Chapter 6
Learning to Reason About Data

> *High on my list of elements of statistical thinking is the claim that data beat anecdotes. This is surely a learned principle, and one much neglected by public opinion.*
> *(Moore, 1998, p. 1257)*

Snapshot of a Research-Based Activity on Distribution

Students arrive in class on the first day of their introductory statistics course. After brief introductions, the instructor asks them to *Meet and Greet* each other. They are asked to stand up, to take a pad of paper with them, and to meet at least five other people by shaking hands and sharing five pieces of data about themselves. This information is to be recorded on their paper:

1. Your name.
2. The number of credits you are taking this semester.
3. Your intended major or field of study.
4. Your gut reaction when you hear the word statistics.
5. If you are a senior (yes/no).

After students have walked around and gathered enough information, they sit back down and engage in a class discussion about their data. They are asked to describe how they recorded their data. Typically, some students write down every response while others may set up a system of categories and tally marks. It is interesting to see what different organization methods were used, which leads to a discussion about what seems to be a good method to organize data and the different types of variables used to collect data (and different types of data collected, e.g., numbers, words, yes/no data; they are asked about how responses differ). Students are asked to look at their data and see which, if any variables, led to a wide variety of responses (variation in the data). There is usually more variation in credits or majors, and less variation in reactions to the word "statistics" (which sadly, are often quite negative). Students are asked about the questions used and if they could have been improved,

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

or if they led to ambiguity of responses. For example, the question: “Are you a senior?” could be answered based on number of credits or year in school, and those might result in different answers.

Finally, the students are asked about what kinds of summaries can be made about the class by looking at the data they collected. This leads to a discussion of their sample of data, and how well their sample represents the entire class. This further leads to the idea that their samples may be too small and possibly biased and that there are better ways to take a sample of data.

Rationale for This Activity

The “Meet and Greet” activity is designed to immerse the students in data and exploratory data analysis from day 1, and to help them think about data collection as well as what you can learn from data. The lesson helps students begin to see data as a classifier, where they collect data to yield frequencies of particular values, and to also begin to see data as an aggregate. These views of data are important and are more advanced ways of thinking about data than simpler, intuitive methods of focusing on individual data values (Konold & Higgins, 2002; Konold, Higgins, Russell, & Khalil, 2003, described later in this chapter). This lesson also helps distinguish statistics from mathematics, by focusing on data and the context of collecting and interpreting data as “noisy” processes. Data that are not numerical are also examined, as three different types of data are informally handled. As we note later in the chapter, there is not much of empirical research on how students develop important ideas about data. This activity and others described in the chapter are often based on implications from studies about the nature of reasoning about data as well as more general research about student learning and effective pedagogical methods for teaching statistics, where specific research studies on learning data are not yet available.

The Importance of Understanding Data Collection and Production

> *Statistics is the science of learning from data. Where the data come from matters.*
>
> (Moore, 2005, p. xviii)

David Moore eloquently reminds us that the most important information about any statistical study is how the data were produced. “Before you trust the results of a statistical study,” he exhorts, “ask about details of how the study was conducted.” Moore goes on to state: “Data enlighten. They shed light in dark places” (Moore, 2005, p. xxiii). He explains that statistics can help guide us in using data to explore the unknown: how to produce trustworthy data, how to look at data (starting with plotting graphs), and how to reach sound conclusions that come with an indication of just how confident we can be. The three important aspects of statistical science are data production, data analysis, and statistical inference (inferring conclusions from data). This chapter deals with the first area, data production.

The Place of Data in the Curriculum

In “traditional” statistics classes, data are introduced on the first day, and distinctions are made between different types of data (categorical and quantitative). In a subsequent unit of graphs, the distinction between these two types of data is revisited as categorical data are tied to pie graphs and bar charts while quantitative data are represented in histograms and dot plots. From that point on, data are used by students as they learn different statistical summaries and procedures. Many students could leave their statistics course never questioning where data come from or realizing that how data are gathered or produced is directly related to methods of analysis and conclusions drawn.

Several years ago, this serious absence began to change as more and more textbooks included chapters on data collection (why and how to take samples and use surveys) and methods of producing data (experimental designs) (e.g., De Veaux, Velleman, & Bock, 2005; Moore, 2004; Peck, Olsen, & Devore, 2007; Rossman, Chance, & Lock, 2001; Watkins, Scheaffer, & Cobb, 2004; Utts, 2004). However, even with this added materials, many times that information was left behind and not referred to later in the course and was treated as an isolated unit. Today’s curriculum recommendations (e.g., the GAISE Project, see Franklin & Garfield, 2006) encourage the study of data production and the integration of this topic throughout a class. Therefore, explorations of data include discussions about the purpose for which the data were gathered and how. While there are many suggestions about how to teach about data, so far there is little empirical research related to this topic. The next section summarizes existing research as well as important work by influential statisticians and uses this literature as a basis for suggesting a progression of ideas and activities to develop reasoning about data.

