Instructor Guide

to accompany

Introduction to Statistical Investigations

by

Nathan Tintle, Beth Chance, George Cobb, Allan Rossman, Soma Roy, Todd Swanson and Jill VanderStoep

Index

Introduction

With any book, the instructor guide serves the important role of getting inside the head of the authors and other former users of the materials to figure out what they're thinking, why they did what they did and how you and your students can make sense of it. Given the novelty of the approach that we are taking, we feel that this instructor guide serves a critical role in transitioning instructors from "That sounds great (in principle)!" to "I see that how that will work in my classroom." The instructor guide is meant to be one of a number of supporting materials that new instructors will use to help prepare for and assist during classroom implementation.

The instructor guide is organized into two main sections:

- **Curriculum wide instructor guide—**covering topics including routes through the curriculum, general classroom pedagogical options and sample syllabi
- **Section-by-section instructor guide—**covering topics including student stumbling blocks, approximate time in class, tips and tricks, technology and materials, and more, for each section of the book

In addition to this instructor guide we remind readers that the preface to the book lays out our big picture rationale for the curriculum, George Cobb's article (2007) lays out a compelling case for randomization and Tintle et al. (2011, 2012) provide assessment data supporting effectiveness of the materials (along with providing some background and motivation). The author team also invites you to participate in our ongoing support via the curriculum blog (email an author for access), or simply email an author if you have a specific question.

Thanks for taking an interest in and using these materials, and we hope that the following suggestions and advice provide you a clear path forward and effective plan for implementation.

The Author Team
August 2014
Revised: May 2015

Curriculum wide instructor guide

Routes through the curriculum

We have tried to design a curriculum that is maximally flexible to instructor preferences and institutional requirements for topics and order, while remaining true to the spirit of simulation and randomization. When designing a sequence in order to engage the topics it is important to keep in mind the following:

1. Preliminaries and Chapters 1-4 contain mainly required topics which we expect all students to complete in order prior to beginning Chapter 5. Some exceptions to this rule are as follows.
    a. Sections 1.5, 2.2, 3.2 and 3.3 which contain treatment of theory-based approaches to inference and confidence intervals for a single proportion and a single mean. These can be de-emphasized in a course focusing mainly or exclusively on randomization/simulation.
    b. Section 3.5 is not really explicitly referenced until later in the course
2. Chapter 5 (**Comparing two proportions**) introduces a basic randomization test and serves as a framework for later chapters. Section 5.3 could potentially be skipped in a course focusing mainly on randomization/simulation.
3. After completing, Preliminaries and Chapters 1-5, students can complete Chapters 6-10 in any order except that Chapter 6 (**Comparing Two Means**) should come before Chapter 7 (**Paired data**), and Chapter 6 should also come before Chapter 9 (**Comparing Multiple Means**) through 7 and 9 can be done in any order.
4. Sections 6.3, 7.2, 8.2, 9.2 and 10.5 cover theory-based approaches which can be skipped or de-emphasized in a course focusing mainly or exclusively on randomization/simulation.

Here are a few sample routes through the curriculum (others exist)

- *Full course:* Preliminaries, Chapters 1-10: 15 week, 3 hour per week class with no projects or extra software package
- *Full course:* Preliminaries, Chapters 1-10: 15 week, 4 hour per week class with projects and an extra software package.
- *Randomization/simulation only course:* Preliminaries, Chapters 1-4 (except those sections noted above), Sections 5.1-5.2, 6.1-6.2, 7.1-7.2, 8.1, 9.1, 10.1-10.4: 15 week, 3 hour per week class with projects

Pedagogical options

What should I do in class? What should I do out of class? What should I grade? These questions and others will need to be addressed as you think about structuring your class. The preface provides some background on our choice to maximize pedagogical alternatives in the

curriculum. Below we provide a couple of specific options. This is not in any way meant to be an exhaustive list but to spark your thinking about how you want to create your classroom.

- *Option #1:* Class periods are mainly interactive lecture/discussions of the examples presented in the book. The instructor leads these presentations using PowerPoint along with brief discussions between students on key questions. Students follow along with applets where appropriate. Explorations are completed partially in-class and finished outside of class. Students are provided solutions to explorations or they are discussed in class.
- *Option #2:* Students do explorations in class as a mix of "work by yourself" and "guided by the instructor" activities. Explorations are discussed in class. Examples are read by students outside of class either before or after completing the exploration. Homework exercises are assigned for additional practice.

Most of us find that we don't just "pick a pedagogy" and stick with it all semester long. Instead class periods vary, often in conjunction with the particular section we are working on or just as the instructor's mood (or prep time!) allows. The section-by-section instructor guide gives some tips on which sections we find to be particularly conducive to a type of pedagogy or approach vs. others. Of course you will need to make decisions about what to grade or not grade. Here are a few specific ideas that way:

1. Grade explorations outright, or just use a "done" vs. "partially done" vs. "not at all done" grading system done quickly at the beginning of class (walk around and mark down +, - or 0 in your gradebook)
2. Have students submit homework exercise periodically for a grade. This could be a daily assignment of a few problems reinforcing the concepts from the previous class, or you could use the Investigations and/or Research Articles as a "capstone/integrative" assignment for each chapter.
3. Daily quizzes are an option, with a short quiz at the beginning of each class. Another option is to just have a few carefully placed quizzes—if you take this latter route, key places are after Chapter 1 and after Chapter 5.

Chapter coverage by week for sample classes

Week	16 week, 3 credit course (no projects; 2 midterms, 2 quizzes, final exam)	15 week; 4 credit course (2 projects; 2 midterms, 2 quizzes, final exam)	10 week course; 4 credit course; class meets 4 times a week for 50 minutes each (3 midterms, quizzes, final exam)
1	Discuss syllabus, Prelims	Discuss syllabus, Prelims	Discuss syllabus, Prelims, Ch 1
2	Ch 1	Ch 1	Ch 1 and 2
3	Ch 1	Ch 1 and Ch 2	Ch 2, Exam #1
4	Quiz on Ch 1, Ch 2	Ch 2 and Ch 3	Ch 3 and Ch 4
5	Ch 2, Ch 3	Ch 3, Exam #1	Ch 5
6	Ch 3 (only one class due to break)	Ch 4	Ch 6, Exam #2
7	Ch 4	Ch 5	Ch 6
8	Exam #1, 5	Ch 6 (only one class due to break)	Ch 7, Ch 9
9	Ch 5	Ch 6, Project #1 presentations	Ch 8, Exam #3
10	Quiz on Ch 5, Ch 6	Project #1 presentations, Exam #2, Ch 7	Ch 8, Ch 10
11	Ch 6, Ch 7	Ch 7, Ch 8	
12	Ch 7, Exam #2	Ch 9, Ch 10	
13	Ch 8, Ch 9	Ch 10	
14	Ch 9 (only one class due to break)	Project #2 (only one class due to break)	
15	Ch 10	Project #2	
16	Ch 10		

Section-by-section instructor guide

Preliminaries

Overview
We call this part of the book Preliminaries because we mean it as a preview, not as a review or list of things to know in advance. Our goal is to orient students by using a concrete example to give them an overall sense of direction, a roadmap for the chapters that follow. Everything in the preliminaries will be re-introduced more formally, and in greater detail, later on.

We wrote these preliminaries because we are convinced that these preliminaries really matter, but we wrote it also with the hope and expectation that most instructors and students can get from it all that matters in one class period and one homework assignment. If you spend more than two days on this material, we have failed you, in a way that we originally failed some of our early class testers. Their valuable feedback has led us to a major rewrite on this section of the book. We, the authors, are grateful to them, especially Gary Kader and Lisa Kay and Julie Legler.

Student stumbling blocks

- **Make sure that students don't view the section as optional.** They shouldn't skip it or tune out. Students should think of Preliminaries as an essential preview, an agenda for the entire course.
- **Make sure that students don't think Preliminaries are a summary of what you need to know in advance.** After all, this is an *introductory* statistics book. We hope that if students know the agenda it will help them make sense of the details as they come along later, but reassure students that all the concepts presented here will come back again and again.
- **Make sure students don't get stuck on the non-intuitive result in Monty Hall problem.** We wanted to choose an example that was both compelling and interesting, but also one that is easy to simulate the probability of interest. The goal of Exploration P.3 is to see how simulation can estimate probabilities, and that probability is a relative frequency. If students are "getting it" (Monty Hall problem), quickly move them on---don't get bogged down.

Approximate class time
The goal is to cover these two examples and one exploration in no more than 2, 50 to 75 minute class periods. One option would be that on day 1 you discuss the syllabus and Example P.1. On day 2 you discuss Example P.2 and have students do Exploration P.3. If you don't quite finish P.3 during class you could have students finish for homework and do a wrap-up discussion at the beginning of class on day 3 for a few minutes.

