



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."

controls, whether enriched early education affects 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.



FRAZZ reprinted by permission of United Feature Syndicate, Inc.

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.

Statistics in a Word

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.

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. 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.

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 market. And statisticians help scientists and social scientists understand how unemployment is related to environmental

¹This 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.

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 an equation on that page? Do that again, say, 10 times. We'll bet you saw only a few pages with equations.

Equations are a great way of expressing a mathematical idea concisely. But they're not the main point of Statistics. Rather than just listing definitions and equations, this book leads you through the entire process of *thinking* about a problem, finding and *showing* results for the problem, and *telling* others about what you have discovered.

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 first. Know where you're headed and why. It will save you a lot of work. **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 what you've learned. Until you've explained your results so that someone else can understand your conclusions, the job is not done.

Each chapter applies new concepts in a worked example called a Step-by-Step. These examples model 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.

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.

One of the interesting challenges of Statistics is that, unlike 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.

Get your pants first, and then you can distort them as much as you please. (Pants are stubborn, but statistics are more pliable.)
—Mark Twain

What Can Go Wrong?

TI Tips



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).

...and the Computer

Doing It by Hand

We'll show you how to calculate the simpler formulas by hand in a marginal note like this one. That way you can find it easily, but it won't be in the way.

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

"... geographers ... crowd into the edges of their maps parts of the world which they do not know about, adding notes in the margin to the effect that beyond this lies nothing ..."

—Plutarch, (48? C.E.—
120 C.E.), *A Life of Theseus*

Onward!

There are a number of statistics packages for computers, too, 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'll present generic output and point out common features that you should look for in an ... And the Computer section in most chapters. We'll also provide an Appendix showing tables of instructions to get you started on any of five commonly used packages.

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.

We'll also highlight **Key Concepts** as they come up, and collect them at the end of each chapter, together with a summary of important **Skills**. Use these to check your knowledge of the important ideas in the chapter. If you have the skills and understand the key 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.

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.

In the real world, there's no chapter just before the question. So in addition to the problems at the ends of chapters we've also collected more problems at the end of each part 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.

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!²

²Or in a footnote.

³So, turn the page.

Chapter 2

Practice



Many years ago, you probably had to go to a store to buy a pair of shoes. If you walked into the store, you would find a new bridge that had come in from Italy. It was made of leather, and it was your dad's size, and the handless or knobby shoes were available. There were still some stores like that around today, but most people go to the big department stores, by phone or on the Internet. If you go to a store, you might see a number to buy new running shoes, customer service representative. If you call, your first name, or ask about the socks, you might see a number. If you go online, you may send an e-mail in October offering to buy a pair of socks for your company. This company has millions of customers, and you called without identifying yourself. How did it know who you are, where you live, and what you had bought?

The answer to all these questions is data. Collecting data on the most common transactions, and sales enables companies to know where their customers are and what their customers prefer. These data can help them predict what their customers may buy in the future and how much of each item to stock. The store can use the data and what they learn from the data to improve customer service, mimicking the kind of personal attention a shopper had 50 years ago.

Eziba was founded in 1999 as an exclusively Web-based bazaar to forge a link between artisans in developing countries and customers in the West. To meet this challenge, they needed to collect and analyze data to track hundreds of inventory items from around the world, tens of thousands of customers, and marketing initiatives that include direct mail and print advertising. A small company must make smart decisions. How could Eziba compare the success of its print advertisement in the *New York Times* magazine with the effectiveness of its new Web site design?

Eziba designed its Web site to collect data on customer behavior. The company wanted to know how long a visitor to the site spent on each page and how likely she was to make a purchase. The Web site asked customers where they had heard of the company and recorded the number who mentioned the *Times* magazine ads. Analyses of these and other data have enabled Eziba to manage customer relationships and expand sales.

EZIBA



Amber Chand (foreground) co-founder of Eziba, with colleagues from Aid to Artisans, a nonprofit organization

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"?

You might say that data are numbers. The amount of your last purchase in dollars is numerical data, but some data record names or other labels. Your name in Eziba's database is data, but not numerical.

