Chapter 13
Learning to Reason About Statistical Inference

Despite all the criticisms that we could offer of the traditional introductory statistics course, it at least has a clear objective: to teach ideas central to statistical inference.
(Konold & Pollatsek, 2002, p. 260)

Snapshot of a Research-Based Activity on Statistical Inference

Students revisit an activity conducted earlier in the semester in the unit on comparing groups with boxplots (*Gummy Bears Activity* in Lesson 2, Chapter 11). Once again, they are going to design an experiment to compare the distances of gummy bears launched from two different heights. The experiment is discussed, the students form groups, and the conditions are randomly assigned to the groups of students. This time a detailed protocol is developed and used that specifies exactly how students are to launch the gummy bears and measure the results. The data gathered this time seem to have less variability than the earlier activity, which is good. The students enter the data into *Fathom* (Key Curriculum Press, 2006), which is used to generate graphs that are compared to the earlier results, showing less within group variability this time due to the more detailed protocol.

There is a discussion of the between versus within variability, and what the graphs suggest about true differences in distances. *Fathom* is then used to run a two sample *t test* and the results show a significant difference, indicated by a small P-value. Next, students have *Fathom* calculate a 95% confidence interval to estimate the true difference in mean distances. In discussing this experiment, the students revisit important concepts relating to designing experiments, how they are able to draw casual conclusions from this experiment, and the role of variability between and within groups. Connections are drawn between earlier topics and the topic of inference, as well as between tests of significance and confidence intervals in the context of a concrete experiment.

The metaphor of making an argument is revisited from earlier uses in the course, this time in connection with the hypothesis test procedure. Links are shown between the claim (that higher stacks of books will launch bears for farther distances), the evidence used to support the claim (the data gathered in the experiment), the quality and justification of the evidence (the experimental design, randomization, sample size), limitations in the evidence (small number of launches) and finally, an indicator of how convincing the argument is (the P-value). By discussing the idea of the

J.B. Garfield, D. Ben-Zvi, *Developing Students' Statistical Reasoning*,

P-value as a measure of how convincing our data are in refuting a contradictory claim (that the lower height resulted in farther distances), students see that the farther they are from this contradictory claim, the more likely we are to win our argument. As they have seen in earlier uses of informal inference throughout the course, the farther in the tails, the smaller the probability of observing what was seen in the sample if the contradictory claim is true and the smaller the P-values. So they link small P-values with convincing evidence and a more convincing argument.

Rationale for This Activity

Unlike many of the topics in previous chapters of this book, there is little empirical research on teaching concepts of inference to support the lessons described in this chapter. However, there are many studies that document the difficulties students have reasoning and understanding inferential ideas and procedures. Therefore, we are much more speculative in this chapter, basing our lessons and activities more on writing by influential statistics educators as well as general research-based pedagogical theories. Later in this chapter, we address the many questions we have about appropriate ways to help students develop good reasoning about statistical inference and some promising new directions that are just beginning to be explored.

This particular activity is introduced near the end of a course that is designed to lead students to understand inferences about one and two means. We use it at a time where the material often becomes very abstract and challenging for students, a time where it is often hard to find a motivating activity for students to engage in. Now that students have already conducted this experiment, they are more aware of the need to use good, consistent protocols for launching gummy bears, to decrease the variability within each condition, and to provide a convincing argument supporting their claim and refuting the alternative claim. Also, now that students are acquainted with formal methods of making statistical inferences, they can do a statistical comparison of the difference in distances using a two-sample test of significance. The use of the argument metaphor helps students connect the confusing terminology used regarding hypothesis tests to something they can understand and relate to, and builds upon earlier uses of this metaphor and associated terms throughout the course.

The Importance of Understanding Statistical Inference

> Drawing inferences from data is now part of everyday life but it is a mystery as to why and how this type of reasoning arose less than 350 years ago.
>
> (Pfannkuch, 2005b, p. 267)

Drawing inferences from data is part of everyday life and critically reviewing results of statistical inferences from research studies is an important capability for all adults. Methods of statistical inference are used to draw a conclusion about a particular population using data-based evidence provided by a sample.

Statistical inference is formally defined as "the theory, methods, and practice of forming judgments about the parameters of a population, usually on the basis of random sampling" (Collins, 2003). Statistical inference "moves beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and the conclusions are uncertain" (Moore, 2004, p. 117). There are two important themes in statistical inference: parameter estimation and hypothesis testing and two kinds of inference questions: generalizations (from surveys) and comparison and determination of cause (from randomized comparative experiments). In general terms, the first is concerned with generalizing from a small sample to a larger population, while the second has to do with determining if a pattern in the data can be attributed to a real effect.

Reasoning about *data analysis* and reasoning about *statistical inference* are both essential to effectively work with data and to gain understanding from data. While the purpose of exploratory data analysis is exploration of the data and searching for interesting patterns, the purpose of statistical inference is to answer specific questions, posed before the data are produced. Conclusions in EDA are informal, inferred based on what we see in the data, and apply only to the individuals and circumstances for which we have data in hand. In contrast, conclusions in statistical inference are formal, backed by a statement of our confidence in them, and apply to a larger group of individuals or a broader class of circumstances. In practice, successful statistical inference requires good data production, data analysis to ensure that the data are regular enough, and the language of probability to state conclusions (Moore, 2004, p. 172).

The Place of Statistical Inference in the Curriculum

The classical approach to teaching statistical inference was a probability theory-based explanation couched in formal language. This topic was usually introduced as a separate topic, after studying data analysis, probability, and sampling. However, most students had difficulty understanding the ideas of statistical inference and instructors realized something was wrong about its place and portion of the curriculum. For example, an important part of Moore's (1997) plea for substantial change in statistics instruction, which is built on strong synergies between content, pedagogy, and technology, was the case to depart from the traditional emphasis of probability and inference. While there has been discussion on whether to start with means or proportions first in introducing inference (see Chance & Rossman, 2001), there has been some mention about ways to bring ideas of inference earlier in a course. The text book *Statistics in Action* (Watkins et al., 2004) does a nice job of introducing the idea of inference at the beginning of the course, asking the fundamental question - 'is a result due to chance or due to design', and using simulation to try to address this question.

We believe that ideas of inference should be introduced informally at the beginning of the course, such as having students become familiar with seeing where a sample corresponds to a distribution of sample statistics, based on a theory or

hypothesis. Thus, the informal idea of *P*-value can be introduced. These types of informal inferences can be part of units on data and on distribution (does this sample represent a population? would it generalize to a population?), comparing groups (do the observed differences lead us to believe there is a real difference in the groups these samples represent?), sampling (is a particular sample value surprising?), and then inference (significance tests and confidence intervals). By integrating and building the ideas and foundations of statistical inference throughout the course, we believe that students should be less confused by the formal ideas, procedures, and language when they finally reach the formal study of this topic; however, there is not yet empirical research to support this conjecture. We also recommend revisiting the topic of inference in a subsequent unit on covariation, where students build on applying their inference knowledge to test hypotheses about correlation coefficients and regression slopes.