Implementation tips and tricks

This section follows a different structure than the rest of the book. In particular, in Preliminaries, the Examples and Explorations do not cover the same material. The reason for this is to encourage instructors and students alike to quickly move through this material and into Chapter 1.

We purposefully do not provide a standard formulaic definition of the standard deviation. Remember that we will revisit the standard deviation repeatedly through the book. The important thing for students to take form Exploration P.2 is that standard deviation is one way to measure the variability in a set of quantitative data and that bigger values mean more variability. The next level of conceptual understanding would be to help students realize that the standard deviation is approximately the average of the deviations of the individual values from the mean.

You could do the game in-class for cheap prizes with randomly selected student participants and with you as host to build excitement about the “solution” (which strategy is best) to the paradox. When talking about Exploration P.3, make sure to emphasize that the first probability of winning is very intuitive and the simulation confirms it. The probability of winning if the student switches is not intuitive, and the simulation is invaluable here to discover what that probability is.

Technology and materials

- Playing cards so students can play the game themselves to simulate
- The Monty Hall simulation applet (available online and linked from the exploration).

UNIT 1: FOUR PILLARS OF INFERENCE: STRENGTH, SIZE, BREADTH, AND CAUSE

Chapter 1: Significance: How strong is the evidence?

Chapter overview
This is where things really get fun! Here students will get their first look at the logic of inference (how to draw conclusions from data), along with hypotheses, p-value, etc. etc. We find that students really "jump in" here and you want to ride the "beginning of the semester" momentum as long as possible. We've worked hard to build on students intuitive notions about drawing conclusions from data and to reduce the technical and notational overhead as much as possible. The key for students is to leave this chapter with a good sense of the general logic used to strength of evidence. Of course, as with most things in the course, these ideas will be revisited over and over again. Many of us like putting a quiz after chapter 1 as an initial in-class assessment that makes sure students are on the right track. This chapter is setting the stage for the rest of the course.

Section 1.1. Introduction to chance models

Overview
The goal of this section is to get students to understand the 3-S strategy (Statistic, Simulate, Strength of Evidence) as quickly and intuitively as possible. Students typically find this section engaging and intuitive. That's really the key here we think: Many students are thinking this will be "another math class" and, for many, that means, abstract, non-intuitive and full of rules that need to be memorized. While statistics certainly has its non-intuitive results (e.g., Simpson's paradox), we argue that's not the place to start. Start intuitive, start simple---where students gut feelings are right, and they get some quick successes which set the stage for the rest of the course.

Student stumbling blocks
Not many stumbling blocks here if you are purposeful about not using technical language or being overly critical of student language if they have the right intuition/idea about simulating chance in order to evaluate the likelihood of chance as an explanation for the observed data.

Approximate class time
One, 50-75 minute, in-class period for Doris and Buzz as guided discussion with tactile simulation by students and if Exploration 1.1 is given as homework. There could also be time to start Exploration 1.1 in class.

Implementation tips and tricks

Most of us have been doing the Doris and Buzz example as an interactive lecture discussion involving (a) explaining the experiment and results (b) having students discuss in small groups whether they think this is evidence that Dolphins can communicate (c) discussing results (d) having students brainstorm how they would convince someone who wasn't convinced that Buzz's choices were something rarely obtained by guessing (e) having each student flip a coin to simulate "just guessing" (f) each student shares individual results from guessing on board (class dot plot) (g) demonstrate applet which simulates sets of coin flips (h) draw conclusions (i) introduce 3-S strategy (j) talk about follow-up study where Buzz doesn't get it right.

This is followed by assigning Exploration 1.1 for homework. Exploration 1.1 is similar enough to Doris and Buzz that students can do it on their own if you've done Doris and Buzz in class. Important note: If you do this, however, you will need to have students skip question 12 which requires students to pool their data with their classmates. A final tip is to have students close their books while you do Doris and Buzz so they aren't reading ahead.

To do this section well you will want to very purposefully remove any of the technical language barriers that we (as experts) have a tendency to creep into our language. The idea of null and alternative hypotheses, p-value, pi, etc. are all coming up soon (in the next section in fact!), but now is not the time. Keep it simple and intuitive and you will be pleasantly surprised at how well your students do at understanding the logic of inference

At this early stage students may not be formulating their hypotheses very well, but that's OK---it will come with practice. The goal here is to build on student intuition on simulating chance and then evaluating the likelihood of the chance explanation for the observed data.

Technology and materials

- Coins to conduct in-class tactile simulation and pool results
- One proportion applet

Section 1.2. Measuring the Strength of Evidence

Overview
The goal of this section is to help students make the 3-S process more familiar, while simultaneously introducing the some of the more formal "language" of inference (null and alternative hypothesis; p-value). The important notion of "parameter" is also introduced for the first time. The notion of a less than alternative and non 50-50 null are also introduced.

Student stumbling blocks
We've tried to be purposeful in the materials to bridge students from the natural and intuitive (Section 1.1) to putting structure and rigor to the 3-S process using the typical language of statistics (hypotheses, p-value and parameter). You should model this to students also as you teach this section, trying to make connections between Doris and Buzz, the new contexts, 3-S and the lingo.

In particular, p-value is merely a convenient way of saying how extreme (in the tail) your statistic is. Hypotheses are convenient ways of expressing two possible 'true statements' about the unknown (the parameter).

The notion of parameter is a tricky one for students and you're really just beginning to introduce students to the idea---they won't get it all at once. Remind students (and yourself!) that you've got all semester to revisit the idea of parameter!

Approximate class time
The water tasting activity can be comfortably completed in a single-class period of 50-75 minutes.

Implementation tips and tricks
- The FAQ on "What p-value should make us suspicious" is a good one to help students understand why the p-value guidelines are what they are. Some of us do an in-class demo where we flip a coin about 8 times in a row (and tell students we get heads every time even though you likely won't)—no introduction, just start flipping. Then talk students through the fact that after about 4-5 heads in a row they got suspicious something was going on (two-sided coin, etc.); 4-5 heads in a row happens around 3-6% of the time, which is where we start saying "strong evidence."
- You can do the water tasting as an in-class activity where your students are the tasters.

Technology and materials
- One proportion applet

Section 1.3. Alternative Measure of Strength of Evidence

Overview
The goal of this section is to introduce an alternative to the p-value---the standardized statistic---to measure strength of evidence. This section acts as a nice place to revisit the language of Section 1.2, while noting that the p-value is really just one way to measure "extremeness."

Student stumbling blocks
Students are still getting comfortable with the notion of parameter so be careful in describing what that is in comparison to the sample proportion.

You'll want to reinforce what the standard deviation is measuring (variability of the sample proportions). Also reinforce that it is the standard deviation of the null distribution we are using. This will become more complicated in the future when we start looking at quantitative data and there will be standard deviations of other things (like the sample data) involved.

Approximate class time
The Bob and Tim activity can be comfortably completed in a single-class period of 50-75 minutes

Implementation tips and tricks
- Bob and Tim is a fun in-class activity. You may want to put the photos on a powerpoint slide and catch your students off guard to get good results, but just going through the book should work fine as well.
- We have found that around 75% of our students will choose Tim for the picture on the left. If you have similar results, you may find your results significant or not depending on the size of your class. This may be a good place to start talking about the effect of sample size on p-value if you haven't started talking about it already.

Technology and materials
- One proportion applet.
- Prepared to collect data on Bob-Tim faces

Section 1.4. What Impacts Strength of Evidence?

Overview
The goal of this section is to introduce students to three things that impact strength of evidence: (a) difference between statistic and null hypothesized value (b) sample size and (c) one vs. two-sided tests. These are listed in order of difficulty for students, though none of these are particularly challenging for students.

Student stumbling blocks
You should plan to provide some justification of the need for two-sided tests. In particular, two-sided tests are really just a convenient way to (quite arbitrarily) penalize the researcher (by doubling the p-value and keeping the criteria for assessing strength of evidence constant) for "knowing less" before the study started; thus, two-sided tests are more objective.

Approximate class time
Exploration 1.4 can be comfortably completed in a single-class period of 50-75 minutes

Implementation tips and tricks
While it may be intuitive for students to understand that increasing the sample size gives us more evidence and thus should increase the strength of the evidence (and thus lower the p-value), it is important to relate that back to the variability in the null distribution. It is nice to show pictures of how as the variability decreases in the null distribution and the observed statistic stays the same, it will be more out in the tail.

