

The next several items are based on simulation (resampling) methods.

ITEM 49

Rosiglitazone is the active ingredient in the controversial type~2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos.

A randomized study compared the rates of serious cardiovascular problems for diabetic patients on rosiglitazone and pioglitazone treatments. The table below summarizes the results of the study.

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

1. What proportion of all patients had cardiovascular problems?

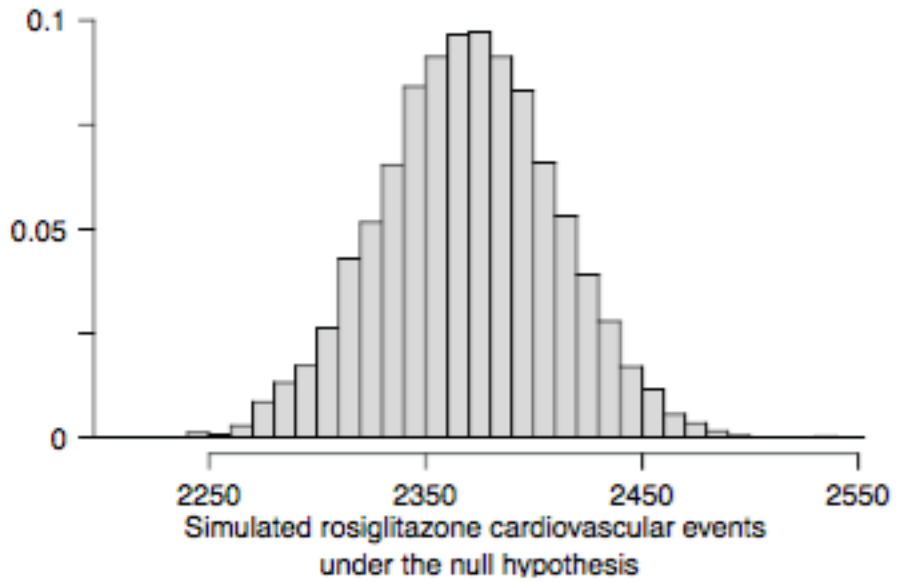
2. If the type of treatment and having cardiovascular problems were independent (null hypothesis), about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?

3. We can investigate the relationship between outcome and treatment in this study using a randomization technique. While in reality we would carry out the simulations required for randomization using statistical software, suppose we actually simulate using index cards. In order to simulate from the null hypothesis, which states that the outcomes were independent of the treatment, we write whether or not each patient had a cardiovascular problem on cards, shuffled all the cards together, then deal them into two groups of size 67,593 and 159,978. We repeat this simulation 10,000 times and each time record the number of people in the rosiglitazone group who had cardiovascular problems. Below is a relative frequency histogram of these counts.

4. What are the claims being tested?

5. Compared to the number calculated in the second part, which would provide more support for the alternative hypothesis, *more* or *fewer* patients with cardiovascular problems in the rosiglitazone group?

6. What do the simulation results suggest about the relationship between taking rosiglitazone and having cardiovascular problems in diabetic patients?



ITEM 50

The Stanford Heart Transplant Study was a randomized trial of a new medical intervention. Of the 34 patients in the control group, 4 were alive at the end of the study. Of the 69 patients in the treatment group, 24 were alive. The contingency table below summarizes these results.

		Group		
		Control	Treatment	Total
Outcome	Alive	4	24	28
	Dead	30	45	75
	Total	34	69	103