Review of the Literature Related to Reasoning About Data

In our review of the published literature, we found that the majority of research studies on reasoning about data were conducted at the primary school level, or identified *how* students reason about data rather than *how to develop* good reasoning about data. In addition, we found articles by statisticians and statistics educators expressing strong opinions about the topics related to data. For example, the statistics education literature stresses the importance of using real (and to a lesser extent, realistic) data, the importance of students planning surveys and experiments and collecting their own data, and reflecting on the processes and considerations involved in formulating a statistical question that can be answered with a survey or experimental study. In particular, the idea of where randomization plays a role in statistical study, either through random sampling or random assignment, is now regarded as an extremely important topic for students to learn (Cobb, 2007; Franklin & Garfield, 2006).

We begin this literature review with an overview of what statisticians and statistics educators have written about the nature and importance of exploratory data analysis and how we would like students to think about data, which have provided a basis for the learning goals related to this topic. Then, we summarize the literature that relates to how students understand and reason about data and methods used to collect or produce data.

The Nature of Exploratory Data Analysis

"Statistics is the science of learning from data" (Moore, 2005, p. xviii). One approach to learn from data is Exploratory Data Analysis (EDA), developed by Tukey (1977). EDA is the discipline of organizing, describing, representing, and analyzing data, with a heavy reliance on informal analysis methods, visual displays and, in many cases, technology. The goal of EDA is to make sense of data, analogous to an *explorer of unknown lands* (Cobb & Moore, 1997, p. 807). The original ideas of EDA have since been expanded by Mosteller and Tukey (1977) and Velleman and Hoaglin (1981), and others. They have become the accepted way of approaching the analysis of data (Biehler, 1990; Moore, 1990, 1992). EDA has been widely adopted by statistics educators, in large part, because it serves the need for more data and what we can learn from them, and does not focus on the underlying theory and complicated recipes (Biehler & Steinbring, 1991; Cobb & Moore, 1997; Scheaffer, 2000).

According to Graham (1987), Kader and Perry (1994), Nicholson, Ridgway and McCusker (2006), and others, data analysis is viewed as a four-stage process: (a) specify a problem, plan, pose a question, and formulate a hypothesis; (b) collect and produce data from a variety of sources (survey, experiments); (c) process, analyze, and represent data; and (d) interpret the results, discuss, and communicate conclusions. In reality, however, statisticians do not proceed linearly in this process, but rather iteratively, moving forward and backward, considering and selecting possible paths (Konold & Higgins, 2003). Thus, "data analysis is like a give-and-take conversation between the hunches researchers have about some phenomenon and what the data have to say about those hunches. What researchers find in the data changes their initial understanding, which changes how they look at the data, which changes their understanding" (Konold & Higgins, 2003, p. 194).

The focus of EDA is not on a set of techniques, but on making sense of data, how we dissect a data set, what we look for, how we look, and how we interpret. EDA postpones the classical statistical inference assumptions about what kind of model to fit to the data, with the more direct approach of "letting the data speak for themselves" (Moore, 2004, p. 1); that is, allowing the data to reveal the underlying structure and model through the translating eyes of a statistically literate viewer.

Statistical Thinking About Data

Based on in-depth interviews with six professional practicing statisticians and 16 statistics students, Wild and Pfannkuch (1999) provide a comprehensive description of the processes involved in statistical thinking, from problem formulation to conclusions. They suggest that a statistician operates (sometimes simultaneously) along four dimensions: investigative cycles, types of thinking, interrogative cycles, and dispositions. In solving statistical problems, the statisticians interviewed were particularly interested in giving prominence to grasping the dynamics of the system under investigation, problem formulation, and measurement issues. In the first stages of the investigation, statisticians also attended to data-related issues such as sampling design, data collection, data management, and data cleaning. Some types of the statisticians' thinking that emerged from the interviews are inherently statistical, such as the recognition of need for data, consideration of variation, reasoning with statistical models, and integrating the statistical and contextual.

Although statistical thinking is a synthesis of statistical knowledge, context knowledge, and the information in data to produce implications, insights, and conjectures, these researchers found that the earliest stages of a statistical investigation are driven almost entirely by context knowledge. The statistical knowledge contributes more as the thinking crystallizes. There is a back and forth shuttle between thinking in the context sphere and the statistical sphere that goes on during all phases of the statistical investigation (Wild & Pfannkuch, 1999).

Research on Students Reasoning About Data and Data Analysis

In a "teaching experiment"[1] conducted with lower secondary school students in Germany by Biehler & Steinbring (1991), data analysis was introduced as "detective" work. Teachers gradually provided students with a data "tool kit" consisting of tasks, concepts, and graphical representations. The researchers concluded that all students succeeded in acquiring the beginning tools of EDA, and that both the teaching and the learning became more difficult as the process became more open. There appears to be a tension between directive and nondirective teaching methods in this study. A study by de Lange, Burrill, Romberg, and van Reeuwijk (1993) reveals the crucial need for professional development of teachers in the teaching of EDA in light of the difficulties teachers may find in changing their teaching strategy from expository authority to guiding and from the precision of mathematics to the messiness of statistics.