Technology and materials
- One proportion applet

Section 1.5. Inference for a Single Proportion: Theory-Based Approach

Overview
The goal of this section is to introduce students to the idea that you don't have to simulate; in fact, before computers it was quite inconvenient (and not really done!). So, instead, people ***predicted what would have happened if you had simulated.*** Using some fancy mathematical proofs, they found that they could make accurate predictions---some of the time!

Student stumbling blocks
Some students want to immediately assume this is the "right way" and forget about simulation. Be clear that simulation is most always appropriate, whereas a theory-based test has extra validity conditions. Of course, simulation isn't always appropriate (e.g., non independence of

observations, poor model of reality, etc.), but the take-home for students is that those same limitations hold for the theory-based approach.

Approximate class time

The calling heads or tails activity can be comfortably completed in a single-class period of 50-75 minutes

Implementation tips and tricks

- Take time to highlight what can go wrong when the validity conditions aren't met both in terms of skewness or "chunkiness" of a null distribution. This seems to really drive home how simulation and theory-based approaches connect.
- We don't have the binomial exact test here, but you could make that connection as well for a more sophisticated student audience. Keep in mind, however, that it adds "one more" approach for generating a p-value (note: the applet does provide the exact p-value with a button click).
- The researchers in Example 1.5 on Halloween treats were hoping to find that the null hypothesis was a plausible explanation. In other words, they were hoping for a large p-value. This is quite unusual in a study and may be worth pointing out to students.
- Some instructors will catch their students off guard before starting Exploration 1.5 and pull out a coin and tell the students to call heads or tails as the coin is tossed in the air. Then ask the students what they called as a way of collecting the data.
- The One-Proportion applet does not do a continuity correction when performing the normal approximation. This may make p-values for theory-based not match up to those of simulation-based as much as would be expected. Depending on your class, you may or may not wish to discuss this with them.

Technology and materials

- One proportion applet
- Ready to collect data on your students (would they choose heads or tails)

Chapter 2: Generalization: How Broadly do the Results Apply?

Chapter overview:
The examples and explorations we have shown in Chapter 1 dealt with process probabilities instead of population proportions. For example, we started out by looking at Buzz's probability of choosing the correct button. These repeated attempts by Buzz (if he was just guessing) could easily be modeled by the process of flipping a coin. We now want to expand our examples to involve making inferences about population proportions like those explored in national polls. That is one of the focuses of this chapter---when can we generalize our results to a larger population. Generalization (or breadth) is one of our four pillars of statistical inference. Chapter 1 covered Significance and now we will tackle generalization. This pillar makes up part of step 5, the scope of inference. This chapter shows similarities and differences of sampling from a finite population versus sampling from an infinite process. *The main goal of this chapter is to explain when and why you can (sometimes) think of choosing your sample as a chance process.* When you can, all of the results from Chapter 1 apply.

Chapter 2 has three sections. The first is about how random sampling works in practice and when you can draw inferences in the same way whether you are sampling from a process or from a population. The second section considers quantitative data, showing how random sampling suggests a way to make inferences about a population mean and that the reasoning process from Chapter 1 works for population means also. Finally, the third section reminds us that when we make inferences, there is always the chance that we will make the wrong decision, but we have some control over how often this happens. This section introduces the idea of significance level, type I and type II errors.

This chapter explores ideas that are quite different than those in Chapter 1 so there is the potential for losing momentum from Chapter 1 and leaving some students behind. We recommend thinking about having some sort of assessment (short quiz) at the end of Chapter 1, before moving on to Chapter 2 in order to make sure you haven't left any students behind (and for those you have---to get them the support they need before moving on).

Section 2.1: Sampling from a Finite Population

Overview
Sampling words from the Gettysburg Address (Exploration 2.1A) is specifically an exploration on sampling methods, not inference. Remind students that they are exploring the concept of sampling and not doing inference. The exploration looks at biased methods for sampling and unbiased methods for sampling (namely a simple random sample). The Gettysburg Address exploration is long, so if you just want to do the length of word quantitative variable and skip the categorical you may do so. This section does make more clear the distinction from a process to population. Example 2.1B and Exploration 2.1B follow up on this distinction.

Student stumbling blocks
Make sure that your students don't talk about biased samples … it is the method of obtaining the sample that is biased. As stated above, make sure they don't think we are doing inference in this section. We are just exploring sampling methods from a known population.

Approximate class time
It will probably take a couple of 50-75 minutes class periods to do cover this section completely. It could be paired down so that some topics are talked about more briefly so that it could be covered in one class period. In doing so, we would recommend you complete the Sampling Words exploration and more briefly talk about the other topics.

Implementation tips and tricks
It is nice to add a bit of history and tell your students that Edward Everett spoke for over two hours as the featured orator in the dedication of the Gettysburg Cemetery. He immediately preceded President Lincoln's two minute address. There is also a picture that you could show your students of Lincoln at Gettysburg after he gave his speech. He finished so quickly that the photographers present weren't able to get a picture of him giving the speech.

Often instructors assign Exploration 3.1B as an out -of-class assignment to be turned in for a grade at the next class period.

Some instructors have printed each of the 268 words of the Gettysburg address on 268 equal sized slips of paper as an example of an unbiased method where one would draw words out of a hat. The equal sized slips of paper make this an unbiased method.

If you have your student produce dotplots on the board of the biased method of sampling from the Gettysburg Address and the unbiased method, it is nice to do this using the same scale with one above the other. This makes for an easy comparison. You can also use a sample size of 10 for the simple random sample so that it matches up with the sample size of the biased method.

Technology and materials

- A random number generator is needed (either the one from our applet collection or another such as random.org).
- Sampling Words applet
- One Proportion applet
- Theory-Based Inference applet

Section 2.2: Inference for a Single Quantitative Variable

Overview
In Section 2.2, students further explore a single quantitative variable by looking at the mean, median, and skewness of a distribution. Simulation-based inference for a single quantitative variable is done by repeatedly sampling from a population that we think is similar to the actual population and calculating the sample mean for each sample to develop a null distribution. This is followed by a theory-based test for a single quantitative variable. While the simulation-based tests we did in Chapter 1 and the ones we will do in the future can be done in practice, this method really isn't. It is used as a segue to the theory-based method and to give some conceptual understanding as to why the theory-based method works.

Student stumbling blocks
Students seem to want to say a distribution is skewed left if the bulk of the data are on the left side of the distribution. Make clear that the direction of the tail dictates the direction of the skew.

Approximate class time
One 50 to 75 minute class period for Exploration 2.2 may not be quite enough, as the exploration on sleep times is a bit longer than most and you will probably need to have students finish this outside of class. Be prepared to discuss questions the next class period.

A brief summary of Section 2.3 can be included during this class period to cut down days needed for chapter 2. See Section 2.3 for further suggestions on this.

Implementation tips and tricks
Sketch a picture of a symmetric distribution. Add a value lower than the center yet near the center and a value higher than the center yet very far away from the center. Add a sketch of the shape of this distribution with these two extra data points. Talk about how the measures of median and mean are affected and how the shape has changed from symmetric to skewed. This helps some students solidify the connection between skewness of a distribution and position of mean and median.

Students enjoy hearing of the history of statistics. Make sure to say something about William Gosset and his student's t-distribution. There is a FAQ about Gosset in section 6.3.

Technology and materials

- Dotplot Summaries applet
- One Mean applet

Section 2.3: Errors and Significance

Overview

Up to this point, we have been focusing on the p-value indicating strength of evidence and not using it to reject the null hypothesis or not. In other words, we have not been discussing significance level yet. We do so in this section. We also discuss type I and type II errors in this section. The probability of type I errors is discussed here, while the probability of type II errors (along with power) is not discussed in detail until Section 3.5.

Student stumbling blocks

The terms type I and type II errors are not descriptive of what they are describing. We like to talk about a type I error as a false alarm to make it more descriptive. A car alarm going off when there is no problem could be thought of as a type I error and something that most students have seen or heard. We also like to describe a type II error as a missed opportunity. The alternative hypothesis is true, but we missed out on concluding it was true. You could relate this to the situation where a high school boy decides not to ask a certain girl to the prom because he determines she will say no, when in fact she was hoping he would ask and would love to go with him. He missed an opportunity here.

Approximate class time

This section can easily be done in one class period of 50-75 minutes, and potentially less (e.g., combined with a partial day on section 2.2). You could just lecture on this material and combine it with another section or some other class activity if you are pressed for time.

Implementation tips and tricks

Besides talking about type I and type II errors as false alarms and missed opportunities, you can also talk about them as false positives or false negatives and relate them to medical tests. Or another related way (as is done in the section) is with a jury trial and the conclusions of the jury that don't match up with the actual guilt or innocence of the defendant.

Technology and materials

- One Proportion applet

Chapter 3: Estimation: How Large is the Effect?