1. What proportion of patients in the treatment group and what proportion of patients in the control group died?

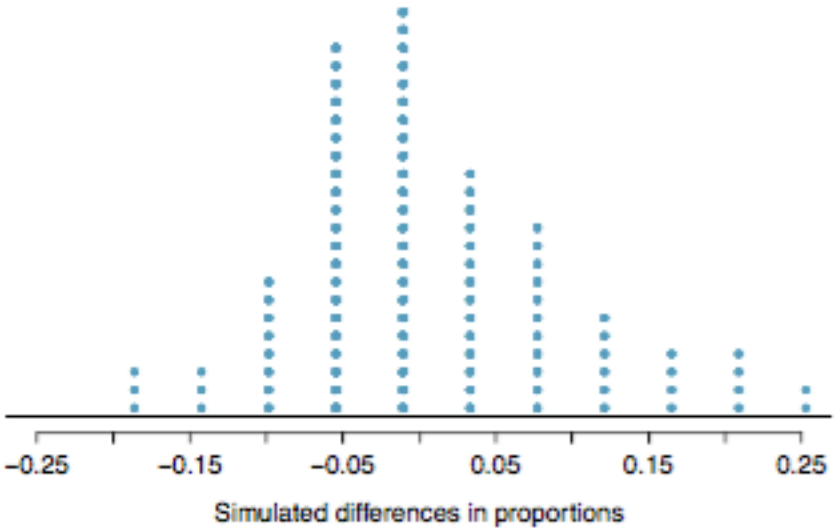
2. One approach for investigating whether or not the treatment is effective is to use a randomization technique.

2.1 What are the claims being tested? Use the same null and alternative hypothesis notation used in the section.

2.2 The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.
We write *alive* on _____ cards representing patients who were alive at the end of the study, and *dead* on _____ cards representing patients who were not. Then, we shuffle these cards

and split them into two groups: one group of size _____ representing treatment, and another group of size _____ representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this many times to build a distribution centered at _____. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are _____. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

2.3 What do the simulation results shown below suggest about the effectiveness of the transplant program?



ITEM 51

Researchers studying the effect of antibiotic treatment compared to symptomatic treatment for acute sinusitis randomly assigned 166 adults diagnosed with sinusitis into two groups. Participants in the antibiotic group received a 10-day course of an antibiotic, and the rest received symptomatic treatments as a placebo. These pills had the same taste and packaging as the antibiotic. At the end of the 10-day period patients were asked if they experienced improvement in symptoms since the beginning of the study. The distribution of responses is summarized below.

		Self-reported improvement in symptoms		
		Yes	No	Total
Treatment	Antibiotic	66	19	85
	Placebo	65	16	81
	Total	131	35	166

1. What type of a study is this?

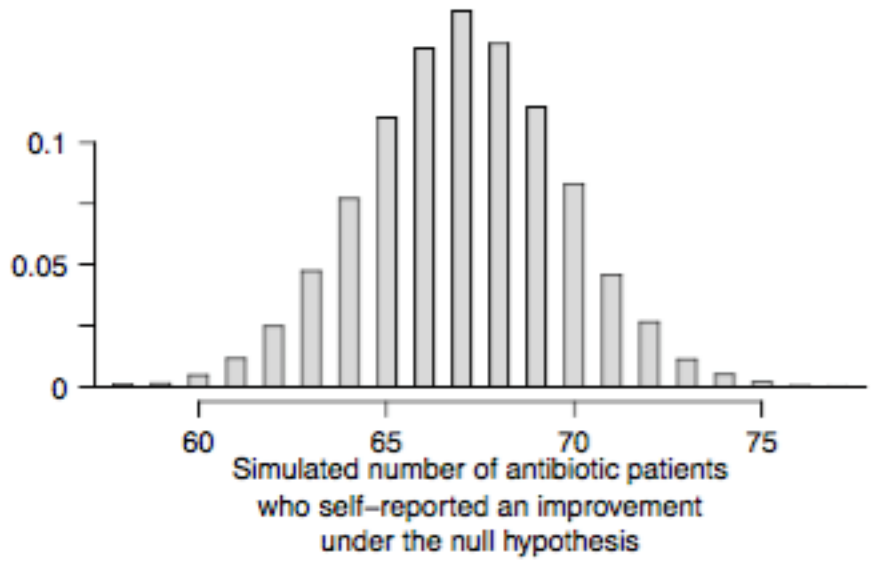
2. Does this study make use of blinding? Justify your answer.