Based on their observations of school students' reasoning about data, Konold, Higgins, Russell, & Khalil (2003) suggest a framework for describing increasing levels of complexity in how people understand data, focusing on how different representations and uses of these representations highlight or de-emphasize the

[1] See page 37 Chapter 2 for a description of a teaching experiment.

aggregate characteristics of data. Konold and colleagues (Konold & Higgins, 2002; Konold et al., 2003) argue that children see data in *several* simpler ways before ever noticing aggregate and emergent features of data sets. Their fourfold schema includes the following different ways of viewing data, which we consider useful for examining the thinking of adults as well as children:

1. Data as a *pointer* to the data collection event but without a focus on actual data values – in this view, data remind children of their experiences, "We looked at plants. It was fun."
2. Data as a focus on the identity of individual *cases* – these can be personally identifiable, "That's my plant! It's 18 cm tall," extreme values, "The tallest plant was 37 cm," or interesting in some other way.
3. Data as a *classifier*, which focuses on frequencies of particular attribute values, or "slices," without an overall view – "There were more plants that were 15–20 cm than 10–15 cm."
4. Data as an *aggregate*, focusing on overall and emergent characteristics of the data set as a whole, for example, seeing it as describing variability around a center, or "noise" around an underlying "signal" (Konold & Pollatsek, 2002) – "These plants typically grow to between 15 and 20 cm."

More information on the aggregate view of data and distribution is provided in Chapter 8.

Generating and Formulating Statistical Questions

In their review of the research literature on teaching students to generate questions, Rosenshine, Meister, and Chapman (1996) wrote:

> "Question generation is an important comprehension-fostering (Palincsar & Brown, 1984) and self-regulatory cognitive strategy. The act of composing questions focuses the student's attention on content. It involves concentrating on main ideas while checking to see if content is understood (Palincsar & Brown, 1984). Scardamalia and Bereiter (1985) and Garcia and Pearson (1990) suggest that question generation is one component of teaching students to carry out higher level cognitive functions for themselves" (p. 181).

Rosenshine, Meister, and Chapman (1996) found that teaching students to generate questions about the material they have read resulted in gains in comprehension, as measured by tests given at the end of the intervention. Generating and formulating a statistical question, which is the starting point of any statistical investigation, is a challenging task for school and college students. In their chapter, "Reasoning about Data", Konold and Higgins (2003) described this challenge:

> "One of the first challenges is to transform that general question about the real world into a statistical one, one that we can answer with data . . . Among other things, the statistical question allows us to develop measurement instruments and data-collection procedures. By analyzing the data, we answer our statistical question, which ideally, but not always, tells us something about the real question we started with.

> In learning how to formulate questions and to collect and analyze data to answer them, students must learn to walk two fine lines. First, they must figure out how to make a statistical question specific enough so that they can collect relevant data yet make sure that in the process they do not trivialize their question. Second, they must learn to see the data they have created as separate in many ways from the real-world event they observed yet not fall prey to treating data as numbers only. They must maintain a view of data as "numbers in context" (Moore, 1992) while at the same time abstract the data from that context" (p. 195).

Studies on Students' Reasoning About Data Collection and Study Design

The guidelines for teaching the introductory college statistics course (The GAISE Project, Franklin & Garfield, 2006) suggest goals for students in an introductory course that include understanding random sampling and random assignment and the distinction between them. *Random sampling* allows results of surveys and experiments to be extended to the population from which the sample was taken. *Random assignment* in comparative experiments allows cause and effect conclusions to be drawn.

Although there is consensus among statistics educators that student data collection and analysis projects are of substantial value, the planning and piloting phases of data collection are often neglected. Short and Pigeon (1998) asked their college freshman introductory statistics, graduate statistics, and pre-service teachers to write protocols or detailed plans for how the data would be collected for a data investigation project, and to plan and conduct pilot studies before embarking on full scale data collection. They found that the protocol and pilot study assignments developed important global problem-solving and communication skills in students. One of the most important advantages of careful planning of data collections and subsequent analyses that is reported is that students keep the structure of the data they collect within the boundaries of their statistical expertise.

Another essential part of effective statistical study design is deciding when and how to conduct experimental studies rather than nonexperimental ones. This can be challenging even for college students. Heaton and Mickelson (2002) found that undergraduates had some difficulty matching appropriate data collection methods to the quantifiable questions they had posed for class projects. Derry, Levin, Osana, Jones, and Peterson (2000) described the development of undergraduates' statistical thinking ability in regard to study design and documented students' tendency to confuse the concepts of random sampling and random assignment after the course.

Given the difficulties college students have exhibited with deciding when and how to conduct experiments, one would expect experimental design to be a nontrivial matter for high school students. Groth (2003) asked high school students how they would go about designing studies to answer several different quantifiable questions. This study provides a picture of levels of thinking one might expect from high school students in regard to the design of statistical studies.

The role of "data creation" in learning about data analysis was studied by McClain and Cobb (2001). The researchers developed an approach in which the teacher talked through the data creation process with the middle school students. These conversations often involved extended discussions during which the teacher and students together framed the particular phenomenon under investigation, clarified its significance, delineated relevant aspects of the situation that should be measured, and considered how they might be measured. The teacher then introduced the data the students were to analyze as being produced by this process. The researchers found that the data creation process grounded the students' activity in the context of a problem or question under investigation and improved their ways of reasoning about data as they made statistical arguments in the course of their analyses.