Review of the Literature Related to Reasoning About Statistical Inference[1]

> Historically, there were huge conceptual hurdles to overcome in using probability models to draw inferences from data; therefore, the difficulty of teaching inferential reasoning should not be underestimated.
>
> (Pfannkuch, 2005b, p. 268)

Difficulties in Inferential Reasoning

Research on students' informal and formal inferential reasoning suggests that students have many difficulties in understanding and using statistical inference. These results have been obtained across many populations such as school and college students, teachers, professionals, and even researchers. Many types of misunderstandings, errors, and difficulties in reasoning about inference have been studied and described (e.g., Carver, 1978; Falk & Greenbaum, 1995; Haller and Krauss, 2002; Mittag & Thompson, 2000; Oakes, 1986; Vallecillos and Holmes, 1994; Wilkerson and Olson, 1997; Williams, 1999; Liu, 2005; Kaplan, 2006). In addition to studies documenting difficulties in understanding statistical inference, the literature contains studies designed to help explain why statistical inference is such a difficult topic for people to understand and use correctly, exhortations for changes in the way inference is used and taught, and studies exploring ways to develop students reasoning about statistical inference.

---

[1] We gratefully acknowledge the contributions of Sharon Lane-Getaz as part of her dissertation literature review with Joan Garfield.

Survey Studies on Assessments of Students' Understanding Statistical Inference

In a study of introductory students' understandings about "proving" the truth or falsity of statistical hypotheses, Vallecillos and Holmes (1994) surveyed more than 400 students from different fields who responded to a 20-item survey. One of the interesting results in this study was that nearly one-third of the answers reflected a faulty belief that hypothesis tests logically prove hypotheses. Additional misunderstandings were found among introductory statistics students at the end of a one-semester introductory statistics course by Williams (1997, 1999). Williams interviewed eighteen respondents and found that statistical ideas of P-values and significance were poorly understood. In an earlier study, Williams (1997) identified several sources of students' misunderstanding of P-values such as inadequate or vague connections made between concepts and terms used, and confusion between P-value and significance level. Williams (1999) also found that many introductory students believed that the P-value is always low.

To assess graduate students' understanding of the relationships between treatment effect, sample size, and errors of statistical inference, Wilkerson and Olson (1997) surveyed 52 students. They found many difficulties students had, such as misunderstanding the role of sample size in determining a significant P-value. Similar results were documented in a study by Haller and Krauss (2002), who surveyed instructors, scientists, and students in psychology departments at six German universities. The results showed that 80% of the instructors who taught courses in quantitative methods, almost 90% of instructors who were not teaching such courses, and 100% of the psychology students identified as correct at least one false meaning of P-value (Haller and Krauss, 2002).

Additional difficulties in reasoning about inference were identified such as confusion about the language of significance testing (Batanero et al., 2000) and confusion between samples and populations, between α and Type I error rate with P-value (Mittag & Thompson, 2000). In sum, survey studies have identified persistent misuses, misinterpretations, and common difficulties people have in understanding of inference, statistical estimation, significance tests, and P-values.

Students' responses to inference items were described as part of an examination of data from a national class test of the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS – delMas et al., 2006). A total of 817 introductory statistics students, taught by 28 instructors from 25 higher education institutions from 18 states across the United States, were included in this study. While the researchers found a significant increase in percentage of correct scores from pretest to posttest on items that assessed understanding that low P-values are desirable in research studies, ability to detect one misinterpretation of a confidence level (95% refers to the percent of population data values between confidence limits), and ability to correctly identify the standard interpretation of confidence interval, there were also items that showed no significant gain from pretest to posttest. For these items, less than half the students gave correct responses, indicating that students did not appear to learn these concepts in their courses. These items included ability to detect

two misinterpretations of a confidence level (the 95% is the percent of sample data between confidence limits, and 95% is the percent of all possible sample means between confidence limits), and understanding of how sampling error is used to make an informal inference about a sample mean. There was also a significant increase in students selecting an incorrect response (26% on pretest and 35% on posttest), indicating that they believed that rejecting the null hypothesis means that the null hypothesis is definitely false. In addition, although there was statistically significant gain in correct answers to an item that assessed understanding of the logic of a significance test when the null hypothesis is rejected (37% correct on the pretest to 47% correct on the posttest), there were still more than half the students who answered this item incorrectly on the posttest.

Why Is Statistical Inference so Difficult to Learn and Use?

Reasoning from a sample of data to make inferences about a population is a hard notion to most students (Scheaffer, Watkins & Landwehr, 1998). Thompson, Saldanha and Liu (2004) examined this difficulty, noting that literature on statistical inference "smudges" two aspects of using a sample.

> The first aspect regards attending to a single sample and issues pertaining to ensuring that an individual sample represents the population from which it is drawn. The second aspect regards the matter of variability amongst values of a statistic calculated from individual samples. The two aspects get "smudged" in this way: (1) we (researchers in general) hope that people develop an image of sampling that supports the understanding that increased sample size and unbiased selection procedures tend to assure that a sample will look like the population from which it is drawn, which would therefore assure that the calculated statistic is near the population parameter; (2) we hope that people develop an image of variability amongst calculated values of a statistic that supports the understanding that as sample size increases, the values of a statistic cluster more tightly around the value of the population parameter.
>
> (Thompson et al., 2004, p. 9)

Thompson et al. (2004) state that they see ample evidence from research on understanding samples and sampling that suggests that students tend to focus on individual samples and statistical summaries of them instead of on how collections of samples are distributed. There is also evidence that students tend to base predictions about a sample's outcome on causal analyses instead of statistical patterns in a collection of sample outcomes. They view these orientations as problematic for learning statistical inference because they appear to "disable students from considering the relative unusualness of a sampling process' outcome" (Thompson et al., 2004, p. 10). These authors report on a study that explored students developing reasoning about inference in two teaching experiments in high school mathematics classes that involve activities and simulations to build ideas of sampling needed to understand inference. They found that those students who seemed to understand the idea and use a margin of error for a sample statistics had developed what. Saldanha and Thompson (2002) called a "multiplicative conception of sample" – a conception of sample that entails recognition of the variability among samples, a hierarchical image of collections of samples that simultaneously retain their individual composition, and the idea that each sample has an

associated statistic that varies as samples varied. This study suggested that if students could be guided to develop this reasoning, they would be better able to understand statistical inference. Indeed, Lane-Getaz (2006) developed a visual diagram to help students develop this type of reasoning that has been adapted and used in the lessons in this book (*Simulation of Samples Model*, see Chapters 6 and 12).

Other studies designed to reveal why students have difficulty learning statistical inference have examined how this reasoning develops and offer suggested ways to help students move toward formal inference (e.g., Biehler, 2001; Konold, 1994b; Liu, 2005; Pfannkuch, 2006a).

Using Simulation to Illustrate Connections Between Sampling and Inference

Recent research suggests that improving the instruction of sampling will help students better understand statistical inference (e.g., Watson, 2004). This can be done by using good simulation tools and activities for teaching sampling distribution and the Central Limit Theorem (e.g., delMas et al., 1999; Chance et al., 2004).

However, using these simulation tools is not enough; they need to be linked to ideas of statistical inference. Lipson (2002) used computer simulations of the sampling process and concept maps to see how college students connected sampling concepts to statistical inference. She found that while the simulations appeared to help students understand some aspects of sampling distributions, students did not appear to be linking these ideas to hypothesis testing and estimation. In a subsequent study, Lipson, Kokonis, and Francis (2003) devised a computer simulation session to support the development of students' conceptual understanding of the role of the sampling distribution in hypothesis testing. The researchers identified four developmental stages through which students progress while using the visual simulation software: (a) *recognition* of the software representations, (b) *integration* of the three concepts of population, sample, and sampling distribution; (c) *contradiction* that the sample may not be typical of the hypothesized population, and (d) *explanation* of results from a statistical perspective. A stumbling block for the students appeared to be that that they looked for a contextual explanation rather than a statistical explanation, even when they acknowledged the low probability of the sample coming from hypothesized population. The researchers concluded that current software supported the recognition stage only, and suggested that students need to have a substantial experience in thinking about samples and sampling.