3. Compute the difference in the proportions of patients who self-reported an improvement in symptoms in the two groups: $\hat{p}_{antibiotic} - \hat{p}_{placebo}$.

4. At first glance, does antibiotic or placebo appear to be more effective for the treatment of sinusitis? Explain your reasoning using appropriate statistics.

5. There are two competing claims that this study is used to compare: the null hypothesis that the antibiotic has no impact and the alternative hypothesis that it has an impact. Write out these competing claims in easy-to-understand language and in the context of the application.

6. Below is a histogram of simulation results computed under the null hypothesis. In each simulation, the summary value reported was the number of patients who received antibiotics and self-reported an improvement in symptoms. Write a conclusion for the hypothesis test in plain language. (Hint: Does the value observed in the study, 66, seem unusual in this distribution generated under the null hypothesis?)



Examples of Assessments for Presentations and Projects

Projects and presentations are an increasingly common component of introductory statistics courses.[53]

[53] Halvorsen (ICOTS, 2010, \url{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 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 F: Learning Environments

Instructors work in a variety of settings, and some readers may question whether they can adopt the GAISE recommendations. Different classroom situations have areas of greater and lesser challenges in the implementation of the recommendations. This appendix provides examples of ways to apply the Guidelines in different environments. Five common instructional conditions explored are

- face-to-face, both large and small class sizes
- flipped (inverted) classes
- distance learning
- cooperative learning
- limited technology

The purpose of this appendix is to provide research references and a few inspiring examples for implementing GAISE teaching in courses where one or more of the Recommendations for Teaching appear difficult to employ.

Face-to-Face Courses

Whether instruction is in a classroom or an individual tutoring setting, instruction in statistics has primarily been in a face-to-face format. While times are changing and other approaches are now available, the majority of college and university teaching of statistics still occurs in a face-to-face environment. In these college settings the class size ranges from small to medium to large and to what some would even call very (or extremely) large. In this appendix, we illustrate how to incorporate the GAISE recommendations in teaching situations made complex by class size. For example, some have questioned the feasibility of active learning in a large class setting. Others have found that with very small classes a simulation completed with manipulatives by the students in the class might not demonstrate the desired principle.

Small Classes:

Collecting data from students during class is suggested as a way to *foster active learning* and *integrate real data with a context and purpose*. Classes with low enrollment, however, cannot collect enough data to be used in the same ways that larger classes can.

<u>Example #1</u>: Physical Exploration

When an active, concrete illustration (e.g., die rolling, card shuffling) is desirable prior to a computer simulation that demonstrates a concept, individual students can repeat the task more than once to help generate additional real data. Another alternative is to have the students complete a process and record the result just once in the classroom in order to understand the process, and then to use technology-based simulations, such as applets, to repeat the simulation

many times quickly. Alternatively, the teacher could prime the pump with a simulated data set and then add class data to those starter data.

Example #2: Project Data from the Class

One way to overcome the issue of collecting a large enough sample for use in a class-focused project is to keep records of the data collected over several semesters. Another option is to collect and share data with colleagues across multiple sections of the course. Beginning with the data collected from the class members, a conversation of the limitations of the small sample size can motivate additional data collection from a larger sample of non-classmates.

To *teach statistical thinking*, *focus on conceptual understanding*, or *foster active learning* peer-to-peer interactions are often an integral part of the educational experience. A small class necessarily means fewer peers to interact with, creating challenges for instructors.

Example #3: Cooperative Groups

Some faculty find that using cooperative groups is a great strategy for teaching statistics (see the last section of this appendix and Appendix C: *Activities, Projects, Data*). Small classes limit the size of the groups and/or the number of groups. Pairing of pairs after initial dyad discussions provides an opportunity to leverage collaboration. It is tempting in a small class to let individual students work to their strengths, rotating the group member's roles ensures all students have opportunities to lead, record, present, etc. Distributing responsibility to individual students for presentation of some of the "light" topics in the course can nurture the sense of ownership for learning among the classmates.