Reasoning About Random Samples and Sampling

Confusion about random samples and sampling applies to school and college students as well as adults. In their seminal paper, "Belief in the Law of Small Numbers," psychologists Tversky and Kahneman (1971) wrote:

> The research suggests that people have strong intuitions about random sampling; that these intuitions are wrong in fundamental aspects; that these intuitions are shared by naïve subjects and by trained scientists, and that they are applied with unfortunate consequences in the course of scientific inquiry . . . People view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. Consequently, they expect any two samples drawn from a particular population to be more similar to one another and to the population than sampling theory predicts, at least for small samples. (p. 24)

Since the publication of this article, many researchers have examined and described the difficulties students have understanding samples, sampling variability, and inevitably, sampling distributions and the Central Limit Theorem (CLT). For example, Pollatsek, Konold, Well, and Lima (1984) administered a questionnaire to 205 undergraduate psychology students in the United States. In one experiment, subjects estimated (a) the mean of a random sample of ten scores consisting of nine unknown scores and a known score that was divergent from the population mean; and (b) the mean of the nine unknown scores. The modal answer (about 40% of the responses) for both sample means was the population mean. The results extend the work of Tversky and Kahneman (1971) by demonstrating that subjects hold a passive, descriptive view of random sampling rather than an active balancing model. This result was explored further in in-depth interviews with 31 additional students, where subjects solved the problem while explaining their reasoning. The interview data replicated the first experiment and further showed (a) that subjects' solutions were fairly stable – when presented with alternative solutions including the correct one, few subjects changed their answer; (b) little evidence of a balancing mechanism; and (c) that acceptance of both means as the population mean is largely a result of the perceived unpredictability of "random samples."

In a summary of articles by psychologists on the topic of reasoning about samples, Well et al. (1990) noted that people sometimes reason correctly about sample size (e.g., that larger samples better represent populations) and sometimes do not (e.g., thinking that both large and small samples equally represent a population). To reveal the reasons for this discrepancy, they conducted a series of experiments that gave college students questions involving reasoning about samples and sampling. The researchers found that students used sample size more wisely when asked questions about which sample size is more accurate than on questions that asked them to pick which sample would produce a value in the tail of the population distribution, indicating that they do not understand the variability of sample means.

Understanding the relationship between sample and population requires grasping the representative nature of a random sample. Although studies of elementary students have found that even young children have rich informal knowledge about samples and sampling, they also have numerous difficulties in reasoning about these ideas (Jacobs, 1999; Metz, 1999; Schwartz, Goldman, Vye, & Barron, 1998; Watson and Moritz, 2000a). For example, students were reluctant to generalize from a sample to a population since they seriously doubted that any inference can be drawn beyond the sample at hand, or believed that information on all cases is necessary to draw a conclusion about a population (Metz, 1999). Studies have found that elementary students were often not able to differentiate between results produced by biased and unbiased sampling techniques, and struggled to grasp the idea of randomness and random sampling, preferring convenience sampling over random sampling (Jacobs, 1999; Schwartz, Goldman, Vye, & Barron, 1998). Although students tend to prefer convenience samples and tend to accept stratified samples when thinking about surveys, they seem comfortable with the concept of randomly generating data when considering games of chance (Konold & Higgins, 2003.)

In a rare study at the college level, Dietz (1993) reports on results of a teaching experiment in several introductory statistics courses of undergraduate mathematics education and statistics students. One activity was designed to stimulate students, who had not yet studied sampling, to think creatively about methods of selecting a representative sample from a population. The students generated possible methods for selecting a representative sample, computed various summary statistics and made plots for the variables in each sample, compared their samples statistics to the population parameters and evaluated the advantages and disadvantages of the proposed sampling methods. Dietz reported that the students have "invented" simple random sampling, systematic sampling, stratified sampling, and various combinations thereof. Students, however, had difficulties in evaluating and discussing the various "invented" sampling methods, since they based their evaluation primarily on sample and population measures of central tendency, but ignored the differences in variability. Being actively involved in their own learning and construction of sampling ideas, students better understood and longer remembered ideas related to sampling methods. Chapter 12 in this book further discusses the literature related to reasoning about samples and sampling.

Implications of the Research: Teaching Students to Reason About Data

The literature reviewed suggests that an important component of a statistics course should be on the nature of data, where data come from, how to produce or collect good data (random samples and sampling), and what types of analyses and conclusions are appropriate for data collected in different ways. In order to do this, good, rich data sets are needed. Some of these may be collected by and about the students in the course to engage them in the data collection process. Other data sets may be used as well but always considering the context and the source of data.

In trying to help students develop statistical reasoning about data, it appears helpful to model for them the kinds of questions we need to ask about data in a study, such as:

1. Was this an observational or experimental study? What types of conclusions are therefore appropriate?
2. What methods or precautions were taken to prevent biased data?
3. How and where was randomization used in the study (random sampling? random assignment?).
4. What other precautions were taken (e.g., was the study double-blinded, how were the questions phrased, was there consistency across all measurements? Are the units and measurements clear, e.g., what "operational definitions" were made along the way?).

After a unit on data collection and production is finished, these questions should be revisited throughout the course, so that students are not presented with data sets to analyze without considering where the data come from and what types of analyses are appropriate.