Chapter Overview
Having spent a fair bit of time in Chapter 1 with the 3-S process and understanding the logic of inference, it's now time to look at estimation and confidence intervals. There is great potential for assisting students in (a) gaining a better understanding of parameter (b) reinforcing the 3-S process from Chapter 1 even further and (c) giving students a solid understanding of confidence intervals/estimation.

We develop confidence intervals three different ways in this chapter. In the first section we will develop a range of plausible values through repeated tests of significance. This is a somewhat tedious way to construct a confidence interval, but one that makes the direct connection between the tests we did in Chapter 1 and the interval. We also present a 2SD method which is a quick "back of the napkin" approach to get an approximate 95% confidence interval. This method is a nice segue from simulation to the traditional theory-based formula that we also present.

Section 3.1 Statistical Inference: Confidence Intervals

Overview
The beauty of this section is that you can generate confidence intervals through repeated application of the 3-S process (just change the null parameterized value) on your data. Essentially, generate a "range of plausible values" for the parameter by testing different null values for the parameter and deeming some "rejected" and others "plausible."

Student stumbling blocks
We use the term *plausible value* here to represent numbers that could be the population proportion or probability. Plausible may not be a word your students use too often. You could also use the word believable to try to represent this concept. Don't use the word possible, because most any number is possible, plausible (or believable) are more restricting.

If students too quickly get to the point of "just pushing buttons" on the applet, they'll leave the section with little understanding of what they just did. It's important to walk students **slowly and repeatedly** through the idea of testing multiple null values. If they can understand that they are testing multiple null values, they will really be in great shape! Some students need encouragement. Students will often say something like, "Oh before we were looking for small p-values and now we are looking for large ones." Remind them we were never really looking for certain p-values, we were finding p-values and making some interpretation from them. We are still making the same interpretation as before, this really didn't change. A small p-value means the value under the null is not plausible and a large p-value means that it is plausible.

Approximate class time
The kissing exploration (2.1) can be done comfortably in one 50-75 minute class period.

Implementation tips and tricks
If students are doing the exploration themselves, having some one on one conversations with weaker students about why they're doing what they're doing can be very enlightening for them.

Make the connection that (as discussed in Chapter 1), a large p-value doesn't mean you've proven the null is true (this section makes that abundantly clear), but simply that it is one plausible value in a range of plausible values.

Really push to drive home the notion of parameter here; again, this section has great potential to help students "get" what a parameter is if they are pushed to think about the process they've gone to generate the range of plausible values (confidence interval).

Technology and materials
- One proportion applet

Section 3.2: 2SD and Theory-Based Confidence Intervals for a Single Proportion

Overview
This section introduces the 2SD method as a quick shortcut to find a 95% confidence based on simulation and also introduces the theory-based approach to get a normal distribution.
Finally the empirical rule is introduced as a way of showing the different multipliers that can be used for different confidence levels.

The appeal of the 2SD rule is you can very quickly generate an approximate 95% confidence interval from a simulation alone (and is a simple formula that is used with other data types). Then, the theory-based approach is merely a "prediction" of what you would have gotten had you simulated.

Student stumbling blocks
Be clear on your expectations of student use of formulas. We present formulas here, however these are not needed to calculate confidence intervals when using the Theory-Based Inference applet. You can decide how much you want your students to use or even see the formulas. For some students, the formula for a confidence interval will probably help them under confidence intervals better, and for some it will not. It is important to focus on the big picture and not have your students get bogged down in details.

Approximate class time
The exploration can be done comfortably in one 50-75 minute class period.

Implementation tips and tricks
Don't get bogged down in your own (deeper) understanding of confidence intervals and the limitations of the 2SD rule. It is a quick and dirty way to generate approximate 95% confidence intervals and is convenient pedagogically, even if not exactly right. We walk a fine line in teaching statistics in trying to balance the approximate and the precise, our feeling is this a good place to help students understanding by being approximate and focusing on the big picture.

We don't really make a big deal of the fact that the standard deviation of the simulated sampling distribution will change depending on the null hypothesized value, or that you need to pick a null hypothesized value (and do a simulation) before you can get a confidence interval using the 2SD method. We don't advise you spend a lot of time on these topics either. Bottom line: 50% null is most conservative and so is the default choice.

Technology and materials
- One Proportion applet
- Theory-Based Inference applet

Section 3.3: 2SD and Theory-Based Confidence Intervals for a Single Mean

Overview
While this section looks a 2SD confidence intervals, they are not derived from simulation as was done in the previous section. This 2SD method is just a short-cut way of finding a multiplier (namely the 2) for a theory-based confidence interval. We then move quickly to the traditional theory-based approach, but use the Theory-Based Inference applet to calculate the interval.

Student stumbling blocks
In dealing with proportions, we only had one standard deviation to consider---the standard deviation of the null distribution (or the standard deviation of the statistic). Now that we are looking at quantitative data, we also have the standard deviation of the sample data as well. Make sure you point this out to the students and they understand the difference between these two standard deviations.

Approximate class time
The exploration in this section should be easily completed in one 50-75 minute class period, and potentially less.

Implementation tips and tricks
Highlight the similarities and differences between confidence intervals for a population mean as those for a population proportion.

It might be good to take a couple of minutes and review the notation involved her for the sample and population mean.

Technology and materials
- Theory-Based Inference applet

Section 3.4: Factors that Affect the Width of a Confidence Interval

Overview
In this section we explore the factors that affect the width of a confidence interval---namely the sample size and level of confidence (with a optional subsection on how the sample proportion affects the width). We do all this mostly with confidence intervals for population proportions. In Exploration 3.4B we explore the definition of a confidence interval from the standpoint of repeated samples of the same size from a population. Emphasize to students that this helps define a confidence interval, but in practice we take one sample and interpret that one confidence interval as 95% confident that the parameter value is captured in the confidence interval constructed from the sample data.

Student stumbling blocks
Some students may still mistakenly think that as we increase confidence our interval gets smaller. Remind them that we looked at this back in Section 3.1 in developing our table of plausible values. Also just talk them through the logic a few times (like if I have a bigger net I am more likely to catch a fish) and most should be and they should be fine.

Approximate class time
Exploration 3.4A on holiday spending should go fairly quickly. If you also have your students do Exploration 3.4B completely then you may need more than one 50-75 minute class period to complete both, but not much more.

Implementation tips and tricks
While introducing new concepts, the concepts are relatively straightforward, so this a good time to step back and point out the big picture to students (strength of evidence vs. confidence intervals, etc).

Technology and materials
- One Proportion applet

- Simulating Confidence Intervals applet

Section 3.5: Cautions when Conducting Inference

Overview
In part A of this section we look at things that can go wrong if you don't have an unbiased sampling method and things that can go wrong even if you do have an unbiased sampling method. In part B we look at the difference between statistically significant and practically important. In Exploration 3.5B we also talk about power and give a way of approximating it using the One Proportion applet. This section or parts of this section can be skipped, or be assigned to students as an out-of-class assignment, or just mentioned in lecture as some extra things to be aware of in conducting observational studies.

Student stumbling blocks
Make sure they understand the different kinds of errors talked about here. We have been focusing on random errors for the most part and this type of error is acceptable. Now we are talking about non-random errors and this type of error is not acceptable and is a mistake.

Approximate class time
Depending on what you cover and how you cover it in this section, you could present these ideas fairly quickly, have them work through parts of the explorations out of class, or cover everything in class. Hence it could just take part of one 50 to 75 minute class period up to two full class periods (depending particularly how much you want to discuss power).

Implementation tips and tricks
There are a number of different topics that you can focus on in this section. Pick the ones that you think are important and focus on those and you can ignore others if you would like.

Technology and materials
- Theory-Based Inference applet
- One Proportion applet

Chapter 4: Causation: Can We Say What Caused the Effect?

Chapter overview

This chapter is about **causation**, the last of the four pillars of inference: strength, size, breadth, and **cause**. Of course to talk about causation, we need to look at studies with two variables, so while the previous chapters focused on a single variable, this chapter focuses on research studies involving two variables. In this chapter, students should see how determining causation is very difficult in observational studies because of confounding variables. They should also see that random assignment done in experiments controls for potential confounding variables and thus determining causation is possible.

The main goals of the chapter are to:

- Explore the concept of association between variables.
- Understand that confounding precludes drawing cause-and-effect conclusions from observational studies.
- Recognize the design and purpose of randomized experiments.

Section 4.1: Association and Confounding

O*verview*

This is the first time we take a good look at studies involving two variables. In this section, the main goals are for the students to understand what **explanatory and response variables** are, what it means for two variables to be **associated**, and what a **confounding variable** is. All these concepts are fairly intuitive for the students and both the example (Night Lights and Near-Sightedness) and the exploration (Home Court Disadvantage?) give the students circumstances where it is fairly easy to see the problem of confounding.

