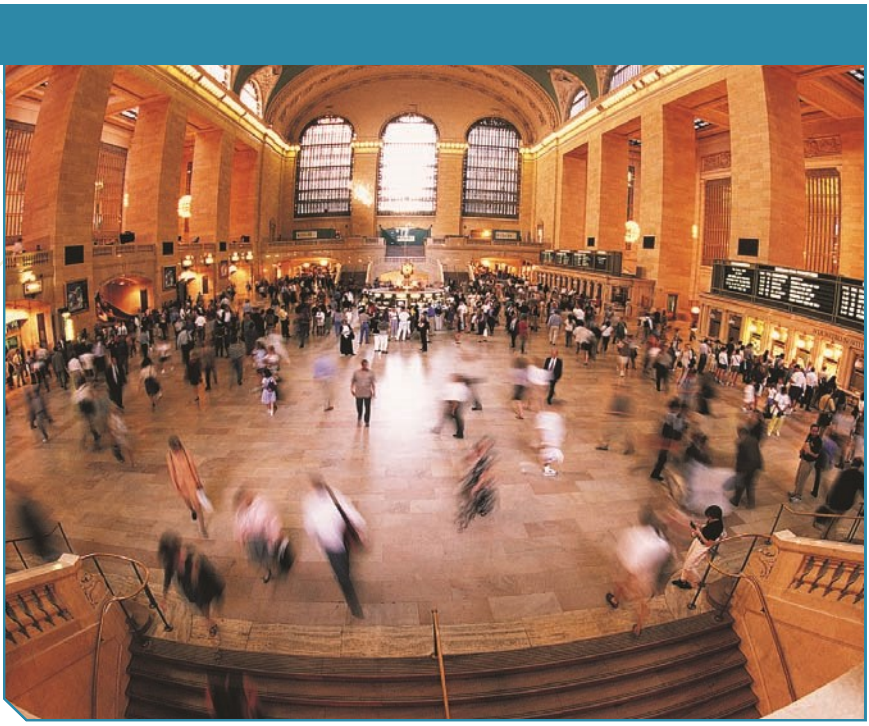
PART

I

Exploring and Understanding Data

CHAPTER 1

Stats Starts Here[1]



"But where shall I begin?" asked Alice. "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."

—Lewis Carroll, *Alice's Adventures in Wonderland*

Statistics gets no respect. People say things like "You can prove anything with Statistics." People will write off a claim based on data as "just a statistical trick." And Statistics courses don't have the reputation of being students' first choice for a fun elective.

But Statistics *is* fun. That's probably not what you heard on the street, but it's true. Statistics is about how to think clearly with data. A little practice thinking statistically is all it takes to start seeing the world more clearly and accurately.

So, What Is (Are?) Statistics?

Q: What is Statistics?
A: Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.
Q: What are statistics?
A: Statistics (plural) are particular calculations made from data.
Q: So what is data?
A: You mean, "what *are* data?" Data is the plural form. The singular is datum.
Q: OK, OK, so what are data?
A: Data are values along with their context.

It seems every time we turn around, someone is collecting data on us, from every purchase we make in the grocery store, to every click of our mouse as we surf the Web. The United Parcel Service (UPS) tracks every package it ships from one place to another around the world and stores these records in a giant database. You can access part of it if you send or receive a UPS package. The database is about 17 terabytes big—about the same size as a database that contained every book in the Library of Congress would be. (But, we suspect, not *quite* as interesting.) What can anyone hope to do with all these data?

Statistics plays a role in making sense of the complex world in which we live today. Statisticians assess the risk of genetically engineered foods or of a new drug being considered by the Food and Drug Administration (FDA). They predict the number of new cases of AIDS by regions of the country or the number of customers likely to respond to a sale at the mall. And statisticians help scientists and social scientists understand how unemployment is related to environmental controls, whether enriched early education af-

[1] This chapter might have been called "Introduction," but nobody reads the introduction, and we wanted you to read this. We feel safe admitting this here, in the footnote, because nobody reads footnotes either.

fects later performance of school children, and whether vitamin C really prevents illness. Whenever there are data and a need for understanding the world, you need Statistics.

So our objectives in this book are to help you develop the insights to think clearly about the questions, use the tools to show what the data are saying, and acquire the skills to tell clearly what it all means.

The ads say, "Don't drink and drive; you don't want to be a statistic." But you can't be a statistic.

We say: "Don't be a datum."

STATISTICS SHOW FAR FEWER SKATEBOARDING INJURIES IN THE WINTER, WHILE SLEDDING INJURIES ARE MORE COMMON.
©2003 Jef Mallett/Distributed by United Feature Syndicate, Inc.
CLEARLY, I'M SAFER GOING DOWN POPSICLE PRECIPICE ON MY SKATEBOARD.
MALLETT
2-24
www.comics.com jefmallett@yahoo.com
FOUR OUT OF FIVE PEOPLE WHO DON'T QUESTION STATISTICS END UP IN JOBS INVOLVING NECKTIES.
REALLY?

FRAZZ reprinted by permission of United Feature Syndicate, Inc.

Statistics in a Word

It can be fun, and sometimes useful, to summarize a discipline in only a few words. So,

Economics is about . . . *Money (and why it is good).*

Psychology: *Why we think what we think (we think).*

Biology: *Life.*

Anthropology: *Who?*

History: *What, where, and when?*

Philosophy: *Why?*

Engineering: *How?*

Accounting: *How much?*

In such a caricature, Statistics is about . . . ***Variation.***

Data vary. People are different. We can't see everything, let alone measure it all. And even what we do measure, we measure imperfectly. So the data we wind up looking at and basing our decisions on provide, at best, an imperfect picture of the world. This fact lies at the heart of what Statistics is all about. How to make sense of it is a central challenge of Statistics.

Statistics is about variation.

Data vary because we don't see everything and because even what we do see and measure, we measure imperfectly.

So, in a very basic way, Statistics is about the real, imperfect world in which we live.

So, How Will This Book Help?

A fair question. Most likely, this book will not turn out to be quite what you expected.