Example #4: Student Presentation of the Results

Fewer students allows time for students to report results from their small group (or individual) work to the entire class. Peer review/evaluation of such presentations offers additional interaction whether written or verbal, immediate or after class.

Large Classes:

While large classes provide a great opportunity for collecting large data sets, they produce their own set of challenges. For example, many excuses have been heard to not *foster active learning* through the use of groups in large classes: "the chairs do not move," "I won't be able to talk to all the groups," "it will be too loud," etc. Carbone (1998) indicates that active/cooperative learning can be effective in large classes as well as small ones and provides suggestions to *foster active learning* in large classes. For very large classes an assistant, who might even be an advanced student, could be helpful for group supervision (Davidson, 1990). In some cases there may be an opportunity to break a large statistics course into separate smaller lab or discussion sections in which group work and activities could be used.

Gelman and Nolan (2002) report that with careful selection, activities can be used successfully in large statistics classes and strongly encourage group work to promote student learning. In

particular, they suggest that when selecting activities for large classes, choose those in which the majority of students remain seated and a limited number of students go to the board or make a presentation to the class.

Regardless of the class size, involving students in the course is important. Gelman and Nolan (2002) provide an example of "Active Homework": Throughout the semester, they suggest assigning pairs of students to go to the library to find data that is needed for class or brought up in a class discussion. For a small class, by the end of the semester, the entire class could have this experience. Another approach to this same activity is to have students find data from the web to bring to class.

Example #1: Working with Partners in Positive, Productive Ways

A modification of the think-pair-share that has been recommended for large classes by Blumberg (2015) can be remembered with the acronym FSLC. These letters help the students remember to **Formulate** the answer on their own first, then **Share** it with a partner. The acronym specifically encourages important partner behaviors and teachers are encouraged to not omit the last two steps of the process which are to ensure that students **Listen** carefully to the answer of their partner and then **Create** a new answer that uses both partner's information in a manner so that the new answer is better than each of the individual answers.

Example #2: Cooperative Groups

In large, tiered lecture halls with fixed chairs, students may find it logistically easier to work in pairs instead of groups. The instructor can use a think-pair-share structure and randomly call on a student to report the thinking for their group. Asking the question and then using a random number generator to determine the student to be selected can help keep students in a large class alert. Sampling with replacement ensures students know they could be called on again at any time.

Example #3: Data Collection Using Class Polling

Zullo and Cline's book "Teaching Mathematics with Classroom Voting: With and Without Clickers" includes three chapters on using clickers in introductory statistics courses. Examples include lesson plans for box plots, hypothesis testing, confidence intervals, and data collection. Furthermore, the text describes how to select lessons that are good for using classroom voting and how to use these approaches for developing conceptual understanding. Specific examples can be found in Appendix D: *Examples of Using Technology*.

Example #4: Using Online Surveys to Maximize Class Time

With the ever increasing number of free or low-cost online tools for developing surveys, professors can maximize class time by setting up online surveys to collect data either in (via a mobile device such as a phone) or out of class as an efficient means of data collection. The article by Taylor and Doehler (2014) includes activity ideas and implementation details for using

survey software for data collection in introductory statistics. Specific examples can be found in Appendix D: *Examples of Using Technology*.

References/Resources:

Blumberg, P. (2008), *Developing Learner-centered Teaching: A Practical Guide for Faculty*. San Francisco: Jossey-Bass.

Blumberg, P. (2015), "Student Participation/Active Learning," University of the Sciences' Teaching and Learning Center. Retrieved Sept. 20, 2015. http://www.usciences.edu/teaching/tips/spal.shtml#within

Carbone, E. (1998), *Teaching Large Classes: Tools and Strategies*, SAGE Publications.