Student stumbling blocks

Students should have a fairly easy time with this section and there shouldn't be any difficult stumbling blocks.

Approximate class time

This section can be completed in a single 50-75 minute class period, with both the example and exploration possible in longer (~75-80 minute) periods, and potentially only one or the other in a 50 minute period.

Implementation tips and tricks

While we do present a two-way table in the example, we don't have students present data in two-way table in the exploration. This is done in chapter 5. However, if you wish to have your students do this, feel free. When we do use two-way tables, we have the columns represent the explanatory variable and the rows represent the response. If you are consistent doing this, it

makes it easier for students to understand the data.

Emphasize that confounding variables are associated with both the explanatory and response variables

Technology and materials

- Just a calculator used to divide is needed.

Section 4.2: Observational Studies versus Experiments

This section show students that a well-designed **experiment** uses **random assignment** to determine which **experimental units** are placed into which groups because this produces groups that are as similar as possible in all respects except for the explanatory variable. Random assignment is motivated by blocking in both the example and exploration. In the exploration (Have a Nice Trip) students initial intuition is to block on gender, in other words they want to put equal numbers of males and females in the two experimental groups. They realize that they can't block on every variable, because they can't take a measure on every variable, and this motivates random assignment. Since this random assignment will likely eliminate confounding variables, cause and effect conclusions are possible.

Student stumbling blocks

Make sure the students know what is going on in the applet used in the exploration. It may be a good idea to work the exploration in small sections making sure the students understand exactly what the applet is doing in each section.

Approximate class time

This section can be completed in a single 50-75 minute class period, with both the example and exploration possible in longer (~75-80 minute) periods, and potentially only one or the other in a 50 minute period.

Implementation tips and tricks

The word random has been used in a variety of contexts so far in the course. We talk about subjects randomly deciding between two things, we model this with flipping a coin at random, we take random samples, and we perform random assignment. It is important that students see the difference between all these things, especially random sampling and random assignment. If you emphasize the difference, they will easily understand.

Technology and materials

- Randomizing Subjects applet

UNIT 2: COMPARING TWO GROUPS

Chapter 5: Comparing Two Proportions

Chapter overview
In the previous chapter we started to look at studies that compared two groups. In this chapter, we will begin doing inference for comparing two groups. Specifically, we will be comparing two proportions, because we will consider studies in which the response variable is categorical. There should be a lot here that is familiar to students. The inference process will still follow the usual six steps. We will use the 3S Strategy to measure the strength of evidence against the null hypothesis, although the statistic will be new. We will still get an interval estimate for the parameter, now a difference in population proportions or process probabilities, using the same methods as in Unit 1. First, we simulate and use the 2SD shortcut, then we use a theory-based shortcut. There are differences, however. Our simulation process for comparing two groups involves what is called a randomization (or permutation) test. We will be using this type of test procedure in the chapters to come and it is important that the students start to get a solid understanding of it in this chapter.

Main goals:

- Perform descriptive analyses of 2×2 tables
- Understand the reasoning process of a randomization test
- Implement a randomization test for comparing proportions in a 2×2 table
- Interpret results for simulation-based and theory-based approaches to compare two proportions
- Produce and interpret confidence intervals for comparing two proportions

Section 5.1: Comparing Two Groups: Categorical Response

Overview
In this section, students learn how to organize data into **two-way tables** of counts and make a **segmented bar graphs** that shows the **conditional proportions** of success and failure across the explanatory variable groups. They should see that comparing counts of successes between two groups is not a valid comparison, because the sample sizes in the two groups could differ substantially. However, comparing conditional proportions is more appropriate and the difference between conditional proportions in the two groups is a reasonable statistic for measuring how different the groups' response. We also introduce the ratio of the conditional proportions, called the **relative risk**, and show it can also be used as a statistic to compare the two groups. The relative risk indicates how many times more likely an outcome is in one group compared to the other.

Student stumbling blocks
The one item that students tend to have difficulty with in this section is filling out a table that shows no association between the variables. Emphasize that to have no association, we need the same conditional proportions for each group and to accomplish this, these conditional proportions must be the same as the overall proportion of successes.

Approximate class time
You should be able to talk about the example and have the students work on the exploration in a single 50 to 75 minute class.

Implementation tips and tricks
Make sure they are setting up the tables in the correct way (we like to have the explanatory variables as the columns and response as the rows) then finding the appropriate conditional proportions should be easy. The next section, is a very important one in this curriculum. It might be a good idea to start foreshadowing the randomization test that is to come with the examples in this section.

Students will be intrigued by the story of Kristen Gilbert who is the nurse discussed in Exploration 5.1. Doing some background reading about her and her motives might help as the students are bound to have some questions about her.

Technology and materials

- No applets are needed---just a calculator to divide.

Section 5.2: Comparing Two Proportions: Simulation-Based Approach

Overview
While the first two chapters focused on comparing one proportion to a fixed number, we are now interested in comparing the conditional proportions (success rates) between two groups. We will continue to use the statistical investigation method to assess whether the difference between two sample proportions is statistically significant and use the same reasoning process (3S Strategy) that we introduced in Chapter 1 to assess whether the two sample proportions differ enough to conclude that something other than random chance is responsible for the observed difference in groups. However, we develop a new simulation approach to approximate the p-value for the group comparison. This approach, sometimes called a permutation test, is the same approach we will use in most of the remaining chapters of our text to develop null distributions. The basic idea to do this is to shuffle the values of the response variable, compute the simulated difference in proportions, and repeat many times. This shuffling simulates values of the statistic under the assumption of no association and with randomized experiments, this shuffling is equivalent to re-randomizing the subjects into the two groups, assuming that the

subjects' responses would have been the same regardless of which group they were assigned to. As always, the p-value is calculated as the proportion of repetitions in which the simulated value of the statistic is at least as extreme as the observed value of the statistic. We will again use the 2SD Method to produce a confidence intervals, but this time for estimating the size of the difference between the two success probabilities (or population proportions). It is important the students see the connection between the confidence interval results and the p-value. For example, if the confidence interval contains 0, then the p-value should be relatively large.

Student stumbling blocks
Sometimes students will still want to compare proportions (or probabilities) in the null hypothesis with some fixed value. For example, they might say the two population proportions are the same and they are both equal to 0.5. Make sure you emphasize that we are doing something very different here than was done in the previous chapters and are not doing any comparison to some constant.

Approximate class time
There probably won't be enough time to cover both the example and exploration in a single 50-75 minute class. However, the students should be able to complete the exploration for homework and you can review it in the next class period.

Implementation tips and tricks
This is the first time the students will see permutation as a randomization method. Take the time to go through the tactile simulation with the cards and show how this relates to what the applet does. It might be a good idea to foreshadow this process in the previous section and review it again before starting the next section.

Technology and materials
- Two Proportions applet.
- You will also need a set of 50 cards. The exploration will ask them to have 14 blue cards and 36 green cards. You can use index cards for this or you can use playing cards by just substituting red and black for blue and green. To make things a bit more concrete, you can also put names on the cards (yawn and no-yawn). This way, students don't have to related colors to the two different responses.

Other comments
This is a very important section. Make sure you provide enough class time so students understand what is going on. You can think about the scrambling approach to develop the null distribution in a number of ways (though all are equivalent). It can be shown through shuffling and dealing cards and thinking about the results in a two-way table. It can also be shown through shuffling the responses in the raw data table. It is important to have the students see this process in a number of ways to help in their understanding.

Section 5.3: Comparing Two Proportions: Theory-Based Approach

Overview
Similar to what was done in In Chapter 1, we show that we can often predict the results obtained via simulation using a theory-based approach that uses normal distributions. We also show how theory-based approaches also gave us much simpler ways to generate confidence intervals for the parameter of interest. We try to simplify and be more consistent with the validity conditions for theory-based tests and confidence intervals. We will say that theory-based techniques should give valid results if there are at least 10 successes and 10 failures for each each category of the response variable. It is important to emphasize that no matter what specific conditions you use, a larger sample size is better.

Student stumbling blocks
If students understood the theory-based techniques from Chapter 1, they shouldn't have any trouble with this section. One difficult concept for students (as well as instructors) is that of a parameter for the difference in proportions that comes from an experiment. Exactly what that is describing is certainly not concrete. Realize this in your and your students explanations of these.

Approximate class time
This section can be completed in a single 50-75 minute class period, with both the example and exploration possible in longer (~75-80 minute) periods, and potentially only one or the other in a 50 minute period.