What's different?

Close your eyes and open the book to a page at random. Is there a graph or table on that page? Do that again, say, 10 times. We'll bet you saw data displayed in many ways, even near the back of the book and in the exercises.

We can better understand everything we do with data by making pictures. This book leads you through the entire process of thinking about a problem, finding and showing results, and telling others about what you have discovered. At each of these steps, we display data for better understanding and insight.

You looked at only a few randomly selected pages to get an impression of the entire book. We'll see soon that doing so was sound Statistics practice and reasoning.

Next, pick a chapter and read the first two sentences. (Go ahead; we'll wait.)

We'll bet you didn't see anything about Statistics. Why? Because the best way to understand Statistics is to see it at work. In this book, chapters usually start by presenting a story and posing questions. That's when Statistics really gets down to work.

There are three simple steps to doing Statistics right: *think, show,* and *tell:*

THINK

Think first. Know where you're headed and why. It will save you a lot of work.

SHOW

Show is what most folks think Statistics is about. The *mechanics* of calculating statistics and making displays is important, but not the most important part of Statistics.

TELL

Tell what you've learned. Until you've explained your results so that someone else can understand your conclusions, the job is not done.

FOR EXAMPLE

STEP-BY-STEP

The best way to learn new skills is to take them out for a spin. In **For Example** boxes you'll see brief ways to apply new ideas and methods as you learn them. You'll also find more comprehensive worked examples called **Step-by-Steps.** These show you fully worked solutions side by side with commentary and discussion, modeling the way statisticians attack and solve problems. They illustrate how to think about the problem, what to show, and how to tell what it all means. These step-by-step examples will show you how to produce the kind of solutions instructors hope to see.

Sometimes, in the middle of the chapter, we've put a section called **Just Checking** . . . . There you'll find a few short questions you can answer without much calculation—a quick way to check to see if you've understood the basic ideas in the chapter. You'll find the answers at the end of the chapter's exercises.

MATH BOX

Knowing where the formulas and procedures of Statistics come from and why they work will help you understand the important concepts. We'll provide brief, clear explanations of the mathematics that supports many of the statistical methods in **Math Boxes** like this.

TI Tips

Do statistics on your calculator!

How do I use this thing?

Although we'll show you all the formulas you need to understand the calculations, you will most often use a calculator or computer to perform the mechanics of a statistics problem. Your graphing calculator has a specialized program called a "statistics package." Each chapter contains **TI Tips** that teach you how to use it (and avoid doing most of the messy calculations).

A S **If you have the DVD, you'll find ActivStats** parallels the chapters in this book and includes expanded lessons and activities to increase your understanding of the material covered in the text.

From time to time, you'll see an icon like this in the margin to signal that the *ActivStats* multimedia materials on the available DVD in the back of the book have an activity that you might find helpful at this point. Typically, we've flagged simulations and interactive activities because they're the most fun and will probably help you see how things work best. The chapters in *ActivStats* are the same as those in the text—just look for the named activity in the corresponding chapter.

TI-*nspire*

If you are using TI-Nspire^{TM} technology, these margin icons will alert you to activities and demonstrations that can help you understand important ideas in the text. If you have the DVD that's available with this book, you'll find these there; if not, they're also available on the book's Web site www.aw.com/bock.

"Get your facts first, and then you can distort them as much as you please. (Facts are stubborn, but statistics are more pliable.)"

—Mark Twain

One of the interesting challenges of Statistics is that, unlike in some math and science courses, there can be more than one right answer. This is why two statisticians can testify honestly on opposite sides of a court case. And it's why some people think that you can prove anything with statistics. But that's not true. People make mistakes using statistics, sometimes on purpose in order to mislead others. Most of the unintentional mistakes people make, though, are avoidable. We're not talking about arithmetic. More often, the mistakes come from using a method in the wrong situation or misinterpreting the results. Each chapter has a section called **What Can Go Wrong?** to help you avoid some of the most common mistakes.

Time out. From time to time, we'll take time out to discuss an interesting or important side issue. We indicate these by setting them apart like this.[2]

A S **Introduction to (Your Statistics Package).** *ActivStats* launches your statistics package (such as Data Desk) automatically. If you have the DVD, try it now.

ON THE COMPUTER

There are a number of statistics packages available for computers, and they differ widely in the details of how to use them and in how they present their results. But they all work from the same basic information and find the same results. Rather than adopt one package for this book, we present generic output and point out common features that you should look for. The . . . **on the Computer** section of most chapters (just before the exercises) holds this information. We also give a table of instructions to get you started on any of several commonly used packages, organized by chapters in Appendix B's Guide to Statistical Software.

You'll find all sorts of stuff in margin notes, such as stories and quotations. For example:

"Computers are useless. They can only give you answers."

—Pablo Picasso

While Picasso underestimated the value of good statistics software, he did know that creating a solution requires more than just *Showing* an answer—it means you have to *Think* and *Tell*, too!

At the end of each chapter, you'll see a brief summary of the important concepts you've covered in a section called **What Have We Learned?** That section includes a list of the **Terms** and a summary of the important **Skills** you've acquired in the chapter. You won't be able to learn the material from these summaries, but you can use them to check your knowledge of the important ideas in the chapter. If you have the skills, know the terms, and understand the concepts, you should be well prepared for the exam—and ready to use Statistics!

Beware: No one can learn Statistics just by reading or listening. The only way to learn it is to do it. So, of course, at the end of each chapter (except this one) you'll find **Exercises** designed to help you learn to use the Statistics you've just read about.

T

Some exercises are marked with an orange T. You'll find the data for these exercises on the DVD in the back of the book or on the book's Web site at www.aw.com/bock.

[2] Or in a footnote.

"Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cookbook, believing the recipes will work without understanding why. A more cordon bleu *attitude . . . might lead to fewer statistical soufflés failing to rise."*

—*The Economist,* June 3, 2004, **"Sloppy stats shame science"**

We've paired up the exercises, putting similar ones together. So, if you're having trouble doing an exercise, you will find a similar one either just before or just after it. You'll find answers to the odd-numbered exercises at the back of the book. But these are only "answers" and not complete "solutions." Huh? What's the difference? The answers are sketches of the complete solutions. For most problems, your solution should follow the model of the Step-By-Step Examples. If your calculations match the numerical parts of the "answer" and your argument contains the elements shown in the answer, you're on the right track. Your complete solution should explain the context, show your reasoning and calculations, and state your conclusions. Don't fret too much if your numbers don't match the printed answers to every decimal place. Statistics is more about getting the reasoning correct—pay more attention to how you interpret a result than what the digit in the third decimal place was.

In the real world, problems don't come with chapters attached. So, in addition to the exercises at the ends of chapters, we've also collected a variety of problems at the end of each part of the text to make it more like the real world. This should help you to see whether you can sort out which methods to use when. If you can do that successfully, then you'll know you understand Statistics.

Onward!

It's only fair to warn you: You can't get there by just picking out the highlighted sentences and the summaries. This book is different. It's not about memorizing definitions and learning equations. It's deeper than that. And much more fun. But . . .

You have to read the book![3]

[3] So, turn the page.

CHAPTER 2

Data

"Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the Web site experience."

—Ronny Kohavi,
Director of Data Mining
and Personalization,
Amazon.com

Many years ago, most stores in small towns knew their customers personally. If you walked into the hobby shop, the owner might tell you about a new bridge that had come in for your Lionel train set. The tailor knew your dad's size, and the hairdresser knew how your mom liked her hair. There are still some stores like that around today, but we're increasingly likely to shop at large stores, by phone, or on the Internet. Even so, when you phone an 800 number to buy new running shoes, customer service representatives may call you by your first name or ask about the socks you bought 6 weeks ago. Or the company may send an e-mail in October offering new head warmers for winter running. This company has millions of customers, and you called without identifying yourself. How did the sales rep know who you are, where you live, and what you had bought?

The answer is data. Collecting data on their customers, transactions, and sales lets companies track their inventory and helps them predict what their customers prefer. These data can help them predict what their customers may buy in the future so they know how much of each item to stock. The store can use the data and what it learns from the data to improve customer service, mimicking the kind of personal attention a shopper had 50 years ago.

Amazon.com opened for business in July 1995, billing itself as "Earth's Biggest Bookstore." By 1997, Amazon had a catalog of more than 2.5 million book titles and had sold books to more than 1.5 million customers in 150 countries. In 2006, the company's revenue reached $10.7 billion. Amazon has expanded into selling a wide selection of merchandise, from $400,000 necklaces[1] to yak cheese from Tibet to the largest book in the world.

Amazon is constantly monitoring and evolving its Web site to serve its customers better and maximize sales performance. To decide which changes to make to the site, the company experiments, collecting data and analyzing what works best. When you visit the Amazon Web site, you may encounter a different look or different suggestions and offers. Amazon statisticians want to know whether you'll follow the links offered, purchase the items suggested, or even spend a

[1] Please get credit card approval before purchasing online.

longer time browsing the site. As Ronny Kohavi, director of Data Mining and Personalization, said, “Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them.”

But What *Are* Data?

We bet you thought you knew this instinctively. Think about it for a minute. What exactly *do* we mean by “data”?

Do data have to be numbers? The amount of your last purchase in dollars is numerical data, but some data record names or other labels. The names in Amazon.com’s database are data, but not numerical.

Sometimes, data can have values that look like numerical values but are just numerals serving as labels. This can be confusing. For example, the ASIN (Amazon Standard Item Number) of a book, like 0321570448, may have a numerical value, but it’s really just another name for *Stats: Modeling the World.*

Data values, no matter what kind, are useless without their context. Newspaper journalists know that the lead paragraph of a good story should establish the “Five W’s”: *Who, What, When, Where,* and (if possible) *Why.* Often we add *How* to the list as well. Answering these questions can provide the **context** for data values. The answers to the first two questions are essential. If you can’t answer *Who* and *What,* you don’t have **data,** and you don’t have any useful information.

THE W’S:
WHO
WHAT
and in what units
WHEN
WHERE
WHY
HOW

Data Tables

Here are some data Amazon might collect:

B000001OAA	10.99	Chris G.	902	15783947	15.98	Kansas	Illinois	Boston
Canada	Samuel P.	Orange County	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.	N
Mammals	10783489	Ohio	N	Chicago	12837593	11.99	Massachusetts	16.99
312	Monique D.	10675489	413	B00000I5Y6	440	B000002BK9	Let Go	Y

***Activity:* What Is (Are) Data?** Do you really know what’s data and what’s just numbers?

Try to guess what they represent. Why is that hard? Because these data have no *context.* If we don’t know *Who* they’re about or *What* they measure, these values are meaningless. We can make the meaning clear if we organize the values into a **data table** such as this one:

Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammals

Now we can see that these are four purchase records, relating to CD orders from Amazon. The column titles tell *What* has been recorded. The rows tell us *Who.* But be careful. Look at all the variables to see *Who* the variables are about. Even if people are involved, they may not be the *Who* of the data. For example, the *Who* here are the purchase orders (not the people who made the purchases).

A common place to find the *Who* of the table is the leftmost column. The other W's might have to come from the company's database administrator.[2]

Who

In general, the rows of a data table correspond to individual **cases** about *Whom* (or about which—if they're not people) we record some characteristics. These cases go by different names, depending on the situation. Individuals who answer a survey are referred to as *respondents*. People on whom we experiment are *subjects* or (in an attempt to acknowledge the importance of their role in the experiment) *participants*, but animals, plants, Web sites, and other inanimate subjects are often just called *experimental units*. In a database, rows are called *records*—in this example, purchase records. Perhaps the most generic term is **cases.** In the Amazon table, the cases are the individual CD orders.

A S ***Activity:* Consider the Context . . .** Can you tell who's *Who* and what's *What?* And *Why?* This activity offers real-world examples to help you practice identifying the context.

Sometimes people just refer to data values as *observations,* without being clear about the *Who.* Be sure you know the *Who* of the data, or you may not know what the data say.