Some statistics educators (e.g., Biehler, 2001; Gnanadesikan et al., 1987; Jones, Lipson & Phillips, 1994; Konold, 1994b; Scheaffer, 1992) advocate that inference should be dealt with entirely from an empirical perspective through simulation methods to help students understand how statistical decisions are made. One such approach is the *resampling* method. Konold (1994b) used his *DataScope* Software (Konold & Miller, 1994) tool to introduce resampling methods to help students develop a more intuitive idea of a *P-value*. Mills (2002) summarizes papers that give examples of how simulation can be used to illustrate the abstract ideas involved in confidence intervals; however, it is difficult to locate research studies that document the impact of these methods on students' reasoning.

Informal Reasoning About Statistical Inference

A topic of current interest to many researchers as well as teachers of statistics is informal inferential reasoning rather than formal methods of estimation and tests of significance (e.g., Pfannkuch, 2005a). As new courses and curricula are developed, a greater role for informal types of statistical inference is anticipated, introduced early, revisited often, and developed through use of simulation and technological tools.

Informal Inferential Reasoning is the cognitive activities involved in informally drawing conclusions or making predictions about "some wider universe" from data patterns, data representations, statistical measures and models, while attending to the strength and limitations of the drawn conclusions (Ben-Zvi et al., 2007). Informal inferential reasoning is interconnected to reasoning about distribution, measures of centre, variability, and sampling within an empirical enquiry cycle (Pfannkuch, 2006a; Wild & Pfannkuch, 1999).

Rubin et al. (2006) conceptualize informal inferential reasoning as statistical reasoning that involves consideration of multiple dimensions: properties of data aggregates, the idea of signal and noise, various forms of variability, ideas about sample size and the sampling procedure, representativeness, controlling for bias, and tendency. Bakker, Derry, and Konold (2006) suggest a theoretical framework of inference that broadens the meaning of statistical inference to allow more informal ways of reasoning and to include human judgment based on contextual knowledge.

Using the Logic of an Argument to Illustrate Hypotheses Testing

Ben-Zvi (2006) points out that informal inference is closely related also to argumentation. Deriving logical conclusions from data – whether formally or informally – is accompanied by the need to provide persuasive explanations and arguments based on data analysis. Argumentation refers to discourse for persuasion, logical proof, and evidence-based belief, and more generally, discussion in which disagreements and reasoning are presented (Kirschner, Buckingham-Shum, & Carr, 2003). Integration and cultivation of informal inference and informal argumentation seem to be essential in constructing students' statistical knowledge and reasoning in rich learning contexts. This view is supported by Abelson (1995), who proposes two essential dimensions to informal argumentation: The act or process of deriving conclusions from data (inference), and providing persuasive arguments based on the data analysis (rhetoric and narrative).

Part of making a statistical argument is to know how to examine and portray the evidence. In statistical inference, this means understanding how a sample result relates to a distribution of all possible samples under a particular null hypothesis. Therefore, one type of informal inference involves comparing samples to sampling distributions to get a sense of how surprising the results seem to be. This type of informal reasoning is based on first having an understanding of sampling and sampling distributions (see Chapter 12).

Students' Dispositions Regarding Statistical Inference

Another important research topic is students' dispositions and their relation to statistical proficiency. Kaplan (2006) studied the extent to which differences in psychological dispositions can explain differences in the development of students' understanding of hypothesis testing. Kaplan investigated undergraduate students who have taken an algebra-based statistics course. She used large samples to find relationships between statistics learning and dispositions and smaller samples to uncover themes and common conceptions and misconceptions held by undergraduate statistics students. No relationships were found between the statistics learning and the dispositions that were studied: "Need for Cognition," and "Epistemological Understanding." The research did identify three emergent themes in the student discussions of hypothesis testing: how students consider the experimental design factors of a hypothesis test situation, what types of evidence students find convincing, and what students understand about P-values.

Teachers' Understanding of Statistical Inference

Content and pedagogical-content knowledge of statistics teachers have a considerable influence on what and how they teach in the classroom. Liu (2005) explored and characterized teachers' understanding of probability and statistical inference, and developed a theoretical framework for describing teachers' understanding. To this end, she analyzed a seminar with eight high school teachers. Liu revealed that the teachers experienced difficulties in understanding almost every concept that is entailed in understanding and employing hypothesis testing. Beyond the complexity of hypothesis testing as a concept, Liu conjectured that teachers' difficulties were due to their lack of understanding of hypothesis testing as a tool, and of the characteristics of the types of questions for which this tool is designed. Although the teachers were able to root the interpretation of margin of error in a scheme of distribution of sample statistics, some of them were concerned with the additive difference between a population parameter and a sample's estimate of it. This study revealed a principle source of disequilibrium for these teachers: They were asked to develop understandings of probability, sample, population, distribution, and statistical inference that cut across their existing compartments.

Implications of the Research: Teaching Students to Reason About Statistical Inference

> Deepen the understanding of inferential procedures for both continuous and categorical variables, making use of randomization and resampling techniques.
>
> (Scheaffer, 2001)

The research suggests that understanding ideas of statistical inference is extremely difficult for students and consists of many different components. Many of these

components themselves are difficult for students to understand (e.g., sampling distributions). Simulation and resampling methods are viewed as having the potential to offer a way to build informal inferences without focusing on the details of mathematics and formulas. In addition, using data sets and questions in early data analysis units to have students consider informal inferences (e.g., what does this sample suggest about the population, what do we believe about the difference in means for these two groups that these two samples come from) may help develop formal ideas of inference in later units.

In studying the difficulties students have reasoning about statistical inference, many different types of errors and misunderstanding have been identified, as well as a detailed description about what it means to reason about different aspects of statistical inference. Being aware of the complexities of the ideas as well as the common misunderstandings can help teachers be on the alert for student difficulties through formal and informal assessments that can be used for diagnostic purposes.

Some of the ideas related to correct (and incorrect) reasoning about two aspects of statistical inference: P-values and confidence intervals have been detailed by the *Tools for Teaching and Assessing Statistical Inference Project* (see http://www.tc.umn.edu/~delma001/stat_tools/). For example, some common misconceptions about P-values and confidence intervals are summarized as follows:

Misconceptions about P-values

- A P-value is the probability that the null hypothesis is true.
- A P-value is the probability that the null hypothesis is false.
- A small P-value means the results have significance (statistical and practical significance are not distinguished).
- A P-value indicates the size of an effect (e.g., strong evidence means big effect).
- A large P-value means the null hypothesis is true, or provides evidence to support the null hypothesis.
- If the P-value is small enough, the null hypothesis must be false.

Misconceptions about Confidence Intervals

- There is a 95% chance the confidence interval includes the sample mean.
- There is a 95% chance the population mean will be between the two values (upper and lower limits).
- 95% of the data are included in the confidence interval.
- A wider confidence interval means less confidence.
- A narrower confidence interval is always better (regardless of confidence level).

Suggestions for Teaching Statistical Inference

As mentioned at the beginning of this chapter, there is little empirical research on the effectiveness of different instructional strategies, sequences of activities, or

technological tools in helping students develop correct reasoning about statistical inference. However, there are many strong and often conflicting beliefs among statistics educators about optimal methods of teaching these ideas. Arguments have been made for teaching inferences on proportions before means, teaching confidence intervals before tests of significance, not teaching students the method of pooling variances in comparisons of two-sample means, and abandoning t-tests altogether and instead using resampling and randomization methods. We describe below some of the suggestions that we believe to be aligned with the approaches described in our book and which we have used to build our suggested sequences of activities, acknowledging that they are not necessarily based on empirical research studies, and that their effectiveness is untested at this point.