Implementation tips and tricks
How much you want to cover formulas for the test statistic and confidence interval is up to you. You can take a minimalist approach and not even discuss them. You may want to do a bit more and just have your students understand the basic structure of these statistics, but not necessarily the details. You could also have your students do calculations with them as well.

Technology and materials
- The Two Proportions applet
- The Theory-Based Inference applet

Chapter 6: Comparing Two Means

Chapter Overview
This chapter mirrors Chapter 5, focusing on comparing two groups, with the difference being that the response variable is *quantitative* rather than categorical. We recommend highlighting to students, early and often, that the big new idea here is having a quantitative response variable and that otherwise, the same kinds of analyses and reasoning processes apply.

Main Goals:

- Perform exploratory analyses (using graphs and statistics, including five-number summary) for comparing two groups with a quantitative response variable
- Simulate randomization test for comparing two groups, using difference in means and difference in medians as relevant statistics/parameters of interest
- Estimate difference between two means with confidence interval
- Recognize role played by within-group variability, in addition to sample sizes and difference in means/medians, in both significance test and confidence intervals
- Apply two-sample *t*-test and two-sample *t*-interval, when appropriate, for conducting significance test and confidence intervals

Section 6.1: Comparing Two Groups: Quantitative Response

Overview
In this section we present exploratory (descriptive) analyses of two independent groups on a quantitative response variable. The ideas of comparing shape, center, and variability can be re-emphasized. The five-number summary is introduced for the first time, along with the use of inter-quartile range (IQR) as an alternative to standard deviation for measuring variability and boxplots as an additional way to display data.

Student stumbling blocks
Be sure to remind students about looking for shape, center, variability, and unusual observations when describing the distribution of a quantitative variable. Many students need practice and feedback with using comparative language when describing what they see, as opposed to simply providing a laundry list of features in each group separately. Also encourage students (as always) to relate their comments to the context. You might also caution students to always be on the lookout for unusual features that do not fall neatly into the “shape, center, variability” labels. For example, the haircut prices in Exploration 6.1A reveal a gap between 0 and the smallest positive values, which can be explained as some students receiving a free haircut (perhaps from friends) but then nobody getting a haircut for below some value such as $15. Exploration 6.1B can be skipped, although it tries to make the point that measures of center do not tell the whole story and that some research questions require looking at the entire distribution.

Approximate class time
This section can be completed in one 50-75 minute class meeting--perhaps less if you are looking to save time.

Implementation tips and tricks
As mentioned above, Exploration 6.1B can be skipped. For Exploration 6.1A, you might choose to have students analyze data from their own class, either after the data presented in the chapter or perhaps instead of analyzing that data.

Technology and materials

- The Descriptive Statistics applet is used in Exploration 6.1A. Alternatively, if providing experience with a commercial statistical software package is a learning goal for your course, you could have students analyze the haircut prices using a software package such as Minitab or JMP.

Section 6.2 Comparing Two Means: Simulation-Based Approach

Overview
This section is very much analogous to Section 5.2. Students are asked to simulate a randomization test for determining whether two groups differ significantly. The key difference, of course, is that the response variable is now quantitative rather than categorical.

Student stumbling blocks
The primary stumbling block here is recognizing that while the reasoning process is exactly the same as in Section 5.2, the difference is that the statistic is slightly more difficult to calculate with quantitative data. Related to this is that we can no longer use cards marked as yes/no in the simulation; we need to have cards with the actual responses values marked. And things are now slightly more complicated than simply counting the number of successes in each group and calculating a difference in proportions. We now need to (analogously) calculate the difference in *means* between the two groups. Emphasize to students that these changes are pretty minor in the grand scheme of things, and it's important to recognize that the reasoning process is exactly the same as in Chapter 5.

Approximate class time
The example and much of the exploration could be covered in a 75 to 80-minute class. With 50-minute class periods, you might want to assign part of the exploration to be done outside of class.

Implementation tips and tricks

Some students struggle to understand the response variable in the sleep deprivation study. Emphasize that the response variable is the *improvement* in reaction time. Most values are positive, indicating that most subjects did improve (react more quickly) in the second instance. But some values are negative, revealing that those subjects reacted more slowly the second time.

With the bicycle study, be sure to help students realize that the study design is not a good one, because the rider was not blind as to which bicycle type he was riding on a particular day.

With Exploration 6.2, you'll want to decide whether to have students do the hands-on part of the simulation analysis (question 9) or to save some time by skipping that part and proceeding directly to the applet simulation (question 10).

Technology and materials

- The Multiple Means applet is used for Exploration 6.2. This applet has the sleep deprivation data pre-entered, and the applet aims to help students visualize the process of combining the two groups' values together, mixing them up, and re-randomizing those values into the two groups. In other words, the applet mimics the by-hand simulation analysis that students perform earlier in the exploration.
- 21 cards with the the improvement times written on them. Alternatively, you can write values of the explanatory variable (sleep dep. or not) and response variable (improvement times) on perforated paper and have students separate the scores from the sleep assignment as a way for them to think of breaking the association. Then they would shuffle the improvement times and place into the two piles as is done with the cards.

Section 6.3 Comparing Two Means: Theory-Based Approach

Overview

Once again this section parallels Section 5.3. Now that a simulation analysis for comparing two groups has been studied in the previous section, this section presents a theory-based approach.

Student stumbling blocks

As with other theory-based approaches, a stumbling block to address is how the theory-based approach relates to the simulation-based approach. Once again you can emphasize to students that when the validity conditions are met, the theory-based approach enables us to predict what the distribution of the statistic (difference in group means, in this case) would look like if a simulation analysis were to be conducted.

Some students are also not comfortable with the t-distribution as opposed to the normal distribution. You can remind them that they first worked with the t-distribution back in Chapter 2,

and make the general point that the t-distribution arises when working with a quantitative variable.

Some students also struggle to understand that the parameter of interest here is a *difference* in population/process means. This difficulty shows up especially when interpreting what a confidence interval reveals. As in Section 5.3, the key question is typically whether the interval is entirely positive, entirely negative, or includes positive and negative values (and zero).

Approximate class time
This section can be completed in a single 50-75 minute class period, with both the example and exploration possible in longer (~75-80 minute) periods, and potentially only one or the other in a 50 minute period.

Implementation tips and tricks
You will want to decide the level of detail to present with regard to the two-sample t-test and t-interval. Formulas are given in the example. They can be ignored, used just to show their similarity of other formulas, or used in calculation.

Technology and materials
- The Multiple Means applet

Chapter 7: Paired Data: One Quantitative Variable

Chapter Overview
The main focus in this chapter is on analyzing paired data. An introduction to how we may be able to control for some of the variability in data through paired design is given in Section 7.1. In the next two sections, we focus on inference for paired data. In Section 7.2 this is done through randomization and in Section 7.3 we look at a theory-based test or the familiar matched-pairs test. Most of the analysis done in this chapter is not new. The Six Step Statistical Investigation Method and the 3S Strategy will continue to be used. The simulation process to develop a null distribution is a bit different than those used previously, but it should intuitively make sense to students.

Section 7.1: Paired Designs

Overview
The main goal of this section is to have students understand the difference between an **independent groups design** for an observational study or experiment and that of a **paired design**. They should also come away with an understanding of the advantages of paired designs. In chapter 4, students saw that to control for confounding variables we need to create groups that are as similar as possible in every aspect except the one that is manipulated by the

experimenter. A paired design does just that. Students should start to understand that this type of design can lead to more powerful tests for differences and narrower confidence intervals. We are not completing a test of significance or computing confidence intervals in this section. That is saved for the following two sections. We are just discussing design issues at this point.

Student stumbling blocks
Note that there are two different types of paired designs given here---paired design using repeated measures and paired design using matching. Give students examples of each to help them understand the difference.

Approximate class time
This section should be easily completed in a single class period of 50-75 minutes, perhaps less.

Implementation tips and tricks
Focus on the fact that paired design can often control for variability. This should be intuitive in Exploration 7.1 in which the paired design controls for differences in speeds for the different runners. You could also bring up other contexts where paired design would control for variability. For example, suppose you wanted to test the freshman 15 theory (college freshman gain about 15 pounds). Would it make more sense to get the weights of a random sample of freshmen and a random sample of sophomores to test this or weight a random sample of freshmen and then weigh them again one year later and look at the differences in all their weights?

Technology and materials
- No technology or additional materials are needed.