Often, the cases are a **sample** of cases selected from some larger **population** that we'd like to understand. Amazon certainly cares about its customers, but also wants to know how to attract all those other Internet users who may never have made a purchase from Amazon's site. To be able to generalize from the sample of cases to the larger population, we'll want the sample to be *representative* of that population—a kind of snapshot image of the larger world.

FOR EXAMPLE Identifying the "Who"

In March 2007, *Consumer Reports* published an evaluation of large-screen, high-definition television sets (HDTVs). The magazine purchased and tested 98 different models from a variety of manufacturers.

Question: Describe the population of interest, the sample, and the *Who* of this study.

The magazine is interested in the performance of all HDTVs currently being offered for sale. It tested a sample of 98 sets, the "Who" for these data. Each HDTV set represents all similar sets offered by that manufacturer.

What and Why

The characteristics recorded about each individual are called **variables.** These are usually shown as the columns of a data table, and they should have a name that identifies *What* has been measured. Variables may seem simple, but to really understand your variables, you must *Think* about what you want to know.

Although area codes are numbers, do we use them that way? Is 610 twice 305? Of course it is, but is that the question? Why would we want to know whether Allentown, PA (area code 610), is twice Key West, FL (305)? Variables play different roles, and you can't tell a variable's role just by looking at it.

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? . . . What kinds of things can we learn about variables like these? A natural start is to *count* how many cases belong in each category. (Are you listening to music while reading this? We could count

[2] In database management, this kind of information is called "metadata."

the number of students in the class who were and the number who weren't.) We'll look for ways to compare and contrast the sizes of such categories.

It is wise to be careful. The *What* and *Why* of area codes are not as simple as they may first seem. When area codes were first introduced, AT&T was still the source of all telephone equipment, and phones had dials.

To reduce wear and tear on the dials, the area codes with the lowest digits (for which the dial would have to spin least) were assigned to the most populous regions—those with the most phone numbers and thus the area codes most likely to be dialed. New York City was assigned 212, Chicago 312, and Los Angeles 213, but rural upstate New York was given 607, Joliet was 815, and San Diego 619. For that reason, at one time the numerical value of an area code could be used to guess something about the population of its region. Now that phones have push-buttons, area codes have finally become just categories.

Some variables have measurement **units.** Units tell how each value has been measured. But, more importantly, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement. The units tell us how much of something we have or how far apart two values are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in euros, dollars, yen, or Estonian krooni.

What kinds of things can we learn about measured variables? We can do a lot more than just counting categories. We can look for patterns and trends. (How much did you pay for your last movie ticket? What is the range of ticket prices available in your town? How has the price of a ticket changed over the past 20 years?)

By international agreement, the International System of Units links together all systems of weights and measures. There are seven base units from which all other physical units are derived:

• Distance	Meter
• Mass	Kilogram
• Time	Second
• Electric current	Ampere
• Temperature	°Kelvin
• Amount of substance	Mole
• Intensity of light	Candela

When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical variable.**[3] When a measured variable with units answers questions about the quantity of what is measured, we call it a **quantitative variable.** These types can help us decide what to do with a variable, but they are really more about what we hope to learn from a variable than about the variable itself. It's the questions we ask a variable (the *Why* of our analysis) that shape how we think about it and how we treat it.

***Activity:* Recognize variables measured in a variety of ways.** This activity shows examples of the many ways to measure data.

***Activities:* Variables.** Several activities show you how to begin working with data in your statistics package.

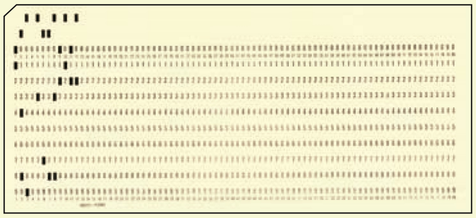
Some variables can answer questions only about categories. If the values of a variable are words rather than numbers, it's a good bet that it is categorical. But some variables can answer both kinds of questions. Amazon could ask for your *Age* in years. That seems quantitative, and would be if the company wanted to know the average age of those customers who visit their site after 3 a.m. But suppose Amazon wants to decide which CD to offer you in a special deal—one by Raffi, Blink-182, Carly Simon, or Mantovani—and needs to be sure to have adequate supplies on hand to meet the demand. Then thinking of your age in one of the categories—child, teen, adult, or senior—might be more useful. If it isn't clear whether a variable is categorical or quantitative, think about *Why* you are looking at it and what you want it to tell you.

A typical course evaluation survey asks, "How valuable do you think this course will be to you?": 1 = Worthless; 2 = Slightly; 3 = Middling; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? Once again, we'll look to the *Why.* A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. When she wants to see whether the course is improving, she might treat the responses as the *amount* of perceived value—in effect, treating the variable as quantitative. But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but we should be careful about treating *Educational Value* as

[3] You may also see it called a *qualitative variable.*

purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational value units" or some similar arbitrary construction. Because there are no natural units, she should be cautious. Variables like this that report order without natural units are often called "ordinal" variables. But saying "that's an ordinal variable" doesn't get you off the hook. You must still look to the *Why* of your study to decide whether to treat it as categorical or quantitative.

One tradition that hangs on in some quarters is to name variables with cryptic abbreviations written in uppercase letters. This can be traced back to the 1960s, when the very first statistics computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and the earliest statistics programs limited variable names to six or eight characters, so variables were called things like PRSRF3. Modern programs do not have such restrictive limits, so there is no reason for variable names that you wouldn't use in an ordinary sentence.



FOR EXAMPLE Identifying "What" and "Why" of HDTVs.

Recap: A *Consumer Reports* article about 98 HDTVs lists each set's manufacturer, cost, screen size, type (LCD, plasma, or rear projection), and overall performance score (0–100).

Question: Are these variables categorical or quantitative? Include units where appropriate, and describe the "Why" of this investigation.

The "what" of this article includes the following variables:

- manufacturer (categorical);
- cost (in dollars, quantitative);
- screen size (in inches, quantitative);
- type (categorical);
- performance score (quantitative).