Connecting Statistical Inference to Data Collection, Description, and Interpretation

Rossman and Chance (1999) offer "Top Ten" list of recommendations for teaching the reasoning of statistical inference. Their goal is to help students to focus on investigation and discovery of inferential reasoning, proper interpretation and cautious use of results, and effective communication of findings. The list includes the following recommendations:

1. Have students perform physical simulations to discover basic ideas of inference.
2. Encourage students to use technology to explore properties of inference procedures.
3. Present tests of significance in terms of P-values rather than rejection regions.
4. Accompany tests of significance with confidence intervals whenever possible.
5. Help students to recognize that insignificant results do not necessarily mean that no effect exists.
6. Stress the limited role that inference plays in statistical analysis.
7. Always consider issues of data collection.
8. Always examine visual displays of the data.
9. Help students to see the common elements of inference procedures.
10. Insist on complete presentation and interpretation of results in the context of the data.

Presenting Statistical Inference as Argumentation

A more recent approach to teaching statistical inference is to connect these ideas to the making of an argument, as described earlier by Ben-Zvi (2006). The logic of arguments can be used to explain the reasoning of a hypothesis test as follows:

- In statistics, we argue about claims (hypotheses) we believe to be true or false. While we cannot prove they are true or false, we can gather evidence to support our argument.
- A hypothesis test can be viewed as a method for supporting an argument.

- An argument (hypothesis test) may originate from two different perspectives: wanting to argue *against* a claim (i.e., the null hypothesis) or wanting to argue for a claim (i.e., the research (alternative) hypothesis).
- Just as in real life, even if we convince someone by our argument, we are only convincing them with evidence, we cannot really establish if our claim is actually true or not. In a hypothesis test, we only decide if the evidence is convincing enough to reject the null hypothesis, but not *prove* it is true or false.
- In order to make a good argument, we need four building blocks:

  1. A clear claim we are making (and a counterclaim that includes all other possibilities).
  2. Data to support our argument.
  3. Evidence that the data are accurate and reliable, not misleading.
  4. A good line of reasoning that connects our data to our argument.

- In real life when we make an argument, the resolution is that we win or lose the argument based on how convincing our argument is. This is based on the strength of our evidence, and how we use the evidence to support our case. In a hypothesis test, the result is to reject or fail to reject the null hypothesis, which is based on the size of the obtained P-value.
- We need to see how far away our data are from the claim we are arguing against. Therefore, we look for data that are far from what we would expect if the claim we are arguing against is true. A low P-value results from data that are far from the claim we are arguing against, and the lower (farther) they are, the stronger the evidence.

Introducing the idea of an argument would seem to be a useful way to help students understand the process of making and testing hypotheses, and may help students better understand this complex and often counterintuitive procedure.

Basing Inference on Simulation and Randomization

While many educators have advocated the use of simulations to help students understand the connections between sample, population, and sampling distribution in inference, to illustrate the abstract ideas of confidence interval (e.g., Mills, 2002) others have suggested that traditional approaches to inference be replaced entirely with resampling methods (e.g., Simon, Atkinson, & Shevokas 1976; Simon, 1994; Konold, 1994b). More recently, in light of flexible and accessible technological tools, educators such as Cobb (2007) and Kaplan (2007) have suggested radically different approaches to statistical inference in the introductory course. Their suggestions place inference as the focus of a course that teaches three R's: Randomize data production, Repeat by simulation to see what's typical, and Reject any model that puts your data in the tail of the distribution (see Cobb, 2007). We find these ideas very appealing but have not yet explored ways to build a sequence of lessons around them and experimented with them in our classes.

Progression of Ideas: Connecting Research to Teaching

Introduction to the Sequence of Activities to Develop Reasoning About Statistical Inference

The sequence of ideas and activities for inference represent one of many possible ways to guide students to develop good inferential reasoning, and we do not have a strong conviction that this sequence is an optimal one. Although we have used these lessons and find them to work well in engaging students, we believe that it might be better to adopt more of an informal and conceptual approach, rather than leading students to learn the formal aspects of testing hypotheses and constructing confidence intervals. However, we provide examples in this chapter of how to build lessons about inference on the previous big ideas and activities, and make connections between foundational concepts and the formal aspects of statistical inference.

We suggest that ideas of informal inference are introduced early in the course and are revisited in growing complexity throughout the course. Underlying the development of this inferential reasoning is a fundamental statistical thinking element, *consideration of variation* (Moore, 1990; Wild & Pfannkuch, 1999), and how variability of data and samples is a key part of making inferences. This means that students have opportunities to see and describe variability in samples throughout the course as they make informal inferences about how these samples relate to the population from which they were drawn, and whether these samples lead us to infer about what that population might be. When ideas of formal inference are eventually introduced, they are devoid of computations and formulas so that students can focus on what the ideas of null and alterative hypothesis mean, the idea of P-value, and types of errors. The computer is used to run tests and generate confidence intervals before students see the formulas. The culmination of this progression of ideas is giving students a set of research questions and associated data and having them use their statistical thinking to choose appropriate procedures, test conditions, arrive at conclusions, and provide evidence to support these conclusions.

In addition to the progression from informal to formal methods of statistical inference, we suggest the use of two important pedagogical methods. One is the modeling by the teaching of statistical reasoning and thinking in making statistical inference. This means, making their thinking visible as they go from claims to conclusions, checking conditions, considering assumptions, questioning the data, choosing procedures, etc. The second is the use of the argumentation metaphor for hypothesis testing as described earlier. This means using the language of arguing about a claim, whether we believe a claim is true, the role of evidence and using that evidence well, and what it takes to be convinced that the claim is true or false. Table 13.1 shows a suggested series of ideas and activities that can be used to guide the development of students' reasoning about statistical inference.

Table 13.1 Sequence of activities to develop reasoning about statistical inference[2]

Milestones: Ideas and concepts	Suggested activities
Informal ideas prior to formal study of statistical inference	
• Making inferences and generalizations from a sample of simulated data	• One Son Activity (Lesson 1, Statistical Models and Modeling Unit, Chapter 7)
• Statistical inference as an argument	❖ An informal discussion early in a course about the nature of statistical inference, and comparing this to making an argument and providing evidence to support your claim. (The symbol ❖ indicates that this activity is not included in these lessons.)
• Random sample and how it is representative of a population	• The Gettysburg Address Activity (Lesson 3, Data Unit, Chapter 6)
• Results being due to chance or due to design (some other factor)	• Taste Test Activity (Lesson 4, Data Unit, Chapter 6)
• As a sample grows, the characteristics become more stable, that with more data you can better generalize to a population	• Growing a Distribution Activity (Lesson 1, Distribution Unit, Chapter 6)
• Two samples of data may or may not represent true differences in the population	• Activities in Lessons 1–4, Comparing Groups Unit (Chapter 11)
• When comparing groups, you must take into account the variability between groups relative to the variability within each group	• Gummy Bears Activity (Lesson 2, Comparing Groups Unit, Chapter 11)
• If the normal distribution provides a good model for a data set we may make inferences based on the Empirical Rule	• Normal Distribution Applications Activity (Lesson 3, Statistical Models and Modeling Unit, Chapter 7)
• We can make inferences by comparing a sample statistic to a distribution of samples based on a particular hypothesis	• Activities in Lessons 1 and 2, Samples and Sampling Unit (Chapter 12)
Formal ideas of statistical inference	
• Hypothesis test as making an argument	• Modeling Coin Tosses Activity (Lesson 1: "Testing Statistical Hypotheses")
• Hypothesis test, null and alternative hypothesis	• Balancing Coins Activity (Lesson 1)
• The idea of a P-value	• P-values Activity (Lesson 2)
• Types of errors and correct decisions	• Types of Errors Activity (Lesson 2)
• What is needed to test a hypothesis?	• Types of Errors and P-values Activities (Lesson 2)

[2] See page 391 for credit and reference to authors of activities on which these activities are based.