Davidson, N. (1990), "Small-group cooperative learning in mathematics," *Teaching and Learning Mathematics in the 1990s. The NCTM Yearbook*. 52-61.

Gelman, A. and Nolan, D. (2002), *Teaching Statistics: A Bag of Tricks*, Oxford: Oxford University Press.

Taylor, L. and Doehler, K. (2014), "Using Online Surveys to Promote and Assess Learning," *Teaching Statistics*, 36(2), 34-40.

Zullo, H. and Cline, K. (2012), *Teaching Mathematics with Classroom Voting: With and Without Clickers*, MAA Notes [Washington, DC]: MAA, Mathematical Association of America.

Flipped (Inverted) Classes

With audio, video, and even graphical materials easier to develop and make available on-line, opportunities for efficiently sharing materials with students are expanding. Some faculty are utilizing those types of technology to restructure in-class and out of-class learning environments. Faculty use technology to provide on-line lectures for students to listen to and learn from outside of class. Now they have evolved into videos of the teaching providing lecture-type instruction or of slides with a voiceover which uses animations to help students visualize a concept. In some cases, students watch these videos prior to class. Then the students come to the classroom ready to engage in *active learning* to solve more sophisticated problems than would be easy to solve at home or in isolation. This type of learning structure has been called the "flipped" or "inverted" classroom. The inverted classroom model offers multiple opportunities, both in and out of the classroom, for helping students develop s*tatistical thinking* and c*onceptual understanding*.

<u>Example 1</u>: Out-of-class Videos and In-class Problem Solving Sections

Lape, et al (2014) found that students who watch videos outside of class and use class time in Engineering and Mathematics for problem solving sessions believe that the class time helped them learn the concepts more than students in the corresponding traditionally taught courses.

Example 2: Motivating Reading

Wilson (2013) provides incentives for reading the textbook outside of class by giving "reading quizzes," and while only around 60% of the students rated the readings as helpful, they did significantly increase the amount of reading they did for the course as compared to when the course was taught in the traditional lecture approach. Wilson used a model where the lecture material was moved outside of class and the homework, which was presented as application-type activities that could be done individually or in groups, was completed during the scheduled class time. She found for both student-reported learning and for student final exam grades that the flipped classroom teaching model was significantly better than the traditional model. Thus, in terms of the GAISE recommendations, instructors could focus reading and the corresponding quizzes on *conceptual understanding*. Furthermore, using the flipped classroom approach including reading quizzes on concepts could, without loss of content make more time during the class period for engaging students in significant *active learning activities* (See also Appendix C: *Activities, Projects, Data*.)

Example 3: Informed Classroom Instruction

Strayer (2014) recommends teachers convey information to students outside of class to gain a response from the students prior to coming to the class. These responses should be viewed by teachers before class to better inform them how to teach the class during the face-to-face time with the students. It is particularly helpful if the teacher can construct the task to reveal students' conceptions as well as their misconceptions. Building on the knowledge gained from the out-of-class material, the teacher can be prepared to efficiently structure class discussions to extend the students' knowledge. While the task for Strayer's research was an algebra lesson for pre-service teachers, the lessons learned are transferable to flipped introductory statistics courses.

References/Resources:

Bishop, J. and Verleger, M. (2013), "The Flipped Classroom: A Survey of the Research," in *120th ASEE Annual Conference and Exposition*.

Flipped Learning Network. http://fln.schoolwires.net/Page/1

Lape, N., Levy, R., Yong, D. H., Haushalter, K. A., Eddy, R., and Hankel, N. (2014), "Probing the Inverted Classroom: A Controlled Study of Teaching and Learning Outcomes in Undergraduate Engineering and Mathematics" Paper presented at 2014 ASEE Annual Conference, Indianapolis, Indiana.