The magazine hopes to help consumers pick a good HDTV set.

Counts Count

In Statistics, we often count things. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases are shipped. They'd probably start by counting the number of purchases shipped by ground transportation, by second-day air, and by overnight air. Counting is a natural way to summarize the categorical variable *Shipping Method.* So every time we see counts, does that mean the variable is categorical? Actually, no.

We also use counts to measure the amounts of things. How many songs are on your digital music player? How many classes are you taking this semester? To measure these quantities, we'd naturally count. The variables (*Songs*, *Classes*) would be quantitative, and we'd consider the units to be "number of . . ." or, generically, just "counts" for short.

So we use counts in two different ways. When we count the cases in each category of a categorical variable, the category labels are the *What* and the individuals counted are the *Who* of our data. The counts themselves are not the

A S ***Activity:* Collect data in an experiment on yourself.** With the computer, you can experiment on yourself and then save the data. Go on to the subsequent related activities to check your understanding.

data, but are something we summarize about the data. Amazon counts the number of purchases in each category of the categorical variable *Shipping Method.* For this purpose (the *Why*), the *What* is shipping method and the *Who* is purchases.

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

Other times our focus is on the amount of something, which we measure by counting. Amazon might record the number of teenage customers visiting their site each month to track customer growth and forecast CD sales (the *Why*). Now the *What* is *Teens,* the *Who* is *Months,* and the units are *Number of Teenage Customers. Teen* was a category when we looked at the categorical variable *Age.* But now it is a quantitative variable in its own right whose amount is measured by counting the number of customers.

Month	Number of Teenage Customers
January	123,456
February	234,567
March	345,678
April	456,789
May	. . .
. . .	. . .

Identifying Identifiers

What's your student ID number? It is numerical, but is it a quantitative variable? No, it doesn't have units. Is it categorical? Yes, but it is a special kind. Look at how many categories there are and at how many individuals are in each. There are as many categories as individuals and only one individual in each category. While it's easy to count the totals for each category, it's not very interesting. Amazon wants to know who you are when you sign in again and doesn't want to confuse you with some other customer. So it assigns you a unique identifier.

Identifier variables themselves don't tell us anything useful about the categories because we know there is exactly one individual in each. However, they are crucial in this age of large data sets. They make it possible to combine data from different sources, to protect confidentiality, and to provide unique labels. The variables *UPS Tracking Number, Social Security Number,* and Amazon's *ASIN* are all examples of identifier variables.

You'll want to recognize when a variable is playing the role of an identifier so you won't be tempted to analyze it. There's probably a list of unique ID numbers for students in a class (so they'll each get their own grade confidentially), but you might worry about the professor who keeps track of the average of these numbers from class to class. Even though this year's average ID number happens to be higher than last's, it doesn't mean that the students are better.

Where, When, and How

***Self-Test:* Review concepts about data.** Like the Just Checking sections of this textbook, but interactive. (Usually, we won't reference the *ActivStats* self-tests here, but look for one whenever you'd like to check your understanding or review material.)

We must know *Who, What,* and *Why* to analyze data. Without knowing these three, we don't have enough to start. Of course, we'd always like to know more. The more we know about the data, the more we'll understand about the world.

If possible, we'd like to know the **When** and **Where** of data as well. Values recorded in 1803 may mean something different than similar values recorded last year. Values measured in Tanzania may differ in meaning from similar measurements made in Mexico.

How the data are collected can make the difference between insight and nonsense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of Statistics, to be discussed in Part III, is the design of sound methods for collecting data.

Throughout this book, whenever we introduce data, we'll provide a margin note listing the W's (and H) of the data. It's a habit we recommend. The first step of any data analysis is to know why you are examining the data (what you want to know), whom each row of your data table refers to, and what the variables (the columns of the table) record. These are the *Why,* the *Who,* and the *What.* Identifying them is a key part of the *Think* step of any analysis. Make sure you know all three before you proceed to *Show* or *Tell* anything about the data.

JUST CHECKING

In the 2003 Tour de France, Lance Armstrong averaged 40.94 kilometers per hour (km/h) for the entire course, making it the fastest Tour de France in its 100-year history. In 2004, he made history again by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and once again set a new record for the fastest average speed. You can find data on all the Tour de France races on the DVD. Here are the first three and last ten lines of the data set. Keep in mind that the entire data set has nearly 100 entries.

1. List as many of the W's as you can for this data set.

2. Classify each variable as categorical or quantitative; if quantitative, identify the units.

Year	Winner	Country of origin	Total time (h/min/s)	Avg. speed (km/h)	Stages	Total distance ridden (km)	Starting riders	Finishing riders
1903	Maurice Garin	France	94.33.00	25.3	6	2428	60	21
1904	Henri Cornet	France	96.05.00	24.3	6	2388	88	23
1905	Louis Trousselier	France	112.18.09	27.3	11	2975	60	24
⋮								
1999	Lance Armstrong	USA	91.32.16	40.30	20	3687	180	141
2000	Lance Armstrong	USA	92.33.08	39.56	21	3662	180	128
2001	Lance Armstrong	USA	86.17.28	40.02	20	3453	189	144
2002	Lance Armstrong	USA	82.05.12	39.93	20	3278	189	153
2003	Lance Armstrong	USA	83.41.12	40.94	20	3427	189	147
2004	Lance Armstrong	USA	83.36.02	40.53	20	3391	188	147
2005	Lance Armstrong	USA	86.15.02	41.65	21	3608	189	155
2006	Óscar Periero	Spain	89.40.27	40.78	20	3657	176	139
2007	Alberto Contador	Spain	91.00.26	38.97	20	3547	189	141
2008	Carlos Sastre	Spain	87.52.52	40.50	21	3559	199	145

There's a world of data on the Internet. These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. Many of the data sets we use in this book were found in this way. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than those we present. The disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die.

Our solution to these challenges is to offer the best advice we can to help you search for the data, wherever they may be residing. We usually point you to a Web site. We'll sometimes suggest search terms and offer other guidance.

Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. you may have to work in your favorite statistics or spreadsheet program to reformat the data into variables. You will also probably want to remove commas from large numbers and such extra symbols as money indicators ($, ¥, £); few statistics packages can handle these.

WHAT CAN GO WRONG?

- **Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.** The same variable can sometimes take on different roles.
- **Just because your variable's values are numbers, don't assume that it's quantitative.** Categories are often given numerical labels. Don't let that fool you into thinking they have quantitative meaning. Look at the context.
- **Always be skeptical.** One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. The context colors our interpretation of the data, so those who want to influence what you think may slant the context. A survey that seems to be about all students may in fact report just the opinions of those who visited a fan Web site. The question that respondents answered may have been posed in a way that influenced their responses.

TI Tips

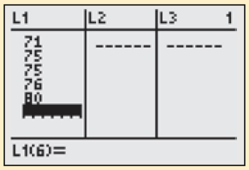
Working with data

You'll need to be able to enter and edit data in your calculator. Here's how.

To enter data:

Hit the STAT button, and choose EDIT from the menu. You'll see a set of columns labeled L1, L2, and so on. Here is where you can enter, change, or delete a set of data.

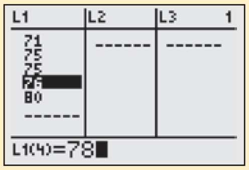
Let's enter the heights (in inches) of the five starting players on a basketball team: 71, 75, 75, 76, and 80. Move the cursor to the space under L1, type in 71, and hit ENTER (or the down arrow). There's the first player. Now enter the data for the rest of the team.



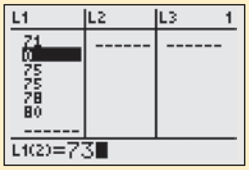
```
L1     |L2     |L3      1
 71    | ------| ------
 75    |       |
 75    |       |
 76    |       |
 80    |       |
 ------|       |
L1(6)=
```

To change a datum:

Suppose the 76" player grew since last season; his height should be listed as 78". Use the arrow keys to move the cursor onto the 76, then change the value and ENTER the correction.



```
L1     |L2     |L3      1
 71    | ------| ------
 75    |       |
 75    |       |
 76    |       |
 80    |       |
 ------|       |
L1(4)=78
```



```
L1      L2      L3     1
 71     ------  ------
 0
 75
 75
 78
 80
 ------
L1(2)=73
```

To add more data:

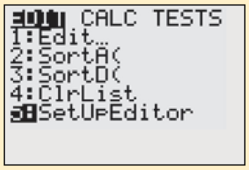
We want to include the sixth man, 73" tall. It would be easy to simply add this new datum to the end of the list. However, sometimes the order of the data matters, so let's place this datum in numerical order. Move the cursor to the desired position (atop the first 75). Hit 2ND INS, then ENTER the 73 in the new space.

To delete a datum:

The 78" player just quit the team. Move the cursor there. Hit DEL. Bye.

To clear the datalist:

Finished playing basketball? Move the cursor atop the L1. Hit CLEAR, then ENTER (or down arrow). You should now have a blank datalist, ready for you to enter your next set of values.



```
EDIT CALC TESTS
1:Edit…
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
```

Lost a datalist?

Oops! Is L1 now missing entirely? Did you delete L1 by mistake, instead of just *clearing* it? Easy problem to fix: buy a new calculator. No? OK, then simply go to the STAT EDIT menu, and run SetUpEditor to recreate all the lists.

WHAT HAVE WE LEARNED?

We've learned that data are information in a context.

- The W's help nail down the context: *Who, What, Why, Where, When,* and *hoW.*
- We must know at least the *Who*, *What*, and *Why* to be able to say anything useful based on the data. The *Who* are the *cases*. The *What* are the *variables*. A variable gives information about each of the cases. The *Why* helps us decide which way to treat the variables.

We treat variables in two basic ways: as *categorical* or *quantitative*.

- Categorical variables identify a category for each case. Usually, we think about the counts of cases that fall into each category. (An exception is an identifier variable that just names each case.)
- Quantitative variables record measurements or amounts of something; they must have *units.*
- Sometimes we treat a variable as categorical or quantitative depending on what we want to learn from it, which means that some variables can't be pigeonholed as one type or the other. That's an early hint that in Statistics we can't always pin things down precisely.

Terms

Context
8. The context ideally tells *Who* was measured, *What* was measured, *How* the data were collected, *Where* the data were collected, and *When* and *Why* the study was performed.

Data
8. Systematically recorded information, whether numbers or labels, together with its context.

Data table
8. An arrangement of data in which each row represents a case and each column represents a variable.

Case
9. A case is an individual about whom or which we have data.

Population
9. All the cases we wish we knew about.

Sample
9. The cases we actually examine in seeking to understand the much larger population.

Variable
9. A variable holds information about the same characteristic for many cases.

Units
10. A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.

Categorical variable
10. A variable that names categories (whether with words or numerals) is called categorical.

Quantitative variable
10. A variable in which the numbers act as numerical values is called quantitative. Quantitative variables always have units.

Skills

THINK

- Be able to identify the *Who, What, When, Where, Why,* and *How* of data, or recognize when some of this information has not been provided.
- Be able to identify the cases and variables in any data set.
- Be able to identify the population from which a sample was chosen.
- Be able to classify a variable as categorical or quantitative, depending on its use.
- For any quantitative variable, be able to identify the units in which the variable has been measured (or note that they have not been provided).

TELL

- Be able to describe a variable in terms of its *Who, What, When, Where, Why,* and *How* (and be prepared to remark when that information is not provided).

DATA ON THE COMPUTER

A S ***Activity:* Examine the Data.** Take a look at your own data from your experiment (p. 12) and get comfortable with your statistics package as you find out about the experiment test results.

Most often we find statistics on a computer using a program, or *package,* designed for that purpose. There are many different statistics packages, but they all do essentially the same things. If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a *tab* character and the delimiter that marks the end of a case to be a *return* character.
- Where to put the data. (Usually this is handled automatically.)
- What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.