Role of Technology in Learning to Reason About Data

The computer has had a major impact on the use of real data in introductory statistics classes. Many rich data sets are available on the World Wide Web, and most are easily accessible by statistical software packages. For example, *Fathom* (Key Curriculum Press, 2006) and *TinkerPlots* (Konold & Miller, 2005) allow easy access to data files stored on the Internet. Data can be loaded from a Website in a variety of ways, and the software will attempt to interpret the incoming data as cases in a collection. While not 100% foolproof, this feature can greatly reduce the amount of work necessary to get data into a suitable form. Furthermore, *Fathom* can directly import samples of U.S. census microdata (data about individual people) from the Integrated Public Use Microdata Series Web site (IPUMS, http://www.ipums.umn.edu/) at the University of Minnesota, which is a coherent US census database spanning

1850–2004. This is a rich source of interesting data, which *Fathom* makes easy to access and explore.

While some Websites provide data sets that have been cleaned up and formatted to use in teaching statistics (e.g., *DASL*, http://lib.stat.cmu.edu/DASL/), others may be messy and need cleaning before their use. Some instructors are now teaching students to clean and manage data as part of their introductory course (e.g., Gelman & Nolan, 2002; Holcomb & Spalsbury, 2005). The activities allow students to develop their reasoning about what the data represent, what constitutes an outlier or an error, where and how the data were produced, and similar questions. With the abundance of data that are now available in downloadable form, it seems inappropriate to have student spend time entering data by hand into the computer or calculator, other than a few examples to learn and experience the basic process of data entry and storage.

Progression of Ideas: Connecting Research to Teaching

Introduction to the Sequence of Activities to Develop Reasoning About Data

Statistics is the science of data. We therefore begin the study of statistics by introducing the basic ideas about data: Any set of data contains information about some group of individuals, the information is organized in variables, and data represent values of a variable and show the variability of something that is measured. Data are to be viewed as numbers with a context, where the context provides meaning. There are different ways to produce data: by taking measurements, sometimes in the context of an experiment, and by asking questions, such as on a survey. Data vary based on how they are collected. The data collection method matters, because it can affect the quality of the data. Therefore, you need to know the source of the data.

Two methods of gathering data are surveys and experiments. These can be studied in any order, but we present samples and surveys first. Students learn about different types of sampling methods and the kinds of data they produce. They learn about a random sample, its characteristics, and how to take a random sample (e.g., simple random sample, stratified random sample). They learn that a random sample is needed to generalize to a larger group (population) and why this is important. They learn about characteristics of good samples and bad samples, the idea of bias, and what can lead to bias or bad data in a sample survey (e.g., voluntary response samples, poorly worded questions).

When studying about how to produce data in an experiment, students learn the importance and purpose of randomization to infer cause and effect. They learn about the basic principles of statistical design of experiments (control, randomization, and replication), and what makes it bad, i.e., *confounding* of the effect of a treatment with other influences, such as lurking variables, *lack of randomization* which causes bias, or systematic favoritism, in experiments.

After understanding basic ideas of surveys and experiments, students can learn what questions to ask about data collection when looking at data, and how to look at a statistical problem by considering the entire process. This is the beginning of statistical reasoning, which can be developed through subsequent units on data exploration and analysis. Therefore, throughout a course, students need to remember the importance of asking questions about where data come from, why they were gathered, and how that relates to the questions being investigated and the methods of analysis.

Data Sets Used in Lessons in This Book

Two main data sets were collected from students and repeatedly used in the lessons in this book. They are a student survey and a set of body measurements. Both are multivariate data sets and while in most cases, activities have students examine one variable at a time, the students also look at the variables together, getting a sense of the multivariate data and how information on one variable may inform understanding data gathered on other variables.

The survey data can be collected through an online form and contains many questions that have to do with time (see Fig. 6.1). The body measurements data were collected as part of an activity in the unit on Center and Spread (see Chapters 9 and 10). The variables measured for this data set are shown in Fig. 6.2.

Table 6.1 shows a suggested series of ideas and activities that can be used to guide the development of students' reasoning about data. While the accompanying Website (http://www.tc.umn.edu/∼aims) includes some activities (organized in lessons) that illustrate these steps, many are described more generically. The activities that are not included in the Website are marked by the symbol ❖.

Introduction to the Lessons

There are four lessons on collecting and producing data. They begin with types of data and types of variables and the variability of data. The lessons provide students with experience using different methods of sampling to develop an appreciation for random sampling and help students understand different sources of variability and bias in data. Students examine surveys and consider the impact of question wording. They design and conduct an experiment to illustrate principles of random assignment as well as issues involved in designing good experiments. The importance of samples and the role of samples in making inferences are revisited throughout these lessons as a preliminary introduction to informal ideas of statistical inference.