Rayens, W. (2012), "Teaching Statistical Concepts in an Inverted Classroom," University of Kentucky. http://www.causeweb.org/webinar/teaching/2012-01/

Strayer, J. (2012), "How learning in an inverted classroom influences cooperation, innovation and task orientation," *Learning Environments Research*, 15(2), 171-193.

Wilson, S. G. (2013), *Teaching of psychology: The flipped class: A method to address the challenges of an undergraduate statistics course.* Lawrence Erlbaum Associates, Inc.

Distance Learning

As technology continues to advance, the use of on-line instruction to teach statistics has also evolved. The instruction in the on-line course can be asynchronous, synchronous or partially synchronous. Some on-line courses are broadcast live to an audience that can both see and hear the professor and the teacher can see and hear all of the students in remote locations at the same time. Since sometimes an on-line class is a result of students having work schedules which make meeting face-to-face difficult, the on-line class can also take on an asynchronous format where students watch videos of the instructor or the text-book author on their own time. Another format for the on-line course is partially synchronous in which there is a combination of face-to-face meetings and on-line instruction. These courses may have different names such as hybrid, blended, or web-enhanced, and they may have different percentages of time spent in the online or the face-to-face environment. Today, there are Massive Open Online Courses (MOOCs) which provide opportunities for learning to tens of thousands of people. Moreover, the MOOCs can be led by an instructor with a set schedule for completing course materials or can be completely self-paced.

Complete definitions and best practices for each of these learning environments can be found at the Hidden Curriculum webpage (Abbott, 2014). Common themes for best practices include maximizing the strengths of each approach to foster interaction between students for discussions and collaborations regarding learning. For the partially synchronous environment, similar to the flipped classroom environment, instructors should carefully consider how to use the in-class time to maximize learning based on the goals of the course.

As faculty design online courses based on the GAISE recommendations, the following examples about technology might be useful in making decisions about the learning environment in which different content is delivered.

<u>Example #1</u>: Data Collection

Teachers can *integrate real data with a context and a purpose* in distance learning environments by using online surveys for collecting data which can be shared with the entire class. This type of data collection about the students can create interest and foster interactions.

<u>Example #2:</u> Discussion Boards

Discussion boards can be used to have students describe how they would use statistics in their major. This can help students connect with other students in their major while they are learning more about applications of statistics in the real world. Critique of journalistic efforts to report scientific research can engender online conversation even asynchronously. The instructor can set up specific "question and answer" assignments where the students can use the discussion board

to help each other better understand the material. Discussing the choice of analytical tool – by students for their coursework or by researchers whose reports are being critiqued – provides opportunity for *statistical thinking, focusing on concepts, using technology,* and *multivariable thinking.*

Example #3: Simulations

The teacher can create short videos demonstrating a simulation using an applet. Then the students can follow the example in the video to run their own simulations for similar problems, using *technology to explore concepts*. (See also Appendix D: *Technology*.) Students can be asked to post decision responses telling what they learned from a simulation to encourage *statistical thinking* and serve as an *assessment to improve and evaluate student learning*.

References/Resources:

Abbott, S. (ed) (2014), "Hidden curriculum" *The glossary of education reform*. Retrieved from http://edglossary.org/hidden-curriculum

Mills, J. D. and Raju, D. (2011), "Teaching Statistics Online: A Decade's Review of the Literature about What Works," *Journal of Statistics Education,* 19(2).

Tudor, G. (2006), "Teaching Introductory Statistics Online – Satisfying the Students," *Journal of Statistics Education*, 14(3).

Ward, B. (2004), "The Best of Both Worlds: A Hybrid Statistics Course," *Journal of Statistics Education*, 12(3).

Utts, J., Sommer, B., Acredolo, C., Maher, M., and Matthews H. (2003), "A Study Comparing Traditional and Hybrid Internet-Based Instruction in Introductory Statistics Classes," *Journal of Statistics Education*, 11(3).