Table 13.1 (continued)

• Confidence interval as an estimate of parameter, with margin of error	• Introduction to Confidence Intervals (Lesson 2)
• Understanding how confidence intervals may be presented in different ways	• Introduction to Confidence Intervals (Lesson 2)
• Understanding what 95% refers to in a confidence interval	• Estimating with Confidence, Estimating Word Lengths, and What Does the 95% Mean Activities (Lesson 3: "Reasoning about Confidence Intervals")
• A statistically significant difference between two groups where randomization of conditions has taken place	• Gummy Bears Revisited Activity (Lesson 4: "Using Inference in an Experiment")
Building on formal ideas of statistical inference in subsequent topics	
• Statistically significant correlation coefficient	• Activities in Lesson 3, Covariation Unit (Chapter 14)
• Statistically significant regression slope	• Activities in Lesson 3, Covariation Unit (Chapter 14)
• There are many types of statistical inferences, and software may be used by correctly choosing the commands	• Research Questions Involving Statistical Methods Activity (Lesson 5: "Applying Methods of Statistical Inference")
• Understanding that the interpretation of *P*-values and confidence depends on assumptions being met	• Research Questions Involving Statistical Methods Activity (Lesson 5)

Introduction to the Lessons

There are five lessons on statistical inference that begin with informal ideas and lead to running tests of significance and confidence intervals on the computer. The focus is on understanding the ideas and methods and interpreting the results, rather than on formulas and computing test statistics. The lessons proceed very slowly, building on informal ideas from previous lessons and also integrating ideas of argumentation. The final lesson provides students with an opportunity to think statistically and to integrate and apply their knowledge, as they are given only research questions and a data set and need to answer the questions using the data and software.

Lesson 1: Testing Statistical Hypotheses

This lesson uses the context of balancing a coin on its edge to introduce formal ideas of testing hypotheses. The proportion of heads obtained when a balanced coin falls is used to test a null distribution based on equally likely outcomes. The idea of the P-value is examined visually and conceptually, and then P-values are found using simulation software. The argumentation metaphor is used to explain the logic of testing hypothesis. Student learning goals for this lesson include:

1. Connect informal to formal ideas of statistical inference.
2. Introduce the process and language of significance tests.
3. Use *Sampling SIM* to conduct an informal test of significance.
4. Understand the use of P-value in a test of significance.

Description of the Lesson

In the *Modeling Coin Tosses* activity, the instructor holds up a penny and asks what students expect if the coin is tossed. It is agreed while the outcome of a toss is unpredictable, that they expect a fair penny to land with Heads up half the time and with Tails up half the time. Students make a conejcture about what would happen if they balance a coin on its edge and let it fall, and if this is done many times, would it also land Heads and Tails in fiarly equal numbers. They are asked how to determine if a balanced coin is just as likely to land Heads up as it is to land Heads down.

Students discuss in pairs and then write down possible numbers of Heads they might expect to get for 8 sets of 10 tosses of a fair penny (e.g., list the number of Heads out of 10 for eight repetitions of this experiment). They are asked whether they expect to get 5 Heads each time, or if they expected some variability between results of each set of 10 tosses, and how variable they expected each set of 10 to be in the number of Heads produced. Students also reason about what outcomes they would consider to be less likely if using a "fair" coin and why.

Next, students use *Sampling SIM* to model tossing a fair coin ten times. They sketch the resulting distribution of sample proportions and describe it in terms of shape, center, and spread. Students shade in areas of the distribution that include what they would consider to be surprising results, so that if they obtained one of those results, they might question the assumption that the coin is equally likely to land Heads up or down (probability of Heads is 0.5).

In the *Balancing Coins* activity, students are asked what they think will happen if they balance sets of 10 pennies on their edge and let them fall, and if they expect the same number of Heads and Tails when flippoing a coin ($p = 0.5$). They are introduced to the idea of testing a *statistical hypothesis*, as shown below:

> Idea 1: Balancing a coin is a "fair" process: Heads and Tails are equally likely to result.
> Idea 2: Balancing a coin is an "unfair" process: There will be a higher percent of Heads or Tails.

These ideas are then written as statistical hypotheses:

> Null hypothesis: The proportion of Heads when we balance a coin repeatedly is 0.5.
> Alternative hypothesis: The proportion of Heads when we balance a coin repeatedly is not 0.5. (In other words the proportion is more, or less, than 0.5.)

The *null hypothesis* is discussed as an *idea of no difference* from the norm or prior belief (e.g., getting the same results as tossing fair coins). The *alternative hypothesis* is discussed as a statement that there will *not* be an equal number of Heads and Tails, something contrary to the first idea.

Students are told that we gather evidence (data) and determine whether or not it supports the null hypothesis or whether it provides convincing support for an alternative hypothesis. To do this, students design an experiment to lead them to make a decision about which of the two hypotheses are supported by the data. They discuss what is needed to test a hypothesis or to make a good argument given this context:

1. *A hypothesis to test* (e.g., the proportion of Heads is 0.5) (The claim).
2. *A sample of data which gives us a sample statistic* (e.g., a sample proportion).
3. *A sampling distribution* for that statistic (based on the null-hypothesis) so we can see how unusual or surprising it is, by seeing if it is far off in one of the Tails (surprising) or in the middle (not surprising). This sampling distribution is based on the null hypothesis and the sample size for our sample data. If our sample statistic is in one of the Tails, that would lead us to reject H_0 (A method to test the claim).
4. *A decision rule*: how far is far off in the Tails? How far in one of the Tails does our sample statistic need to be for us to decide it is so unusual and surprising that we reject the idea stated in H_0, that the coin is equally likely to land Heads up or Heads down when we balance it? (How to evaluate the strength of the evidence.)

Students then get in groups and balance coins, counting the result when the coins fall. The numbers of Heads and Tails are tallied, proportions of Heads for each set of 10 balances are found and gathered for the class. The sample proportions typically range from 0.5 to 0.9.

The next discussion regards an appropriate sampling distribution to use to judge whether their results are due to chance or whether the chances of getting Heads when balancing a coin is greater than 0.5. They decide to refer to the simulation created earlier (in the *Modeling Coin Tosses Activity*, Lesson 1), which allows a comparison of their sample statistics to what they would expect if the coin is equally likely to turn up Heads or Tails when balanced. Students use their sketches made earlier in the activity to determine whether or not this result is in a tail. They mark the sample proportion for their group in the graph and discuss whether they think this result is surprising, and why or why not. This leads to an examination of what percent of the distribution has values more extreme than theirs. They use *Sampling SIM* to find this area.

This value is discussed as the chance of getting the result students got or a more extreme one, and is referred to as a *P-value.* The role of the *P*-value in making

a decision is seen as helping determine which of the two hypotheses seems more likely. Students discuss how small a P-value must be to be judged surpassing and leading them to reject the null hypothesis. Again, the argument metaphor is used, and the P-value is described as an indicator of how convincing the evidence is against the claim (null hypothesis). The farther it is in the tails, the more we are convinced that the null hypothesis (claim) is false. So the smaller the P-value, the stronger is the evidence. Students evaluate their P-values and determine whether they reject the claim that the coin is equally likely to land Heads up or Heads down when balanced on its edge. The class then combines their data to get a better, more stable estimate of the proportion of Heads, and test this result using the *Sampling SIM* software and finding the P-value via simulation.