Section 7.2: Analyzing Paired Data: Simulation-Based Approach

Overview
In this section, students learn how to use a randomization approach to investigate whether the mean difference or change in response obtained from paired samples is statistically significant. The simulation method of developing a null distribution for paired tests is a bit different than those that were previously done. A null hypothesis with a mean difference of zero implies it doesn't matter which outcome, for each pair of outcomes, belongs to which group. Therefore to simulate what could have happened if the null hypothesis were true, we can toss a coin for every observational unit in our sample, and swap their responses if the coin lands heads (and not swap the responses if the coin lands tails). After every repetition, we record the mean difference that was obtained by chance alone. This process of shuffling and redistributing is repeated many times to get an idea of what could happen in the long run.

Student stumbling blocks
Make sure you point out the randomization test here is different than those from Chapters 5 and

6. It is also different from the simulation of binomial distributions that were done in Chapter 1. We will use the randomization test method done in Chapters 5 and 6 in the last three chapters of the book.

Approximate class time
This section can be completed in a single 50-75 minute class period, with both the example and exploration possible in longer (~75-80 minute) periods, and potentially only one or the other in a 50 minute period.

Implementation tips and tricks
Don't skip the tactile method of developing a null distribution with the students. In other words, have them flip coins and develop a null distribution in class. It is important for them to go through this process so they understand that is exactly what the applet is doing and how this process relates to the null hypothesis. The heart rate data you get from your class when they do the exploration can contain lots of variability. Some of the variability is in the actual variability of their heart rates, but some is in how they calculated their heart rates and mistakes they might have made along the way. It may give you a natural opportunity to discuss difficulties that can arise when collecting data and some precautions that should take place in the process. If you don't want to collect data from your class, there is a data set that can be used posted on the text's data webpage.

Technology and materials
- Some sort of timer to calculate their heart rates (or some sort of heart rate monitor).
- Coins
- The Matched Pairs applet
- The Multiple Means applet

Section 7.3: Analyzing Paired Data: Theory-Based Approach

Overview
Students should have noticed that the null distributions found in the previous section were bell-shaped and centered at zero---things that could be predicted. They are told in this section that the variability in these null distributions can also be predicted and that we can use a t-distribution to model these types of null distributions. Hence a theoretical matched pairs test can be done. Students should have already seen t-distributions for a single mean in Section 2.2, so this is a place where you could save some time by covering it fairly quickly.

Student stumbling blocks
It is important to keep relating the results of a test of significance with the resulting confidence interval. Some students will come up with some crazy ideas as to what should and should not be included in this interval (like the p-value or the standard deviation). Make sure they understand that the interval is our estimate for some parameter and in this case a mean

difference.

Approximate class time
Leveraging the fact that they have seen a one-sample t-test in chapter 3, you could cover this section fairly quickly. It should be able to be covered fairly easily in a single 50-75 minute class, potentially less.

Implementation tips and tricks
The exploration data set comes from an auction that involves Magic: The Gathering cards. You don't need to know anything about these cards and the cards are not the important part of the research. They are just a clever way to have a pairing to compare a couple of auction formats. However, some students in your class may have quite a bit of knowledge about these cards and it might be interesting for them to share some if it.

Technology and materials
- The Matched Pairs applet
- The Theory Based Inference applet

NOTE: At this point you may do the following four chapters (Ch 7, 8, 9, and 10) in any order, though you should do the unit introduction first.

UNIT 3: ANALYZING MORE GENERAL SITUATIONS

Chapter 8: Comparing More than Two Proportions

Chapter Overview
This chapter starts Unit 3 in which tests of significance analyze more general situations. This is one of the chapters (7 through 10) that can be done in any order. This chapter, in which we compare several proportions, is an extension of Chapter 5, where we compared two proportions. Students will learn a new statistic called the MAD (mean absolute value of the differences) for the randomization method. For the theory-based test, the students will learn a new statistic called the Chi-square. Both these statistics are used in an overall or omnibus test to see if any differences between the proportions exist. If this overall test is significant, students will need to follow up to find where exactly the differences exist. A need for an overall test is also discussed here (as will as in Chapter 9) to guard against inflating the probability of making a type I error. Both sections in this chapter deal with a binary response variable. The more general case, with multiple groups and multiple responses, is discussed in an appendix.

Section 8.1: Comparing Multiple Proportions: Simulation-Based Approach

Overview
In this section we look at inference for multiple proportions using a randomization-based method. Thinking about the null hypothesis as no association between the explanatory and response variables and the alternative as there is an association is encouraged. The statistic used to measure the overall differences in proportions is the MAD (mean absolute difference). This should be quite intuitive for the students to understand and perhaps, with a little prodding, come up with on their own.

Student stumbling blocks
There is a mean absolute deviation statistic that is also called the MAD statistic that you or your students may have heard of. This statistic is used as a substitute for standard deviation and is the mean distances values in a data set are away from their mean or median. This is not the MAD statistic we are using here. We use the term differences for the D and not deviation. Our MAD statistic is the average distance our proportions are away from each other.

Approximate class time
This section should be easily completed in a 50 minute class. If you have a longer class period (e.g., 75-80 minutes), it can even be combined with the following section, though if this is done there will not be time to talk about both examples and complete both explorations.one class period. If you are at a point where you need to save some time, you could use the example and exploration from this section to go right into the theory-based test. The applet definitely lends itself to an easy transition to theory-based. The validity conditions follow intuitively from comparing two proportions. Thus, this is a spot where you can save a day if you need to.

Implementation tips and tricks
Have students come up with the test statistic that combines information from all groups into one statistic. Some guidance may be needed after any false starts, but the students are pretty good at coming up with the MAD statistic. Make sure to point out differences between the null distribution for the MAD and other null distributions students have seen. It should be clear to the students why the null distribution starts at 0 and is not centered at 0. If you show an example where you use the MAD statistic to compare just two proportions, it should become clear as to why the distribution is right skewed.

Have your students compute the MAD statistic on a quiz or test. This should make sure they have an understanding of what it is measuring.

Technology and materials
- Multiple Proportions applet

Other
If you have students with a strong mathematics background, challenge them to calculate the probabilities of making at least one type I error when doing multiple tests comparing 3, 4, or 5 groups.

Section 8.2: Comparing Multiple Proportions: Theory-Based Approach

Overview
In this section we introduce the theory-based chi-square test. This provides a shortcut to the randomization-based test as long as validity conditions are met. We segue to the chi-square test by first using the chi-square statistic in a randomization-based test. They should see that while the null distribution for this statistic is similar to that of the MAD statistic, it is also quite different.

Student stumbling blocks
We present a different formula than is usually shown to calculate the chi-square statistic. We do so to emphasize that we are looking at finding how far proportions are from each other (or in this case, the overall proportion of successes). Feel free to talk about the more traditional formula that involves the difference between observed and expected values, especially if your students have seen this before. This formula should give the same chi-square statistic.

Approximate class time
This section should be easily completed in a 50 minute class, although the exploration may need to be completed outside of class if you only have 50 minutes. This section can be combined with Section 8.1 or assigned out of class for a short discussion the following period if you need to save a day.

Implementation tips and tricks
Many students will have seen the chi-square from biology classes. Use this to your advantage. Point out that the chi-square test statistic will generate a right-skewed null distribution just like the MAD did.

Technology and materials
- Multiple Proportions applet

Other
Point students to appendix for situations where there are three or more categories in both the explanatory variable and the response.

Chapter 9: Comparing More than Two Means

Chapter Overview
In Chapter 6 we looked at how to compare two groups with a quantitative response using the difference in means as our statistic. We expand upon this to compare more than two groups with a quantitative response using one overall test. In our simulation-based test in Section 9.1 we will again use the MAD statistic as done in Chapter 8 except this time with means instead of

proportions. If your students have already covered Chapter 8, than the simulation-based techniques should look quite similar. Make sure you stress the need for one overall test as opposed to using many tests on all the different ways you can pair the groups. Also like Chapter 8, the intuitive test statistic used in the simulation-based test (MAD) does not lead to a theory-based shortcut. So in Section 9.2 we use the F-statistic first in simulation and then in the traditional ANOVA test.

Section 9.1: Comparing Multiple Means: Simulation-Based Approach

Overview
The focus of this section is a simulation-based approach to inference for more than two groups with a quantitative response. As was done in Chapter 8, we again use the mean of the absolute differences (MAD) statistic as our intuitive test statistic except now using means instead of proportions. Students should easily see that larger values of the MAD statistic indicate a greater difference in sample means across the groups. The simulation method is again to shuffle the values of the response variable as was done in Chapters 6 and 8. The resulting null distribution, similarly to that in Chapter 8, will have a low value of zero and tends to be skewed to the right.

Student stumbling blocks
If the students have already completed Chapter 8, this section is fairly straightforward. Many of the concepts of comparing multiple proportions are similar to those of comparing multiple means.

Approximate class time
This section can be easily completed in an 80-min class and could also be covered quickly in a 50-min class.