Cooperative Learning

A cluster of teaching/learning techniques (with a variety of names and purposes) that involve students working together can provide opportunities for implementing GAISE recommendations into statistics courses. Team-based (St. Clair & Chihara, 2012), student-driven (Sovak, 2010), cooperative (Garfield, 1993) or collaborative (Roseth, Garfield, & Ben-Zvi, 2008) learning, and guided investigations (Bailey, Spence, & Sinn, 2013) have nuances as outlined in the given references, but all come down to opportunities to *foster active learning* in the classroom and *integrate real data with a context and a purpose*, often necessitating the use of *technology to analyze it*. The actual tasks assigned to small groups of students might incorporate the remaining recommendations by focusing on *statistical thinking* and *conceptual understanding*.

For institutions or instructors who design entire courses around these types of instruction, we provide a few more examples in addition to the larger collection in Appendix C: *Activities, Projects, Data*.

Example #1 - Histogram Comparisons

Each student is assigned a pair of histograms (out of four such pairs) for which they must determine which has more variability. They then discuss their reasoning with a partner until consensus is gained on both pairs of histograms. New partnerships are made and each student must explain the reasoning to the new partner for both their own and their original partner's histograms. In the end, every member of the foursome has four well-reasoned examples for determining the relative size of variability. *Active learning* and a *conceptual understanding* of variability are inherent in this activity.

Inspired by Roseth, Garfield, & Ben-Zvi (2008).

Example #2 – Coin Distribution

Students bring coins from home (specify pennies, nickels, etc.) and in their groups they sort them by minting dates, calculating the ages. Several descriptive graphs, tables, or measures might be made on the small collection before compiling the data from all the groups into the classroom sample and further descriptives, perhaps using technology. The use of real data in practicing the construction of dot plots, histograms, and other graphs brings *active learning* to the classroom. This might be a review of earlier descriptive topics and/or serve as a launching point for whole class discussion of sample size, limitations due to sampling methods, outliers, and/or informal inference.

Inspired by National Council of Teachers of Mathematics (2014). *Principles to Actions*. Reston, Virginia: National Council of Teachers of Mathematics.

Example #3 – NFL Quarterback Salaries

In general, a jigsaw activity gives different information to different group members so that it requires cooperation and discussion to fit the pieces together before the final question(s) can be answered. Determining the best predictor (Pass Completion %, Touchdowns, or Yards per Game) of quarterback salary through r and r^2 can be just such a task. Providing each group member with one data set to compare to the salary allows each student the practice in calculating r and r^2. In order to come to consensus on the *best* predictor; however, comparisons of numbers, graphs, and appropriate use of vocabulary is required. Follow-up questions can tap *statistical thinking* and *conceptual understanding* in addition to the *use of real data*, *active learning*, and *technology* that was used to analyze the data. There is also opportunity to address the reality of multi-variate predictors.

Inspired by STatistics Education Web, http://www.amstat.org/education/STEW/.

References/Resources:

Bailey, B., Spence, D. J., and Sinn, R. (2013), "Implementation of Discovery Projects in Statistics," *Journal of Statistics Education*, 21(3).

[NSF-funded curriculum available at http://faculty.ung.edu/DJSpence/NSF/materials.html]

Garfield, J. (1993), "Teaching Statistics Using Small-Group Cooperative Learning," *Journal of Statistics Education*, 1(1).

Roseth, C. J., Garfield, J. B., and Ben-Zvi, D. (2008), "Collaboration in Learning and Teaching Statistics," *Journal of Statistics Education*, 16(1).

Sovak, M.M. (2010), *The Effect of Student-Driven Projects on the Development of Statistical Reasoning* (Doctoral Dissertation). Retrieved from http://iase-web.org/Publications.php?p=Dissertations.

