Chapter 2
Data

What's It About?

Chapter 2 introduces students to data. We talk about the importance of context (the W's), about variables, and make the distinction between categorical and quantitative data. We begin to introduce the vocabulary of Statistics.

Comments

It is valuable to get students involved with data from the start. We don't take a "big picture" approach at this time. There will be plenty of time to build models and draw inferences later. For now, let's just get our hands dirty playing with the data. When students have a good sense of what kinds of things data can say to us, they learn to expect to listen to the data. Throughout the course, we insist that no analysis of data is complete without telling what it means. This is where that understanding starts.

Rather than head directly for the "real purpose" of the course in the inference chapters, we prefer to emphasize the connection between our work with data and what they tell us about the world. No analysis is complete without a connection back to the real-world circumstances. Setting that stage is the underlying motivation for this chapter. We'll spend the next 10 chapters or so looking at and exploring data without making formal inferences.

Looking Ahead

You might have the students thumb through the book and read the opening of some chapters. Each one starts with a story and data, and most have additional stories and more data inside. Statistics is about the real world. Among other topics, we'll be discussing how well we can predict hurricane tracks, whether generic batteries perform as well as their brand name competitors, the debate about the safety of a widely used drug, and whether or not people over 30 snore more. We need to get students thinking about the context of data and able to make the distinction between categorical and quantitative data. These are fundamental skills for everything that follows, and they'll be used throughout the course.

This is also a good time to introduce whatever technology your students will be using. Have them summarize the results of the survey you did on day 1 using the software package or calculator they will be using for the course. Point them to the computer grid found near the end of most chapters. It offers instructions on several common statistics packages and calculators. The *ActivStats* DVD is available in versions that teach *Data Desk*, *JMP*, *Minitab*, *SPSS*, and *Excel* (with an add-in provided on the disk). You may want to give students a customized handout to get them started in your school's particular computing environment.

Class Do's

Get the class thinking about what the term "data" means. Students need to understand that data are not just numbers and that they must have a context (the W's). When data are quantitative, they should also have units. There are two ways we treat data: ***categorical*** and ***quantitative***. Don't get distracted by worrying about ratio, interval, and other distinctions. These are problematic and don't matter for the concepts and methods discussed in this book.

Emphasize that the distinction between treating data as categorical or quantitative may be more about how *we* display and analyze data than it is about the variable itself. The variable "sex" is data, but just because we might label the males as 1 and the females as 0 doesn't mean that it's quantitative. On the other hand, taking the average of those 0's and 1's does give us the percentage of males. How about *age*? It is often quantitative, but could be categorical if broken down only into *child*, *adult*, and *senior*. Zip code is usually categorical, but if one business had an "average" zip code for their customers of 10000 while another had 90000, we'd know the latter had more customers in the western United States. Emphasize the importance of the context and the W's in summarizing these data.

Students should recognize that every discipline has its own vocabulary, and Statistics is no exception. They'll need to understand and use that vocabulary properly. Unfortunately, many Statistics words have a common everyday usage that's not quite the same. We'll be pointing those out as we go along.

Write vocabulary words on the board as they come up. One of the first should be *variable*. Make the point that it does not mean exactly the same thing as it did in Algebra. There, we call "*x*" a variable, but often that means that we just don't currently know its value. In Statistics a variable is an attribute or characteristic of an individual or object whose value varies from case to case.

A *statistic* is a numerical summary of data; one of the first you'll likely hear is that the class is x% male. Point out the difference between statistics and data. The book wisecracks that, contrary to an advertised saying, you can't be a statistic, only a datum.

Some students will suggest pie charts or histograms. It's sufficient for now to point out that graphical displays are useful visual summaries of data.

Point out that summaries of data can be verbal, visual, and numerical. All are important. In fact, any complete analysis of data almost always includes all three of these.

After looking at the data from your class survey, some students may say things like, "The males are more conservative." Point out the difference between *univariate* and *bivariate* analysis. Note that bivariate is a lot more interesting.

Hope that someone objects to finding an overall average shoe size or to comparing men's and women's sizes—shoe sizes are inconsistent in terms of units. This adds emphasis to the importance of units and the W's.

The Importance of What You Don't Say

We are laying a foundation here. Stretching up to the attic at this point just makes everyone feel unsafe. Many fundamental Statistics terms are left unmentioned in this chapter. You'll find it best to leave it that way. We'll get to them when the students have a safe place to file them along with their other knowledge. So we have an unusually long list of terms we recommend leaving for later in the course. In particular, avoid saying the following:

Hypothesis, Inference. These are certainly important in this course, but we have no background for discussing them honestly now, so they would just be confusing and intimidating.

Nominal, Ordinal, Interval, Ratio. "Nominal" is used by some software packages as a synonym for "categorical" as "continuous" is used for "quantitative." These distinctions arise from studies of measurement scales. But it isn't correct to claim that each variable falls into one of these categories. It is the use to which the data are put that determines what properties the variable must have. Ordinal categorical data may come up, but there are no special techniques for dealing with ordered categories in this course. And any differences between interval- and ratio-scaled data are commonly ignored in statistical analyses. If any of these terms were mentioned now, they'd never come up again anyway.

Random, Probability, Correlation. Everyone has some intuitive sense of these terms, and we'll deal with them formally—but not for a while. Students may want to use these terms, but at this early stage in the course, we don't need them. Without background and careful definition, they are likely to be misused and can simply be frightening.

Class Examples

1. Ask students to tell some things they learned about the class from inspecting the data collected in the opening day's survey. You can use that discussion to develop several of the important points of the chapter.

2. Consider 17, 21, 44, and 76. Are those data? Context is critical—they could be test scores, ages in a golf foursome, or uniform numbers of the starting backfield on the football team. In each case, our reaction changes.

3. Run through some other examples of data, asking about the W's, the variables (what are they, what type is each used as, and what are the units), and so on.

   •A Consumer Reports article on energy bars gave the brand name, flavor, price, number of calories, and grams of protein and fat.

   •A report on the Boston Marathon listed each runner's gender, country, age, and time.

Solution:

Consumer Reports
Who: energy bars
What: brand name, flavor, price, calories, protein, fat
When: not specified
Where:not specified
How: not specified. Are data collected from the label? Are independent tests performed?
Why: information for potential consumers
Categorical variables: brand name, flavor
Quantitative variables: price (US$), number of calories (calories), protein (grams), fat(grams)

Boston Marathon
Who: Boston Marathon runners
What: gender, country, age, time
When: not specified
Where: Boston
How: not specified. Presumably, the data were collected from registration information.
Why: race result reporting
Categorical variables: gender, country
Quantitative variables: age (years), time (hours, minutes, seconds)

Resources

ActivStats

• Start with Lesson 1 to let students familiarize themselves with the features of the software. Lesson 2 examines types of data and context.

Web Links

• The Data and Story Library (DASL) is a source of data for student projects and classroom examples. dasl.datadesk.com

• The U.S. Census Bureau, www.census.gov

Other

• Read polls, studies, or other reports in newspaper and magazine articles. It's always interesting to see how well (or poorly) they provide information about the W's.

If you have a computer and projection capabilities in class, you can find daily surveys at Gallup and other polling organizations. Current data are often particularly interesting to students. But don't use results of voluntary-response online surveys. We'll be making the point that these are fatally flawed—but we can't say that clearly without concepts and terms that we haven't developed yet.

Three chapter quizzes are provided.