Sometimes, just to make things confusing, data can have values that look like numerical values but are just numerals serving as labels. The item stock numbers Eziba uses to track inventory are really just names.

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 those questions, you don't have data, and you don't have any useful information.

Data Tables

Here are some customer records from another company's database:

7O28YT	24		305	Boston	18	Kansas	5
Veterans		Orange	Y	CKJ245		413	Y
Garbage	43	Wrigley	Y	Chicago	N		Penway
610		130		7TY734	368	JKN234	312

Try to guess what they represent. Why is that hard? Because these data have no context. It's impossible to know what they're about or what they refer to without knowing the W's. We can make the context clear if we organize the values into a data table such as this one.

Table 2.1

Name	Age (yr)	Time Since Last Purchase (days)	Area Code	Nearest Stadium	Internet Purchase?	Catalog Number of Last CD Bought	Artist
Katharine H.		130	312	Wrigley	Y	7TY734	Kansas
Samuel P.	24	18	305	Orange	N	CKJ245	Boston
Chris G.	43	368	610	Veterans	Y	JKN234	Chicago
Monique D.		5	413	Penway	Y	7O28YT	Garbage

Now we can see that these are four customer records from an Internet CD store (that advertises at sporting events). The column titles tell *What* has been recorded. The rows tell us *Who*. The other W's might have to come from the company's database administrator.¹

¹In database management, this kind of information is called "metadata."

Who

In general, each row of a data table corresponds to an individual about whom (or about which—if they're not people) we record some characteristics. These individuals 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, customer records. Perhaps the most generic term is cases. In the table, the cases are the customers: Katharine, Samuel, Chris, and Monique.

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.

What

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 807, 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. Since the advent of push-button phones, area codes have finally become just categories.

Chicago 312, and Los Angeles 213, but rural upstate New York was given 807, 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. Since the advent of push-button phones, area codes have finally become just categories.

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.

Whether a data value is a number or a name may depend on how we use it. Although area codes are numbers, do we use them that

way? Is 610 twice 305? Of course it is, but is Allentown, PA (610), equal to 2 times Key West, FL (305)?

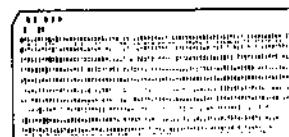
The numbers here are just labels and their values are arbitrary. They represent categories of the variables. We call such variables **categorical**.² A dog's gender, color, and breed are all categorical.

Variables recorded in numbers that we use as numbers are called **quantitative**. Familiar examples include incomes, heights, weights, ages, and counts.

Quantitative variables have measurement units. Units tell how a quantitative value has been measured. Units are such things as yen, cubits, carats, angstroms, nanoseconds, miles per hour, and degrees Celsius. Without units, the values of a quantitative variable have no meaning: It does little good to be promised a salary of 40,000 a year if you don't know whether it will be paid in euros, dollars, yen, or Estonian kroon. Knowing the attribute may not be sufficient. You might be surprised to see someone whose age is 72 listed in a database on childhood diseases until you find out that age is measured in months. No quantitative variable is complete without its units.

²You may also see them called *qualitative*.

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 1890s 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 PRBFF3. Modern programs do not have such restrictive limits, so there's no reason for variable names that you wouldn't use in an ordinary sentence.



Often, just seeking the units can reveal a variable whose definition is dubious. For example, how should we measure "friendliness," "success," "study effort," or "commitment"? It may not be clear how to measure these concepts, yet people make scientific-sounding claims about such "variables." For example, a claim such as "Performance in school is highly correlated with self-esteem" might sound scientific—until we ask in what units one might measure "school performance" or "self-esteem."

Although we've described categorical and quantitative as if they were properties of variables, really they're properties of how we use a variable. For example, if we measure someone's age, we might measure it in years and use it as a quantitative variable. Taking the average age would make sense. But sometimes we treat age as categorical, as, for example, in the categories "child" and "adult." Many variables can be treated either way, depending on the use we want to put them to.