First Day Student Survey Questions

1. Which statistics course are you in?
2. Which section are you in?
3. What is your gender? [Male, Female]
4. What is your age in years?
5. In which month of the year were you born?
6. Which day of the month were you born on?
7. How many statistics courses are you enrolled in this semester?
8. In which year did you start college?
9. Which semester did you start college? [Fall, Winter/Spring, Summer]
10. In which year do you expect to graduate from college?
11. How many credits are you registered for this semester?
12. How many college credits have you completed?
13. What is your cumulative GPA?
14. How many hours per week do you typically study, on the average?
15. How many miles do you travel (one way) from your current home to campus each day, to the nearest mile?
16. How many minutes do you estimate it will take you to travel to school each day this semester, on the average?
17. What type of transportation will you use most often to get to school this semester? [Walk, Car, Bus, Bike, Other]
18. How many minutes do you exercise each week, on the average?
19. Estimate the number of minutes you typically spend each week communicating with your parents (email, phone, in person, etc.).
20. Estimate the number of minutes you spend each day eating (meals and snacks).
21. How many minutes each day do you typically spend on the Internet?
22. How many hours of sleep do you get on a typical week night (Monday through Thursday)?
23. About how many emails do you *send* each day?
24. About how many emails do you *receive* each day?
25. How many minutes do you talk on a cell phone on a typical week day (Monday through Friday)?

Fig. 6.1 First-day student survey questions

Body data collection sheet
(Each student should complete this sheet and enter the measurements into the Instructor's computer.)

1. Head circumference _____
2. Student's head _____
3. Height _____
4. Arm span _____
5. Kneeling height _____
6. Hand length _____
7. Hand span _____

Fig. 6.2 The variables measured for the body measurement data set

Table 6.1 Sequence of activities to develop reasoning about *data*[2]

Milestones: ideas and concepts	Suggested activities
Formal ideas of data	
● Data are values of a variable	● Meet and Greet Activity (Lesson 1: "Data and Variability")
● Measurements produce data	● Meet and Greet Activity (Lesson 1)
● Data show variability	● Meet and Greet Activity (Lesson 1)
● Data are numbers with context	● Variables on Back Activity (Lesson 1)
● There are different kinds of data	● Meet and Greet Activity (Lesson 1)
● Some variability in data is due to measurement process	● Meet and Greet, Variables on Back, and Developing a Class Survey Activities (Lesson 1)
● Importance of taking good measurements by asking clear questions	● Developing a Class Survey Activity (Lesson 1)
● It is important to look at multiple variables (Multivariate data) to better understand and describe a group	● Developing a Class Survey Activity (Lesson 1)
● Sources of bias in questions	● How you Ask a Question Activity (Lesson 2: "Avoiding Bias")
● Importance of asking clear, unambiguous questions in collection survey data	● Critiquing the Student Survey Activity (Lesson 2)
● Idea, purpose and importance of random sampling	● The *Gettysburg Address* Activity (Lesson 3: "Random Sampling")
● Different methods and reasons to take samples	● Student Survey Sampling Activity (Lesson 3)
● Purpose of experiments to produce data to determine cause and effect	● Taste Test Activity (Lesson 4: "Randomized Experiments")
● Purpose of randomization in an experiment	● Taste Test Activity (Lesson 4)
● Idea of making an inference based on a result of an experiment (using simulation)	● Taste Test Activity (Lesson 4)
● Importance of randomization in drawing inferences about results of an experiment	❖ Activity involving random assignment, with introduction to permutation test to informally test if results of the experiment are surprising or due to chance. (The symbol ❖ indicates that this activity is not included in these lessons.)
● Importance of knowing sources of data: data coming from samples or from experiments	❖ Activity where students identify whether the research is a survey (observational data) or an experiment

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

Table 6.1 (continued)

● Good data vs. bad data	❖ Activity where students identify potential sources of bias or confounding
● What type of conclusions can be drawn based on the type of data	❖ Activity identifying the type of conclusion given a study description
● What kinds of questions to ask about where data come from	❖ Activity where students ask appropriate questions for given sets of data
Building on formal ideas of data in subsequent topics	
● Two sources of variation in measurement data	● How Big is Your Head Activity (Lesson 1 in the Variability Unit, Chapter 10)
● Reducing variability in measurement data	● Gummy Bears Activity (Lesson 2 in the Comparing Groups Unit, Chapter 11)
● Determining cause and effect from an experiment	● Gummy Bears Revisited Activity (Lesson 4 in the Statistical Inference Unit, Chapter 13)
● Correlation does not imply causation	● Credit Questions Activity (Lesson 1 in the Covariation Unit, Chapter 14)

Lesson 1: Data and Variability

The goal of the first lesson is to help students see that there are different types of data and different ways to aggregate and display data. This lesson also helps students to see the importance of context and how statistics differs from mathematics in the emphasis of context. Student learning goals for this lesson include:

1. To get started with the statistical process of gathering and interpreting data.
2. To see that there are different types of data and that data vary.
3. To see and consider different sources of variability in data.
4. To develop a survey to use to gather data for future activities (student survey).
5. To see that statistics is different from mathematics and that context of the data is important.

Description of the Lesson

This first class of the course begins with a question about what kinds of students enroll in this class. That leads into the *Meet and Greet Activity*, which is described in detail in the beginning of this chapter. In this activity, students informally discuss many important statistical ideas, such as, methods for data recording, types of variables, types of data, variation in data, question wording, data summaries and representations, sample of data, sample represents the population, sample size, bias, and sampling processes.

The students are then asked to think about other types of information that would be interesting to gather from members of the class (*Developing a Class Survey Activity*). They get into small groups and brainstorm a set of five questions, with the restriction that these questions collect different kinds of data, so that there is at least one question that asks for numerical data, one that asks for categorical (nominal) data and one that asks for yes/no data. Students are encouraged to think of interesting ideas to ask that would produce data they would care about, so not just "year in school" or "gender." The students discuss and create questions that are turned into the instructor, who (before the next class) compiles and edits them into a class survey that they will complete online. This final version of the survey includes additional questions added by the instructor, which will be an important source of data to use in activities on other statistical topics.