EXERCISES

1. Voters. A February 2007 Gallup Poll question asked, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?" The possible responses were "Democrat", "Republican", "Independent", "Other", and "No Response". What kind of variable is the response?

2. Mood. A January 2007 Gallup Poll question asked, "In general, do you think things have gotten better or gotten worse in this country in the last five years?" Possible answers were "Better", "Worse", "No Change", "Don't Know", and "No Response". What kind of variable is the response?

3. Medicine. A pharmaceutical company conducts an experiment in which a subject takes 100 mg of a substance orally. The researchers measure how many minutes it takes for half of the substance to exit the bloodstream. What kind of variable is the company studying?

4. Stress. A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?

(Exercises 5–12) *For each description of data, identify Who and What were investigated and the population of interest.*

Skills

THINK

- Be able to identify the *Who, What, When, Where, Why,* and *How* of data, or recognize when some of this information has not been provided.
- Be able to identify the cases and variables in any data set.
- Be able to identify the population from which a sample was chosen.
- Be able to classify a variable as categorical or quantitative, depending on its use.
- For any quantitative variable, be able to identify the units in which the variable has been measured (or note that they have not been provided).

TELL

- Be able to describe a variable in terms of its *Who, What, When, Where, Why,* and *How* (and be prepared to remark when that information is not provided).

DATA ON THE COMPUTER

A S *Activity:* **Examine the Data.** Take a look at your own data from your experiment (p. 12) and get comfortable with your statistics package as you find out about the experiment test results.

Most often we find statistics on a computer using a program, or *package,* designed for that purpose. There are many different statistics packages, but they all do essentially the same things. If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a *tab* character and the delimiter that marks the end of a case to be a *return* character.
- Where to put the data. (Usually this is handled automatically.)
- What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.

EXERCISES

1. Voters. A February 2007 Gallup Poll question asked, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?" The possible responses were "Democrat", "Republican", "Independent", "Other", and "No Response". What kind of variable is the response?

2. Mood. A January 2007 Gallup Poll question asked, "In general, do you think things have gotten better or gotten worse in this country in the last five years?" Possible answers were "Better", "Worse", "No Change", "Don't Know", and "No Response". What kind of variable is the response?

3. Medicine. A pharmaceutical company conducts an experiment in which a subject takes 100 mg of a substance orally. The researchers measure how many minutes it takes for half of the substance to exit the bloodstream. What kind of variable is the company studying?

4. Stress. A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?

(Exercises 5–12) *For each description of data, identify Who and What were investigated and the population of interest.*

5. **The news.** Find a newspaper or magazine article in which some data are reported. For the data discussed in the article, answer the questions above. Include a copy of the article with your report.

6. **The Internet.** Find an Internet source that reports on a study and describes the data. Print out the description and answer the questions above.

7. **Bicycle safety.** Ian Walker, a psychologist at the University of Bath, wondered whether drivers treat bicycle riders differently when they wear helmets. He rigged his bicycle with an ultrasonic sensor that could measure how close each car was that passed him. He then rode on alternating days with and without a helmet. Out of 2500 cars passing him, he found that when he wore his helmet, motorists passed 3.35 inches closer to him, on average, than when his head was bare. [*NY Times*, Dec. 10, 2006]

8. **Investments.** Some companies offer 401(k) retirement plans to employees, permitting them to shift part of their before-tax salaries into investments such as mutual funds. Employers typically match 50% of the employees' contribution up to about 6% of salary. One company, concerned with what it believed was a low employee participation rate in its 401(k) plan, sampled 30 other companies with similar plans and asked for their 401(k) participation rates.

9. **Honesty.** Coffee stations in offices often just ask users to leave money in a tray to pay for their coffee, but many people cheat. Researchers at Newcastle University alternately taped two posters over the coffee station. During one week, it was a picture of flowers; during the other, it was a pair of staring eyes. They found that the average contribution was significantly higher when the eyes poster was up than when the flowers were there. Apparently, the mere feeling of being watched—even by eyes that were not real—was enough to encourage people to behave more honestly. [*NY Times,* Dec. 10, 2006]

10. **Movies.** Some motion pictures are profitable and others are not. Understandably, the movie industry would like to know what makes a movie successful. Data from 120 first-run movies released in 2005 suggest that longer movies actually make *less* profit.

11. **Fitness.** Are physically fit people less likely to die of cancer? An article in the May 2002 issue of *Medicine and Science in Sports and Exercise* reported results of a study that followed 25,892 men aged 30 to 87 for 10 years. The most physically fit men had a 55% lower risk of death from cancer than the least fit group.

12. **Molten iron.** The Cleveland Casting Plant is a large, highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. The company is interested in keeping the pouring temperature of the molten iron (in degrees Fahrenheit) close to the specified value of 2550 degrees. Cleveland Casting measured the pouring temperature for 10 randomly selected crankshafts.

(Exercises 13–26) *For each description of data, identify the W's, name the variables, specify for each variable whether its use indicates that it should be treated as categorical or quantitative, and, for any quantitative variable, identify the units in which it was measured (or note that they were not provided).*

13. **Weighing bears.** Because of the difficulty of weighing a bear in the woods, researchers caught and measured 54 bears, recording their weight, neck size, length, and sex. They hoped to find a way to estimate weight from the other, more easily determined quantities.

14. **Schools.** The State Education Department requires local school districts to keep these records on all students: age, race or ethnicity, days absent, current grade level, standardized test scores in reading and mathematics, and any disabilities or special educational needs.

15. **Arby's menu.** A listing posted by the Arby's restaurant chain gives, for each of the sandwiches it sells, the type of meat in the sandwich, the number of calories, and the serving size in ounces. The data might be used to assess the nutritional value of the different sandwiches.

16. **Age and party.** The Gallup Poll conducted a representative telephone survey of 1180 American voters during the first quarter of 2007. Among the reported results were the voter's region (Northeast, South, etc.), age, party affiliation, and whether or not the person had voted in the 2006 midterm congressional election.