Some variables fall right in between categorical and quantitative. These are variables whose values are not categorical, but not quite quantitative either. For example, think about a survey that asks a question, "What did you think about the pace of the Statistics course you took?" 1 = Way too slow; 2 = A little too slow; 3 = About right; 4 = A little too fast; 5 = Way too fast. Is this variable categorical or quantitative? There is certainly an *order* of perceived speed here. Higher numbers indicate higher perceived speed. A course that averages 4.5 is perceived as going faster than one that averages 2, but we should be careful about treating them as purely quantitative. A course that averages 4.0 is not necessarily *twice* as fast as one that averages 2.0.

These variables are often called **ordinal variables**. Again, depending on what one wants to do with them, they might be treated as numerical (with caution) or categorical, or sometimes even both. For example, the figure numbers in this book can be categorical labels (as in "see Figure 2.1") or they could be quantitative and count the number of figures in a chapter, although there's little reason to want to know that.³ The main point is to think about what the values mean and how they are being used. Be careful of automatically treating values as quantitative just because they look like numbers.

What is measured tells us the meaning of the values. In the Internet CD database, there are seven variables measuring various characteristics about the customers. The variable names often give a lot of information about the *What* if they are well chosen. For this reason, it's a good idea to avoid names like x_1 , TEMP, or even XPECHR8 (as tempting as this may be). Use names that clearly show what the variable is about.

Where, When, How, and Why

We must know *Who* and *What* to analyze data. The more we know, the more we'll understand.

³Figure numbers are identifiers—a special type of categorical variable.

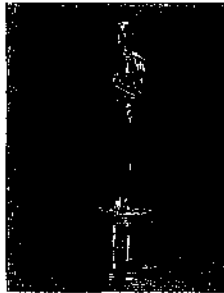
For example, we'd like to know the *When* and *Where* of data as well. Values recorded in 1803 may mean something different than similar values recorded last year. Values measured in Tanzania may differ in meaning from similar measurements made in Mexico.

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, discussed in Part III, is the design of sound methods for collecting data.

Often the most revealing *W* is *Why*. Knowing why data were collected can alert us to potential problems in *How* the data were measured or observed or in the decisions made about the *Who* and *What* of the data. For example, knowing that Levi Strauss conducted the survey may make us skeptical when we hear that 90% of students predicted Levi's 501 jeans to be the most popular clothing item on campus.⁴

Throughout this book, whenever we introduce data, we will provide a marginal note listing the *W*'s (and *H*) of the data. It is a habit we recommend. Whenever you encounter data (or values masquerading as data) ask yourself whether you know the *W*'s. You may be surprised to see how often this vital information is missing. That simple question will protect you from many of the most common misuses of Statistics.

An Example



A study compared the lifetimes of actors who had won Academy Awards (Oscars) with those of actors who had been nominated but had not won and with actors who had been in the movies that led to the awards but had not been nominated. The study found that actors who had won tended to live longer than the other actors considered.

What are the *W*'s in this account? Pause for a moment and try to determine them to the extent that we can tell.

The *Who* are actors, and specifically actors who won Oscars, were nominated for Oscars, or acted in films for which others won Oscars. The *What* of concern in this study is the length of the lives of these actors (which suggests that we may need to modify the *Who* to be only dead actors who meet the other criteria). The *When* isn't clear, although the study may well have spanned the several decades of Academy Awards. *Where* is not really an issue here, although most of the actors nominated are American. The *Why* is, as far as we can tell, a scientific concern with whether fame, or winning, affects lifetimes.

Always refer to variables by name. Your conclusions from statistical analyses should be in clear sentences and about the variables. For example, here is a brief description of a study that was analyzed with Statistics:

Researchers gave 117 people either echinacea or a placebo (sugar pill) for two weeks, then exposed them to cold viruses. Those who took the echinacea were just as likely to develop a cold as those who took a placebo. The researchers concluded that echinacea treatments do not reduce the risk of catching colds.

⁴Learning that these were the only jeans on the list just might increase our skepticism.