Implementation tips and tricks
It is nice to read the ambiguous prose passage to the class before they see the picture and then read it again after they have seen the picture. This gives them a clearer picture of the research study.

Similar to the last chapter, tell your students that the MAD statistic measures the average distance the means are from each other and have them calculate this on a test or a quiz.

Make sure you cover the need for one overall test and why we don't just run a number of pairwise tests when comparing more than two means. If you didn't discuss it much in the last chapter, this is a good place to discuss it.

Technology and materials

- The Multiple Means applet

Section 9.2: Comparing Multiple Means: Theory-Based Approach

Overview

The intuitive MAD statistic used in Section 9.1 does not lead to a nice theory-based method of inference. Hence, the traditional F-statistic is used in this section as is the traditional ANOVA test. We show that the F-statistic is composed of the ratio of two measures of variability (the variability *between* groups and the variability *within* groups). Just like the MAD statistic, the F-statistic has a low value of zero and grows larger as the groups differ more. We use the following two validity conditions that need to be met in order to have valid results from the test:

- All the sample sizes are at least 20 for all the groups or that the quantitative response variable has an approximately normal distribution in the populations.
- The standard deviations of the samples are similar (the largest is no more than twice the smallest).

When the overall test reveals strong evidence that at least one of the population means differs from the others, confidence intervals for the difference in population means can be calculated for pairs of groups. This can easily be done in the Multiple Means applet.

Student stumbling blocks

The applet gives the traditional ANOVA table output. A lot of information is given here and perhaps you only want your students to focus on the p-value. Make sure they know where to look.

Approximate class time

This section could easily be done in an 80-min class and could also be completed quickly in a 50-min class.

Implementation tips and tricks

Be aware that the data set used in Example 9.2 is similar, but different than that used in Example 9.1. While Example 9.1 used comprehension scores, this one uses recall scores. Formulas for the F-statistic are given, but you can cover them at any level you see appropriate from having your students use them, to having your students understand what they are measuring in the big picture, to ignoring them altogether. If you follow this last option, you will want to skip questions 10-13 in Exploration 9.2.

Technology and materials

- The Multiple Means applet
- The Descriptive Statistics applet

Chapter 10: Two Quantitative Variables

Chapter Overview
When summarizing the relationship between two quantitative variables, we look at scatterplots and correlation (Section 10.1) as well as least squares regression (Section 10.3). We test these relationships through simulation using the correlation coefficient as our statistic (Section 10.2) and the slope of the regression line as our statistic (Section 10.3). Finally in Section 10.5 we use test these relationships with a theory-based approach.

The applet used in this chapter (Correlation/Regression), while similar to those used in the past two chapters, has more capabilities. These capabilities are explored in Section 10.3 as students discover what a least squares line is really doing to make it the line of "best fit".

Section 10.1: Two Quantitative Variables: Scatterplots and Correlation

Overview
In this section, we take a look at how to display data in a scatterplot as well as how to describe the direction, form, and strength of the relationship between two quantitative variables. We also show how the correlation coefficient can be used to describe the strength and direction of a linear relationship.

Student stumbling blocks
There shouldn't be any real stumbling blocks in this section as it is pretty straight forward. We don't present a formula for calculating the correlation coefficient, but just rely on the applet to determine its value.

Approximate class time
This section can easily be completed in a 50 minute class period. You can also combine it with Section 10.2 in a longer class period (e.g., 75-80 minutes).

Implementation tips and tricks
As mentioned above, this section can easily be combined with Section 10.2. In doing so, you could present Example 10.1 with a little explanation how the simulation-based inference works with correlation and then have the students work on Exploration 10.2.

The dinner plate example given at the beginning of Exploration 10.1 is an example of the Delboeuf illusion. Showing your students more images illustrating this illusion makes for a more interesting class.

Technology and materials

- The correlation/regression applet

Section 10.2: Inference for the Correlation Coefficient: Simulation-Based Approach

Overview
In this section we use the 3S Strategy to assess whether a sample correlation coefficient is extreme enough to provide strong evidence that the variables are associated in the population. Just as with other data types, we shuffle the values of the response variable to produce simulated values of the statistic. This shuffling simulates values of the statistic under the assumption of no underlying association between the two variables.

Student stumbling blocks
While the scatterplot shown in Figure 10.6 may appear to show no association, if you look at the corners you should be able to see a more dense collection of dots in the upper left and lower right. This is thus causing the negative association.

Approximate class time
This section could easily be done in a 50 minute class. As mentioned earlier, it can also be combined with Section 10.1 in a single, longer, 75-80 minute class period.

Implementation tips and tricks
Often researchers are hoping to determine that the alternative hypothesis is the more plausible one. In the draft lottery (Exploration 10.2), however, it is the null that is hoped to be true. It is worth pointing this out to students.

To get the median draft number for each month, you can ask your class to find the median draft number for the month they were born. Then ask who has a January birthday and then ask that person what the median draft number is for that month. Then do the same for February and so on.

It is nice to give your class some background on the draft lottery. Many of them will have never heard of it or don't really know much about it.

Technology and materials
- The correlation/regression applet

Section 10.3: Least Squares Regression

Overview
In this section we discuss how a least squares regression line minimizes the sum of the squared

vertical deviations between the observations and the line to make what we call the best fitting line. We also have students explain what the slope and intercept coefficients of the regression line mean in the context of the data. Problems with extrapolating and what extrapolation is are discussed. The square of the correlation coefficient (r^2) or coefficient of determination is also covered in both how to can be calculated with the aid of the applet and what it means.

Student stumbling blocks
Some students may need a review of lines and in particular, what the slope and y-intercept mean. They may also be used to the mathematical form of $y = mx + b$ and may need some adjustment to the statistical form of $y = a + bx$.

Understanding what the coefficient of determination means is a fairly difficult concept. Take your time in explaining this.

Approximate class time
This section can be done in one 50-75 minute class period if the primary focus is on the exploration.

Implementation tips and tricks
You could do little with the example in this section and just focus on the exploration. It is best if the students work with the applet to help them understand the concepts.

Emphasize that a residual is the *vertical* distance between the point and a line. The applet shows this nicely, but it this is also worth emphasizing.

Technology and materials
- The Correlation/Regression applet

Section 10.4: Inference for the Regression Slope: Simulation-Based Approach

Overview
In Section 10.2, we saw how to use the sample correlation coefficient in a simulated-based test about a null hypothesis of no association. In this section, we see how we can do the same type of inference, but now with the population slope as the parameter of interest and hence the sample slope as our statistic.

Student stumbling blocks
No real stumbling blocks here.

Approximate class time
This section can easily be done in one 50 to 75 minute class period. You might consider

combining it with the following section in longer class periods (e.g., 75-80 minutes)

Implementation tips and tricks
There is not much new in this section. It combines what was learned with regression along with a simulation-based test that is equivalent to what was done in Section 10.2 using correlation as the statistic.

You could easily combine this section with the next on theory-based inference. You just need to talk about how you are now using the slope as your statistic instead of correlation as was done earlier in a simulation-based test. From that you can launch right into theory-based tests.

Technology and materials
- The Correlation/Regression applet

Section 10.5: Inference for the Regression Slope: A Simulation-Based Approach

Overview
A theory-based method can be used to conduct inference for the population slope coefficient or population correlation coefficient. We consider the two methods identical, and either the observed slope or correlation coefficients could be thought of as the statistic. The applet, however, will only give a confidence interval for the slope. Formulas are given to compute the t-statistic using either the observed correlation coefficient or slope. They are given to show that the formulas give the same t-statistic.

The validity conditions for the theory-based test are a bit more complicated than validity conditions given earlier. To use a theory-based test the general pattern of the points should follow a linear trend, the response variable should have approximately the same distribution of points above the regression line as below the regression line, and the variability of the points around the regression line should be similar regardless of the value of the explanatory variable (equal variance).

Example 10.5B is included to show an situation where a validity condition is not met and the resulting null distribution can not be modeled nicely by a t-distribution.

Student stumbling blocks
The validity conditions as stated above are more complex than for other tests. It is a good idea to show graphs/pictures that might help explain them.

The Correlation/Regression applet displays a typical table, similar to one displayed with other statistical software packages, when performing a theory-based test. There is more information

given in this table than what the student needs. Make sure you point your students to look at the proper p-value in the table and you might (or might not) want to explain what the other numbers represent.

Approximate class time
You can complete this section in one 50-75 minute class period.

Implementation tips and tricks
You may have noticed that up until this section, we have not done any confidence intervals for either correlation or the regression slope. The reason is because we didn't have a validity condition that includes a linearity assumption. This issue is addressed in FAQ 10.5.1.

Technology and materials
- The Correlation/Regression applet