Students are asked what conclusion can be drawn about the original research question, and then apply the same procedure in determining whether or not they believe a Euro coin is equally likely to land Heads up or down when tossed, using data from a series of 100 tosses of a Euro coin.

A wrap-up discussion reviews the process of hypothesis testing (hypotheses, data-sample statistic, sampling distribution, and decision rule) and how this process maps to making a convincing argument. The *Simulation of Samples (SOS)* Model is revisited and used to map the different levels of data: population, sampling distribution, and sample value.

Lesson 2: P-values and Estimation

This lesson builds on the previous lesson, using the context of balancing coins to test hypothesis and learn the language of tests of significance. This lesson also introduces the idea of a confidence interval, helping students see the two parts of the interval (e.g., sample statistic and margin of error) and different ways or reporting confidence intervals. Students also begin to interpret a confidence interval. Student learning goals for this lesson include:

1. Review use of simulations for inference.
2. Review the process for hypothesis testing.
3. Learn about the two types of errors when conducting tests of significance.
4. Use *Fathom* to conduct a test of significance.
5. Understand the idea of a confidence interval as a way to estimate a parameter.

Description of the Lesson

After balancing pennies in the previous lesson, students are asked if they think that balancing the Euro coin will yield equally likely chances of getting Heads and Tails. In the *P-values* activity, they are given a sample result from a person who balanced a Euro 100 times and got 31 Heads. First, students repeat the process they used earlier, of finding this sample proportion and comparing it to the simulated

sampling distribution for a null hypothesis of equally likely outcomes. Next they use *Fathom* software to find P-values without simulation. These two P-values are compared and students reason about why the P-value from the simulation is not exactly the same as the one produced by *Fathom* (*Sampling SIM* ran 500 simulations while *Fathom* is basing their result on the true sampling distribution of all possible samples).

In the *Types of Errors* activity, students review the steps of the previous lesson (Lesson 1 on *Testing Statistical Hypotheses*) discussing the components needed to test a hypothesis and how these compare to the argumentation process. They map the types of errors (Type 1 and Type 2) to this context of balancing the Euro coin. For example:

1) We select H_a but it is the wrong decision because H_0 is true (Type 1 error).
2) We select H_0 but it is the wrong decision because $H_0$is not true (Type 2 error).

Another context is provided and students try to reason about what the different types of errors would mean in that context and the importance of keeping the chance of making these errors small. The idea of alpha as the chance of making a Type 1 error is contrasted to the idea and role of the P-value, and what is meant by the term "statistically significant." This term is compared to winning an argument because the evidence is strong, and compelling. However, winning an argument by presenting strong evidence may also result in an error, if the claim being disputed is actually true. So this parallel is drawn to rejecting a hypothesis when it is actually true.

The next activity, *Introduction to Confidence Intervals*, examines what happens after a null hypothesis is rejected. In this case, balancing a Euro will result in an equal number of Heads and Tails. Students are referred back to the Euro data and make a conjecture about the proportion of Heads they would expect to find in a large number of repetitions of this experiment. When students give different answers or ranges of answers, it is suggested that because we are unsure about giving a single number as our estimate, due to variability of our sample data, we might feel more confident about offering a range of values instead. Students are asked what interval, or range of values, might give an accurate estimate of possible values for this "true" proportion of Heads when a Euro coin is balanced on its edge and falls down. To move to the formal idea of a confidence interval, students are given the following news clip to read:

> *A recent poll of people in the military stated: While 58% say the mission (being in Iraq) is clear, 42% say that the U.S. role is hazy. The survey included 944 military respondents interviewed at several undisclosed locations throughout Iraq. The margin of error for the survey, conducted from Jan. 18 through Feb. 14, 2006, is* $\pm$ *3.3 percentage points.*

Students are guided to use the information stated above to obtain an interval estimate for the percentage of *all people in the military* who believe the mission is hazy. They construct a confidence interval using this information. They see that they need two pieces of information that are given in this article: This information is then related back to the problem of finding a confidence interval for the proportion of Heads when balancing a Euro coin. This includes:

- A *sample statistic* (e.g., the class proportion of Heads when balancing coins), and,
- A *margin of error* (an estimate of how much this statistic varies from sample to sample for a given sample size, calculated from the sample data and information from the sampling distribution for the sample statistic).

Students are shown two ways to present confidence intervals:

- The *sample average, plus or minus a margin of error* (e.g., estimating the average textbook price for statistics, $\$80 \pm \$15$).
- The *two endpoints* (low and high values) of the interval. (e.g., \$65–\$95).

The relationship of the confidence level to the idea of error is examined, and students reason about what a confidence interval tells about estimating a parameter and possibly making an error about that estimate. Students see that a confidence interval provides two kinds of information: an *interval estimate* for the population parameter (rather than a single number estimate) and a *level of confidence* (how confident we are that our interval includes the population value we want to estimate).

A wrap-up discussion includes what the term "margin of error" means, and how this term is used when interpreting results from a poll. Students describe the sample and the population for the survey reported above and critique it, referring back to material from the unit on Data related to designing good surveys (Lessons 1 and 2 in the unit on Data, Chapter 6). Students also consider and discuss different interpretations of the poll results, such as: Can we use our interval to give a guess about the true percentage of all people in the military that believe the mission is hazy? How? How sure are we? Are there any problems with generalizing from our sample of 944 military respondents to all people in the military?

Lesson 3: Reasoning About Confidence Intervals

This lesson helps students develop their reasoning about confidence intervals by using simulation to make and test conjectures about factors that affect confidence intervals. They also have opportunities to discuss common misconceptions as they critique interpretations of confidence intervals. Student learning goals for this lesson include:

1. Develop reasoning about confidence interval.
2. Understand what 95% *confident* actually means.

3. Understand how sample size and confidence level affect the length of the confidence interval.
4. Become familiar finding a confidence interval using *Fathom* software.
5. Understand connections between confidence intervals and hypothesis tests.

Description of the Lesson

In the *Estimating with Confidence* activity, students return to the question from the previous lesson: "What is the true (expected) proportion of Heads when a Euro is balanced?" Now that they believe that the proportion of Heads when a Euro balanced is *not* equal to 0.5, then what is it? Students now know the idea of a confidence interval. *Fathom* is used to produce a confidence interval for the sample of data based on balancing a Euro coin. The class discusses how to interpret this result and are asked what type of estimate might be more informative about the location of the actual population parameter, a narrower or wider interval, and why.

Connections are then made between testing a hypothesis and estimating with a confidence interval, and students see how a confidence interval can be used to test a hypothesis. Students make a conjecture about how the confidence interval would be different if they had only 50 pieces of data rather than 100, and then if they had 1,000 data values and why. This conjecture will be examined later in a simulation activity. Students reflect on the previous unit on sampling and distinguish between the sample statistic and a population parameter for the Euro coin example, and how much they would expect a sample statistic to vary from a population parameter.

In the *Estimating Word Lengths* activity, students return to the *Gettysburg Address* activity from the unit on Data (Lesson 3 in Chapter 6) in which they sampled words from the Gettysburg Address. They use the Gettysburg Address as a population and take samples and construct confidence intervals to see how they behave and how to interpret them. They use the Gettysburg Address Web applet to take a random sample of 25 words and then use *Fathom* to find a 95% confidence interval to estimate the true mean word length for all of the words in the Gettysburg Address. Next, the students draw their confidence intervals on the board, one on top of another. These intervals are compared to the true population mean word length, and students examine how many of the intervals generated by the class overlap the true population mean. Students are asked what percentage of all the intervals in the class they would expect to *not* overlap the population mean and find it is close to what they have generated.