It is easy to tell the *Who* (117 people) and *What* (susceptibility to colds) from this description and to understand the conclusions of the study in terms of the variables (*treatment*: a categorical variable with values of either echinacea or placebo; and *cold*: a categorical variable recording whether the people got colds).

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 be slanted.

You'll need to be able to enter and edit data in your calculator. It's easy!

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			

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			
78			
76			
80			

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.

L1	L2	L3	1
71			
73			
75			
75			
76			
80			

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.



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.



Data and the Computer

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.

Key Concepts

Context	The context ideally tells <i>Who</i> was measured, <i>What</i> was measured, <i>How</i> the data were collected, <i>Where</i> the data were collected, and <i>When</i> and <i>Why</i> the study was performed.
Data	Systematically recorded information, whether numbers or labels, together with its context.
Data table	An arrangement of data in which each row represents a case and each column represents a variable.
Case	A case is an individual about whom or which we have data.
Variable	A variable holds information about the same characteristic for many cases.
Categorical variable	A variable that names categories (whether with words or numerals) is called categorical.
Quantitative variable	A variable in which the numbers act as numerical values is called quantitative. Quantitative variables always have units.
Units	A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.

Skills

When you complete this lesson you should:

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 classify a variable as categorical or quantitative.
- 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).

Exercises

For each description of data, identify the *W's*, name the variables, classify each variable as categorical or quantitative, and, for any quantitative variable, identify the units in which it was measured (or note that they were not provided).

1. 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.
2. According to an article in *Fortune* (Dec. 28, 1992), 401(k) plans permit employees 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.
3. Owing to several major ocean oil spills by tank vessels, Congress passed the 1990 Oil Pollution Act, which requires all tankers to be designed with thicker hulls. Further improvements in the structural design of a tank vessel have been proposed since then, each with the objective of reducing the likelihood of an oil spill and decreasing the amount of outflow in the event of a hull puncture. To aid in this development, *Marine Technology* (Jan. 1995) reported on the spillage amount and cause of puncture for 50 recent major oil spills from tankers and carriers.
4. *Ages of Oscar-Winning Best Actors and Actresses* by Richard Brown and Gretchen Davis gives the ages of actors and actresses at the time they won Oscars. We might use these data to see whether actors and actresses are likely to win Oscars at about the same age or not.
5. 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.
6. The Cleveland Casting Plant is a large, highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. According to an article in *Quality Engineering* (7 [1995]), Cleveland Casting 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 a random sample of 10 crankshafts.
7. *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.
8. A study was conducted to compare the strenuous abilities of men and women to perform the strenuous tasks required of a shipboard firefighter (*Human Factors* 24 [1982]). The study reports the pulling force (in newtons) that a firefighter was able to exert in pulling the starter cord of a P-250 water pump. The study also gives the weight and, of course, the gender of the firefighters.
9. 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 gender of the baby, and whether the baby exhibited health problems (none, minor, major).
10. 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 indicate that flowering has advanced an average of 15 days per decade, an indication of climate warming according to the authors.
11. 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 38 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. 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 the student may have.
13. 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 0-5. They found no evidence of the benefits of the compound.
14. *Yee-Kung sales*. A start-up company is building a database of customers and sales information. For each

14 Part I Exploring and Understanding Data

customer it records name, ID number, region of the country (1 = East, 2 = South, 3 = Midwest, 4 = West), date of last purchase, amount of purchase, and item purchased.