17. **Babies.** Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births during 1998–2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (cesarean, induced, natural), the level of prenatal care the mother had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).

18. **Flowers.** In a study appearing in the journal *Science*, a research team reports that plants in southern England are flowering earlier in the spring. Records of the first flowering dates for 385 species over a period of 47 years show that flowering has advanced an average of 15 days per decade, an indication of climate warming, according to the authors.

19. **Herbal medicine.** Scientists at a major pharmaceutical firm conducted an experiment to study the effectiveness of an herbal compound to treat the common cold. They exposed each patient to a cold virus, then gave them either the herbal compound or a sugar solution known to have no effect on colds. Several days later they assessed each patient's condition, using a cold severity scale ranging from 0 to 5. They found no evidence of the benefits of the compound.

20. **Vineyards.** Business analysts hoping to provide information helpful to American grape growers compiled these data about vineyards: size (acres), number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.

21. **Streams.** In performing research for an ecology class, students at a college in upstate New York collect data on streams each year. They record a number of biological, chemical, and physical variables, including the stream name, the substrate of the stream (limestone, shale, or mixed), the acidity of the water (pH), the temperature (°C), and the BCI (a numerical measure of biological diversity).

22. **Fuel economy.** The Environmental Protection Agency (EPA) tracks fuel economy of automobiles based on information from the manufacturers (Ford, Toyota, etc.). Among the data the agency collects are the manufacturer, vehicle type (car, SUV, etc.), weight, horsepower, and gas mileage (mpg) for city and highway driving.

23. **Refrigerators.** In 2006, *Consumer Reports* published an article evaluating refrigerators. It listed 41 models, giving the brand, cost, size (cu ft), type (such as top freezer), estimated annual energy cost, an overall rating (good, excellent, etc.), and the repair history for that brand (percentage requiring repairs over the past 5 years).

24. **Walking in circles.** People who get lost in the desert, mountains, or woods often seem to wander in circles rather than walk in straight lines. To see whether people naturally walk in circles in the absence of visual clues, researcher Andrea Axtell tested 32 people on a football field. One at a time, they stood at the center of one goal line, were blindfolded, and then tried to walk to the other goal line. She recorded each individual's sex, height, handedness, the number of yards each was able to walk before going out of bounds, and whether each wandered off course to the left or the right. No one made it all the way to the far end of the field without crossing one of the sidelines. [*STATS* No. 39, Winter 2004]

25. **Horse race 2008.** The Kentucky Derby is a horse race that has been run every year since 1875 at Churchill Downs, Louisville, Kentucky. The race started as a 1.5-mile race, but in 1896, it was shortened to 1.25 miles because experts felt that 3-year-old horses shouldn't run such a long race that early in the season. (It has been run in May every year but one—1901—when it took place on April 29). Here are the data for the first four and several recent races.

Date	Winner	Margin (lengths)	Jockey	Winner's Payoff ($)	Duration (min:sec)	Track Condition
May 17, 1875	Aristides	2	O. Lewis	2850	2:37.75	Fast
May 15, 1876	Vagrant	2	B. Swim	2950	2:38.25	Fast
May 22, 1877	Baden-Baden	2	W. Walker	3300	2:38.00	Fast
May 21, 1878	Day Star	1	J. Carter	4050	2:37.25	Dusty
......						
May 1, 2004	Smarty Jones	2 3/4	S. Elliott	854800	2:04.06	Sloppy
May 7, 2005	Giacomo	1/2	M. Smith	5854800	2:02.75	Fast
May 6, 2006	Barbaro	6 1/2	E. Prado	1453200	2:01.36	Fast
May 5, 2007	Street Sense	2 1/4	C. Borel	1450000	2:02.17	Fast
May 3, 2008	Big Brown	4 3/4	K. Desormeaux	1451800	2:01.82	Fast

26. **Indy 2008.** The 2.5-mile Indianapolis Motor Speedway has been the home to a race on Memorial Day nearly every year since 1911. Even during the first race, there were controversies. Ralph Mulford was given the checkered flag first but took three extra laps just to make sure he'd completed 500 miles. When he finished, another driver, Ray Harroun, was being presented with the winner's trophy, and Mulford's protests were ignored. Harroun averaged 74.6 mph for the 500 miles. In 2008, the winner, Scott Dixon, averaged 143.567 mph.

Here are the data for the first five races and five recent Indianapolis 500 races. Included also are the pole winners (the winners of the trial races, when each driver drives alone to determine the position on race day).

Year	Winner	Pole Position	Average Speed (mph)	Pole Winner	Average Pole Speed (mph)
1911	Ray Harroun	28	74.602	Lewis Strang	.
1912	Joe Dawson	7	78.719	Gil Anderson	.
1913	Jules Goux	7	75.933	Caleb Bragg	.
1914	René Thomas	15	82.474	Jean Chassagne	.
1915	Ralph DePalma	2	89.840	Howard Wilcox	98.580
...					
2004	Buddy Rice	1	138.518	Buddy Rice	220.024
2005	Dan Wheldon	16	157.603	Tony Kanaan	224.308
2006	Sam Hornish Jr.	1	157.085	Sam Hornish Jr.	228.985
2007	Dario Franchitti	3	151.744	Hélio Castroneves	225.817
2008	Scott Dixon	1	143.567	Scott Dixon	221.514

JUST CHECKING
Answers

1. Who—Tour de France races; What—year, winner, country of origin, total time, average speed, stages, total distance ridden, starting riders, finishing riders; How—official statistics at race; Where—France (for the most part); When—1903 to 2008; Why—not specified (To see progress in speeds of cycling racing?)

2.

Variable	Type	Units
Year	Quantitative or Categorical	Years
Winner	Categorical	
Country of Origin	Categorical	
Total Time	Quantitative	Hours/minutes/seconds
Average Speed	Quantitative	Kilometers per hour
Stages	Quantitative	Counts (stages)
Total Distance	Quantitative	Kilometers
Starting Riders	Quantitative	Counts (riders)
Finishing Riders	Quantitative	Counts (riders)