The next activity (*What Does the 95% Mean*?) leads students use *Sampling SIM* to make and test conjectures about confidence intervals. They sample data from different populations such as a normal curve as well as for a skewed distribution, which is shown in Fig. 13.1.

Students generate 200 95% confidence intervals for samples of size 25 and examine how many do not include the population mean (shown as red lines) and how close the proportion of intervals that include the mean is to 95% (see Fig. 13.2).

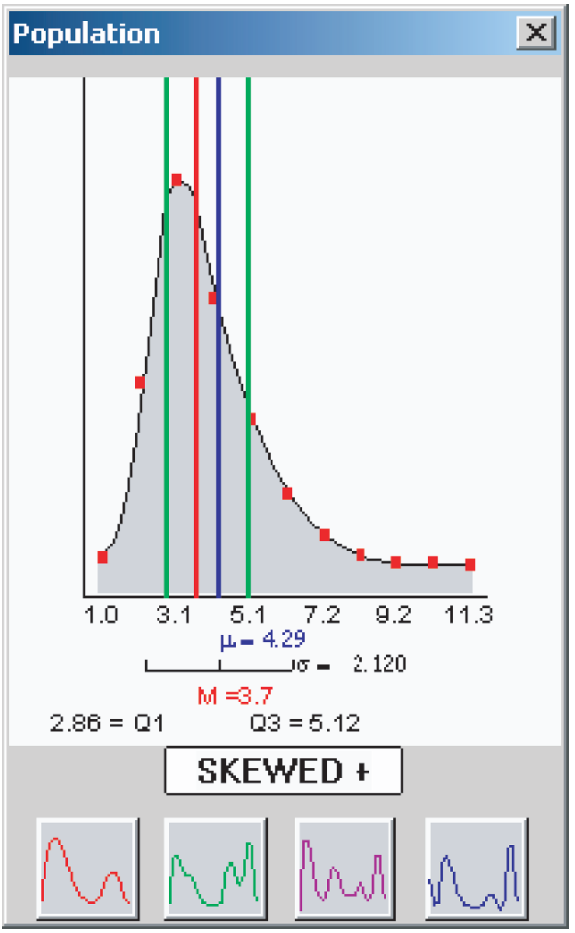

Fig. 13.1 A right-skewed population produced by *Sampling SIM*

They use the results from *Sampling SIM* to help answer the following questions that target common misconceptions about confidence intervals:

1. Does the level of confidence, 95%, refer to the percent of data values in the interval?
2. Does the level of confidence, 95%, refer to the location of the *sample mean* or locating the *population mean*? Explain.
3. Does the level of confidence, 95%, refer to a *single interval* (e.g., the one you found in *Fathom*) or to the *process or creating many intervals* (e.g., all possible intervals)? Explain.

Next, students use the *Sampling SIM* to make and test conjectures about what factors affect the width of the confidence interval. They then test these conjectures by increasing and decreasing the level of confidence, and changing the sample size,

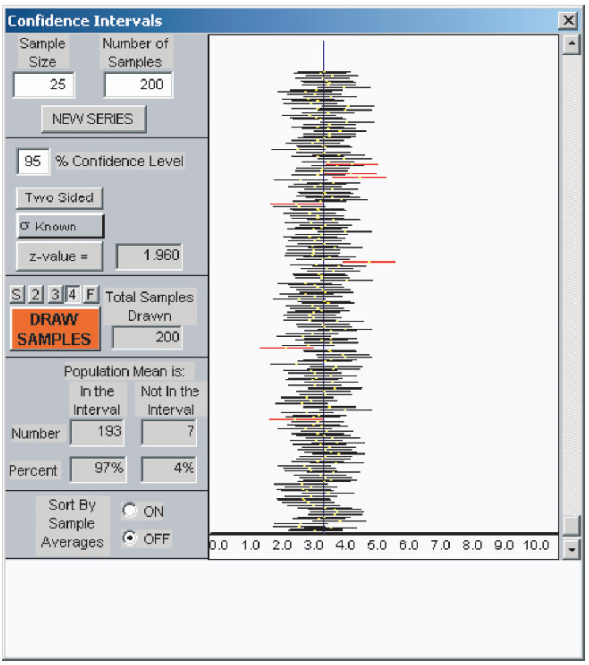

Fig. 13.2 Two hundred 95% confidence intervals (sample size 25) from a right-skewed population in *Sampling SIM*

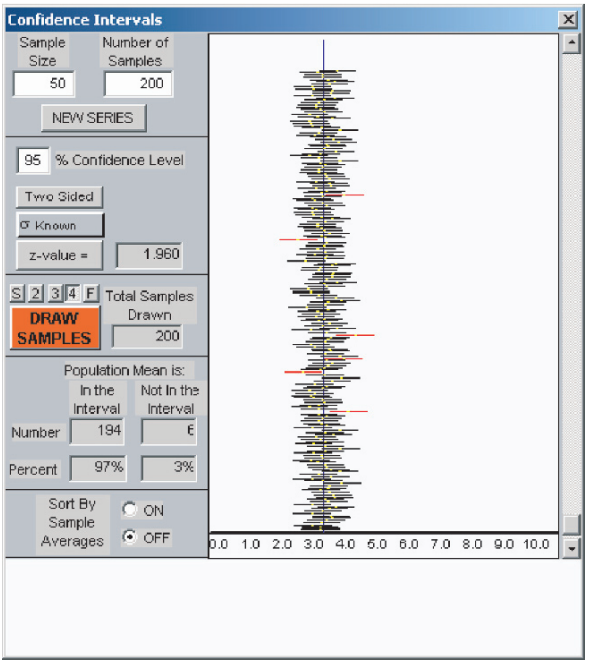

Fig. 13.3 Two hundred 95% confidence intervals (sample size 50) from a right-skewed population

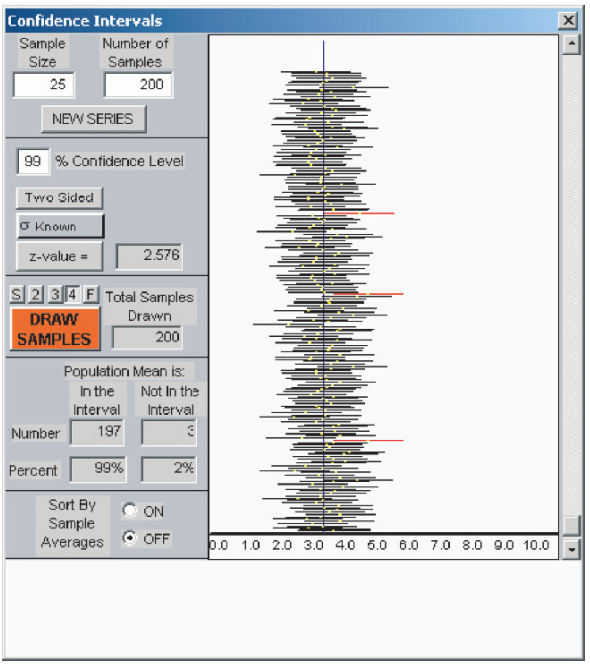

Fig. 13.4 Two hundred 99% confidence intervals (sample size 25) from a right-skewed population

generating new simulated intervals each time. See Fig. 13.3 for larger sample size and Fig. 13.4 for larger confidence level.