A final activity of the first day is to give students an opportunity to reason about data (*Variables on Backs* Activity). Students each have a card taped to their back that has a question on it. Once again, they stand up and walk around the room, this time recording students' responses to the question on their back without knowing what the question is. All the questions are numerical and no units are allowed to be given. Examples of questions are "how many hours did you sleep last night?," "How old is our current president?," "How many counties are in this state?," and "what is the last digit of your ID number?"

After the data are collected, students sit down and look at their data and draw a graph of their choosing (any graph will do). They use this graph to make a guess about what question is taped to their back. They take turns standing up, showing their graph to the class and explaining their reasoning (e.g., "I think my question is how many pets you have because the data I got have lots of 0s, a few ones, and 2s"). After they have explained their reasoning, they can take off the card and see what question was actually taped to their back. The class discussion after this activity involves how students reasoned about their data, and what they thought about and considered as they investigated their question.

A wrap-up discussion includes comparison of statistics to mathematics, how numbers in statistics have a context, and the importance of considering the context of data.

Students are also told that this is the kind of activity they will be doing in class: gathering and analyzing data, and using samples of data to make inferences. It is stressed that *data* are the focus of the course, that we strive to collect good data, and that we are interested in studying the variability in data. They can be asked to summarize why data vary and sources of variability in data.

Lesson 2: Avoiding Bias

The focus of the second lesson is on helping students understand the idea of biased data and ways to avoid biased data in question wording and survey administration. Students suggest and discuss methods of obtaining unbiased data from a survey. Student learning goals for this lesson include:

1. To recognize common instances of bias resulting from how questions are worded or by methods or taking surveys.
2. To learn characteristics of good questions that can be answered by data.
3. To see that how you ask a question makes a difference in the quality of data collected.

Description of the Lesson

Students first see a cartoon that shows a character answering a question on a survey, which illustrates the idea of biased data from a survey and leads to a general discussion about other factors that can bias survey results, leading to bad data. After an initial discussion of how and why we take samples, and the use of surveys, students begin the first activity.

Next, in the *How you Ask a Question* activity, students respond to a set of three questions. What they do not know is that there are two different sets of these questions, worded in different ways. After the questions are answered, a show of hands is asked for the answers to each question, and it is clear that students have responded differently. One student is asked to read aloud their question 1, and then a student reads the other version of question 1. Students are then allowed to see the two sets of questions, and the data are summarized and compared for the two surveys. A discussion ensues on wording effects, how the wording of a question makes a difference in how people respond, the idea of bias in data, and different sources of bias.

This leads into the next activity *Critiquing the Student Survey*, where students read and critique the student survey they helped develop on the first day of class and determine if any of the questions was poorly worded and could lead to biased data. They suggest ways to improve question wording. (Note: they will take the revised survey online outside of class). Finally, students work in groups to discuss what kinds of questions might be posed and answered using the survey data, such as relationship and comparison type questions. They are encouraged to use the new statistical vocabulary they are learning as they talk about surveys, samples, populations, and related terms.

A wrap-up discussion summarizes what students have learned about the meaning and sources of bias in data. They begin to consider ideas of samples in taking surveys as a segue to the next day's topic on types of sampling methods.

Lesson 3: Random Sampling

Lesson 3 focuses on methods of taking samples: why they are important, how to take good samples, how samples differ from each other, and the importance of random sampling. Students take what they think is a representative sample using their judgment, and then compare this to a random sample. They see that nonrandom samples are usually biased. Student learning goals for this lesson include:

1. To understand reasons for using samples in statistical work.
2. To learn to use the basic vocabulary of sampling and surveys.
3. To understand why good samples are important and how we use samples to make inferences.
4. To understand why we rely on chance rather than our own judgment to pick a sample.
5. To learn how to take a Simple Random Sample (SRS) and why SRSs are so important.
6. To recognize and implement several kinds of probability samples (stratified random sample cluster sample, multistage sample, systematic sample).