St. Clair, K. and Chihara, L. (2012), "Team-Based Learning in a Statistical Literacy Class," *Journal of Statistics Education*, 20(1).

Limited Technology

Technology has had many forms and definitions over the years. Today, certain forms of technology are considered standard in many introductory statistics courses; however, that definition of standard varies by institutions. Some classrooms for teaching statistics have statistical analysis software and software for visualizing statistical simulations on computers while other institutions might not have a computer lab large enough to seat an entire statistics class. Some courses meet in computer labs, while others might have a weekly lab sessions. At some schools, the only computer is at the instructor's station, while at others students bring their own devices to class. And unfortunately there are schools where students don't have access to any technology at all. Even for instances when an instructor feels that their classroom environment might be "technology deprived" there are still ways to provide instruction which supports the GAISE recommendations.

<u>Example #1</u>: Teacher Demonstrations

If the course instructor has access to a computer projection system, the teacher can bring a laptop into the classroom to demonstrate using statistical software to analyze large data sets and provide an opportunity for students to see that computation is the least important task of a statistician. Instructors can often receive complementary or discounted statistical analysis software. Many of these tools include instructional videos to help the student learn how to use the software outside of teacher instructional time. Teachers can also demonstrate applets which allow students to quickly observe results of a simulation that would be too time consuming using physical manipulatives. Online statistical analysis tools and applets can be found in the statistics education digital library, *CAUSEweb.org*, using the advanced search tool. (See Appendix D: Technology for additional examples.)

After demonstrating how to use statistical software, the teacher can provide the students with a handout of examples of statistical output for practice at interpreting analysis results. Students

should be able to answer questions on exams interpreting the output of statistical software, such as the question below.

> The body temperatures of a random sample of 65 healthy male college students were taken. Researchers wanted to know if the body temperature of college-age males is different from the "normal" body temperature of 98.6 degrees Fahrenheit. Use the output from a statistical software package to test the researcher's hypothesis at the 5% significance level. Then state your conclusion in the context of the problem.

Hypothesis test results:
μ : Mean of variable
H_0 : μ = 98.6
H_A : μ ≠ 98.6

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
male	98.104615	0.086669986	64	-5.7157574	<0.0001

Inspired by the ARTIST Assessment Builder, https://apps3.cehd.umn.edu/artist/user/login.asp.
Note: A free registration is required for using the Assessment Builder.

Example #2: Calculators for Statistical Analysis

When statistical software packages are not an option, graphing calculators with pre-programmed statistical functions can be used to minimize time spent by students on computation and maximize time spent on *conceptual understanding* and interpreting the statistical output in the context of the given problems. In conjunction with the demonstration of statistical software, students should understand that calculators are not the technology of practicing statisticians or researchers from other fields doing data analysis.

Example #3: Physical Manipulatives

Classrooms with limited technology should not be deprived of physical manipulatives or opportunities to actively engage students in learning statistics. For example, having students make a "living boxplot" based on student data is a great way to get the students to actually be the manipulatives and to visualize what it means to have 25% of the data in a given region. Scheaffer, et.al. (2004) provide instructions for that and other rich activities which might only require paper and a ruler to *teach statistical thinking, focus on conceptual understanding,* and *foster active learning*.

References/Resources:

ARTIST Assessment Builder. https://apps3.cehd.umn.edu/artist/user/login.asp. Assessment Builder FAQ https://apps3.cehd.umn.edu/artist/QA.html.

CAUSEweb.org. The Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) Digital Library for Statistics Education Resources, Research, and Professional Development.

Scheaffer, R., Watkins, A., Witmer, J., and Gnanadesikan, M. (2004), *Activity-based Statistics, Instructors Guide, 2nd Ed.*, revised by Erickson, T. Key College Publishing.

Texas Instrument, TI 84 Activity Central. *Statistics: Find Activities that Support your Lessons*. https://education.ti.com/en/84activitycentral/us/Statistics. Retrieved 09/22/2015.