A discussion follows about what type of width (narrow or wide) gives the most precise estimate of the population parameter, and what level of confidence (lower or higher) most often includes the true population parameter being estimated.

A wrap-up discussion includes when and why we use a confidence interval in a statistical analysis and why we say "95% confident" instead of "95% probability." Students consider why and how confidence intervals and hypothesis tests are connected, and what is unique about each approach and the information it provides.

Lesson 4: Using Inference in an Experiment

This lesson described at the beginning of this chapter revisits an earlier experiment, giving students a chance to try to reduce within group variation and better detect a difference in the two conditions. Data are gathered and analyzed first graphically and then using *Fathom* to run a two sample t-test. The logic of hypothesis tests and comparison to making an argument are revisited for this context. Student learning goals for the lesson include:

1. Understand the idea of a two-sample hypothesis test.
2. Differentiate between a one-tailed and a two-tailed test.

3. Use *Fathom* to conduct a two-sample test.
4. Understand the idea of a two-sample confidence interval (difference in means).
5. Use *Fathom* to conduct a confidence interval to estimate a difference in means.
6. Revisit the ideas of designing an experiment and making cause and effect inferences.
7. Revisit ideas of within and between group variation and how they affect a two sample comparison.
8. Revisit ideas of how to reduce variation within a condition, and ideas of signal and noise in repeated measurements within each condition.

Description of the Lesson

In the *Gummy Bears Revisited Activity*, students reflect on the earlier *Gummy Bear* activity (Lesson 2 in the Comparing Groups Unit, Chapter 11) and discuss how to determine if there is a difference between two conditions in an experiment, in this case, if there are different average launching distances for the one book or four book launching pads. Students are asked, in light of recent discussions and activities on statistical inference, to suggest how, if a difference in sample means is observed, this is not just due to chance.

The students redo the experiment after first discussing a careful and systematic protocol to follow in launching the Gummy bears. Treatments are assigned to groups and each group produces data for 10 launches. Students use *Fathom* to produce side by side boxplots, discussing what the boxplots suggest about the differences in flight distances for the two conditions. Students are asked how to determine if the observed difference in group means is statistically significant and what this means. The null and alternative hypotheses are constructed and *Fathom* is used to run the test. Students contrast one and two tailed tests for this experiment, and run the test both ways using Fathom, contrasting the difference in results. Students explain what the results of the hypothesis test suggest about the difference between the two launching heights. Next, students use a confidence interval to estimate the mean difference in average launch. They discuss what it means if a difference of 0 is in the interval or is not in the interval. Since 0 was not in the interval, they concluded that this is a statistically significant difference in flight distances.

In a wrap-up discussion, students suggest reasons to use a one-tailed or two-tailed test of significance, and advantages and disadvantages of each method. They reason about how the type of test (one or two tailed) affects the P-values obtained and which method is more conservative. Finally, students give a full statistical conclusion about the comparison of flight distances for short vs. high launching pads.

Lesson 5: Solving Statistical Problems Involving Statistical Inference

This lesson comes at the end of a course, after the study of covariation (see Chapter 14) and helps students connect and integrate concepts and processes in statistical

inference, developing their statistical thinking. Student learning goals for the lesson include:

1. Review the process of conducting and interpreting a test of significance.
2. Review the process for finding, reporting, and interpreting confidence intervals.
3. Review the conditions/assumptions that are necessary for our inferences to be valid.
4. Be able to research questions to appropriate inferential procedures.
5. Practice using *Fathom* to conduct tests of significance and to find confidence intervals.
6. Be able to interpret and justify results of statistical inferences.

Description of the Lesson

Discussion begins by looking back at the first few days of the course when students simulated data to estimate whether a sample statistic might be due to either chance or to some other factor. For example, if a student was able to correctly identify Coke of Pepsi in a blind taste test vs. the student was a lucky guesser. The discussion then proceeds to when students learned how to use *Fathom* to generate P-values and confidence intervals to help in making inferences and decisions about population parameters. Now that software can be used to generate statistical results for inferences, students consider the decisions that have to be made, for example:

a. What type of analysis to run (e.g., test or estimate, one or two samples, etc.).
b. What conditions to check.
c. How to interpret the results (and also know if we made a mistake).

In the *Research Questions Involving Statistical Methods* activity, students are reminded that the computer will generate P-values for tests of significance and construct confidence intervals for population parameters, even if the conditions are not checked and met. The class discusses how one should interpret the results of a procedure where the conditions are not met. Next, students are given the following table (Table 13.2) to discuss and complete it together, which will serve as a guide for running different analyses to produce inferential statistics in *Fathom*.

Students are then given a set of research questions (as shown below in Table 13.3) and a data set to use in answering the questions, using *Fathom* software. The data set contains the body measurements for a random sample of 50 college students. First, the instructor models statistical thinking, talking out loud and demonstrating the questions and steps and interpretations involved in answering one or two of the questions on the list below. Working together, students then discuss each question, select appropriate procedures, test conditions, generate graphs and analyses, and interpret their results.

Table 13.2 A guide for running different analyses to produce inferential statistics in *Fathom*

Type of procedure	Example of research question	Fathom instructions
One sample confidence interval for proportion	What is the proportion of college students who graduate in 4 years from your school?	
One sample confidence interval for a mean	What is the average number of credits earned by students when they graduate with a bachelor's degree?	
One sample hypothesis test for a proportion	Is the proportion of students who withdraw during their first year equal to 0.15 (The proportion who withdrew 5 years ago)? Is the proportion of students who withdraw during their first year less than 0.15?	
One sample hypothesis test for a mean	Is the average number of years it takes to finish a degree equal to 5? Is the average number of years it takes to finish a degree greater than 4?	
Two sample confidence interval for the difference between two means	What is the difference in the average number of hours spent studying each week between physics majors and English majors?	
Two sample hypothesis test to compare two means	Is there a difference in the mean GPAs of first year and fourth year students?	

Table 13.3 Selecting appropriate procedures and hypotheses to given research questions

Research question	Type of procedure	Null and alternative hypothesis (if appropriate)
What proportion of students in this class has a larger arm span than height?		
What is the average hand span for students in this class?		
What is the difference in hand spans for males and females?		
Is the average height for female students greater than 163 cm?		
Is the proportion of male students who are taller than 172.72 cm different from 0.5?		
Is there a difference between males and females in head circumference?		

After students complete this work, group answers to the questions are shared and justified.

Summary

Most students studying statistics encounter great difficulty when they reach the topics of statistical inference. Some instructors have compared student response to lecturers on this topic as "the deer in the headlight" phenomena, as students seem frozen, confused, and scared when learning these difficult topics. The research literature documents the difficulty students have understanding inference, and typical misconceptions that persist regarding P-values and confidence intervals.

Although many statistics courses put statistical inference at the end of a first course in statistics, we have illustrated a research-based approach that first presents informal ideas of inference early in the class and revisits these ideas again and again, so that when the formal ideas are introduced later they are more intuitive and easier to understand. The idea of statistical hypotheses as making arguments is used to help make this difficult topic more accessible to students. At the end of the course, students are given a set of research questions and need to integrate and apply all that they have learned to determine what procedures are needed and appropriate, to provide answers, and to justify their conclusions. This process is first modeled by their instructor and then they have the opportunity to use and develop their own statistical thinking by approaching these questions as statisticians, rather than just solving a series of textbook problems for each given procedure. This approach also differs from more standard approaches because the computational procedures are not emphasized. Instead, the process of using the computer to test hypotheses and estimating parameters is stressed, along with how to do this wisely and how to justify and interpret results.