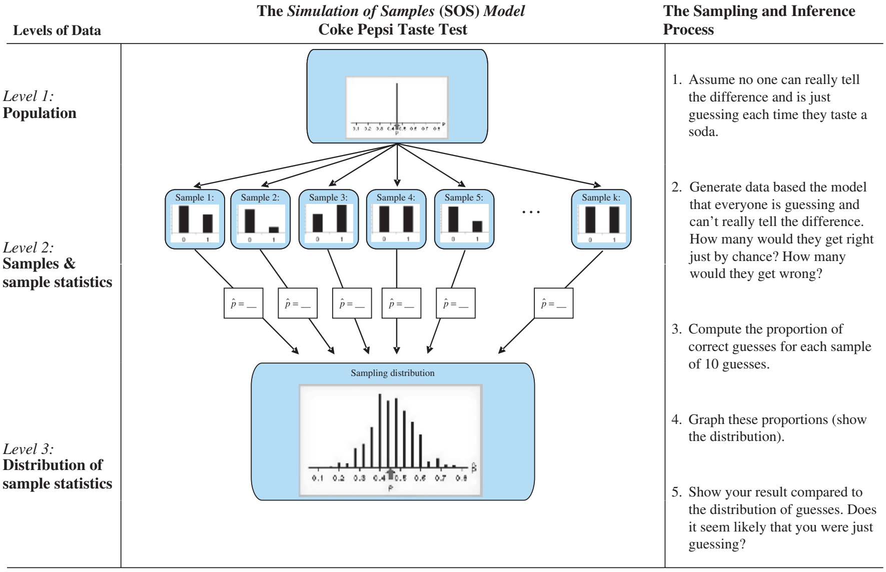
Description of the Lesson

The lesson begins by asking students to suggest good ways to take a representative sample of five students from the class, and they discuss methods of obtaining fair and representative samples for surveys or research purposes. This leads into the *Gettysburg Address* activity. Students are shown the famous *Gettysburg Address* by Abraham Lincoln and told that statistics are often used in analysis of writing style and to identify authors of different writings. Their task is to take a good, representative sample of words from the Gettysburg Address. They do this, and then compute the average word length, being told that average word length is one statistical characteristics of a writer's style. A dot plot of the different sample averages is constructed and examined. The true population average word length is then compared to this plot, and typically, it is not anywhere the center of the graph.

Students then take Simple Random Samples of words from the Gettysburg Address, which is quickly and easily done using *Sampling Words* Java applet (http://www.rossmanchance.com/applets/index.html). They can plot distributions of sample averages and see that these samples are unbiased, and that the true population mean is in or near the center. They repeat this activity with a larger sample size and see the effect. A discussion of bias, representative samples, and the effect of sample size is followed by an activity (*Student Survey Sampling*) where students discuss how they would apply different sampling strategies to taking samples of data from students who have completed the Student Survey.

Lesson 4: Randomized Experiments

This final lesson in the Data unit involves carrying out a randomized experiment, and then considering what is a surprising result. After the randomized experiment is complete and data are gathered, students run a simulation so that they have a distribution of possible results under a chance model to compare with their sample, in order to judge if a particular result is likely to be due to chance, or is too surprising to be attributed to chance. Student learning goals for this lesson include:



1. Assume no one can really tell the difference and is just guessing each time they taste a soda.
2. Generate data based the model that everyone is guessing and can't really tell the difference. How many would they get right just by chance? How many would they get wrong?
3. Compute the proportion of correct guesses for each sample of 10 guesses.
4. Graph these proportions (show the distribution).
5. Show your result compared to the distribution of guesses. Does it seem likely that you were just guessing?

Fig. 6.3 The *Simulation of Samples (SOS) Model* for the taste test

1. To apply the characteristics of a well-defined experiment.
2. To experience the difference between an experiment and an observational study.
3. To learn to recognize instances of confounding in an experiment.
4. To learn the importance of randomization, in randomizing the assignment of treatments, and how that protects against confounding and makes cause and effect statements possible.
5. To develop an informal idea of statistical inference, as the extent to which a result is surprising given a certain claim or theory.
6. To become introduced to and familiar with the *Simulation of Samples (SOS)* Model (Fig. 6.3) as a way to represent data in a simulation.

Description of the Lesson

The lesson begins with the students being asked how many of them believe that they can correctly identify Coke and Pepsi in a blind taste test. They discuss how to determine if somebody really knows how to distinguish the two or if they are just guessing.

After a discussion about the characteristics of a good experiment and what is needed (randomization, control, and replication), students are given the details of the experiment to be performed (*Taste Test* activity). Those students who think they can identify the colas become the tasters. A group of students are assigned to pour the tastes in paper cups, with the order assigned by coin tosses. Another group of students are runners, who bring the tastes to the tasters, not knowing which is which.

The fourth group of students is recorders, who write down what the tasters think each taste is: Coke or Pepsi. The experiment has 10 trials: each taster has 10 blind tests of the soda. After the data are collected, the results are analyzed and each student receives a score for their total of correct identifications.

Next, the class discusses how high a score must be to believe that the student really was not just guessing. The *Sampling SIM* program is run to simulate what we could expect from 10 trails of this experiment if students really were guessing. They can then compare the results of the student guesses to this distribution, to see if the result is in the tails (surprising) and how far in the tails, or in the center (not surprising).

A visual model of the simulation process, adapted from Lane-Getaz (2006), the *Simulation of Samples* (SOS) Model is introduced (see Fig. 6.3). This model is used to help students understand and distinguish between the statistical model used to generate the simulation, the sample data generated, and the distribution of sample statistics for these samples (see Chapter 13 for more detail on this model). The *SOS Model* is also used to substrate the process of comparing students' experimental results to those generated by a particular model or theory (i.e., what if the student was just guessing).

Summary

The four sample lessons in the Data unit help students realize the importance of data and data collection methods in statistics. The class survey that students helped to design was used to collect data that will be analyzed in several subsequent lessons and introduces the importance of a multivariate data set. The ideas of sources of data, data collection methods, measurement issues, and variability of data will be revisited and emphasized again in many of the subsequent lessons. Finally, the importance of randomization and the use of simulation to make an inference about a surprising result are introduced.