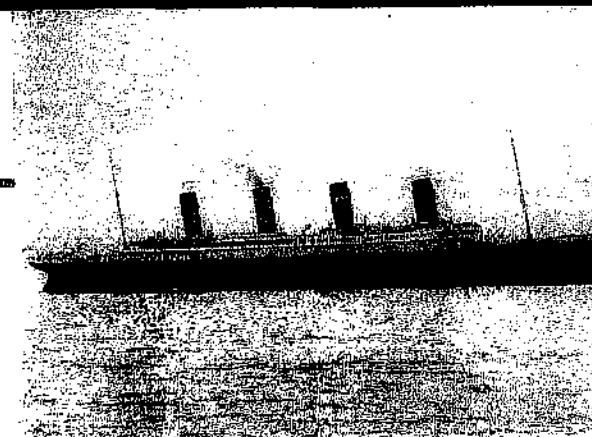
15. A survey of autos parked in student and staff lots at a large university recorded the make, country of origin, type of vehicle (car, van, SUV, etc.), and age.
16. Business analysts hoping to provide information helpful to grape growers compiled these data about vineyards: size (acres), number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.
17. As 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).
18. The Gallup Poll conducted a representative telephone survey of 1180 American voters during the first quarter of 1999. 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 1998 midterm Congressional election.

19. The Federal Aviation Administration (FAA) monitors airlines for safety, and customer service. For each flight the carrier must report the type of aircraft, number of passengers, whether or not the flights departed and arrived on schedule, and any mechanical problems.
20. The Environmental Protection Agency (EPA) tracks fuel economy of automobiles. Among the data they collect are the manufacturer (Ford, Toyota, etc.), vehicle type (car, SUV, etc.), weight, horsepower, and gas mileage (mpg) for city and highway driving.
21. In 2002, *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).

Chapter 3

Displaying and Describing Categorical Data



- WHO** People on the Titanic
- WHAT** Survival status, age, sex, ticket class
- WHEN** April 14, 1912
- WHERE** North Atlantic
- HOW** A variety of sources and internal files
- WHY** Historical interest

What happened on the *Titanic* at 11:30 on the night of April 14, 1912, is well known. Frederick Fleet's cry of "Iceberg, right ahead" and the three accompanying pulls of the crew's nest bell signaled the beginning of a nightmare that has become legend. By 2:15 a.m. the *Titanic*, thought by many to be unsinkable, had sunk, leaving over 1500 passengers and crewmembers on board to meet their icy fate.

Here are some data about the passengers and crew aboard the *Titanic*. Each record of the data table (each row) shows the data for one person on board the ship. The variables are whether or not the person *Survived* (Dead or Alive), the person's *Age* (Adult or Child), *Sex* (Male or Female), and ticket *Class* (First, Second, Third, or Crew).

The problem with a data table like this—and in fact with all data tables—is that you can't see what's going on. And seeing is just what we want to do. We need ways to show the data so that we can see patterns, relationships, trends, and even exceptions.

Survived	Age	Sex	Class
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Alive	Adult	Female	First
Dead	Adult	Male	Third
Dead	Adult	Male	Crew

Part of a data table showing four variables for nine passengers aboard the *Titanic*. Table 3.1

The Three Rules of Data Analysis

So, what should we do with data like these? There are three things you should always do first with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *think* clearly about the patterns and relationships that may be hiding in your data.
2. **Make a picture.** A well-designed display will *show* the important features and patterns in your data. And a picture will show you the things you did not expect to see: the extraordinary (possibly wrong) data values or unexpected patterns.
3. **Make a picture.** The best way to *tell* others about your data is with a well-chosen picture.

These are the three rules of data analysis. There are pictures of data throughout the book, and new kinds keep showing up. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules.

A Picture to Tell a Story

Florence Nightingale, a founder of modern nursing, was also well versed in statistics. To argue forcefully for better hospital conditions for soldiers, she invented this display, which showed that in the Crimean War, far more soldiers died of illness and infection than died of battle wounds. Her campaign succeeded in improving hospital conditions and nursing for soldiers. Figure 3.1



What to Do First: Make Piles

In order to make a picture, the first thing we have to do with data is to make piles. Believe it or not, making piles is the start of all science and the beginning of all understanding about data. We pile together things that seem to go together. Later we'll ask how they go together, why they go together, and how the piles relate and compare with one another. But first, we pile.

Frequency Tables

Class	Count	Class	%
First	325	First	14.766
Second	285	Second	12.949
Third	706	Third	32.076
Crew	885	Crew	40.209

A frequency table of the Titanic passengers.
Table 3.2

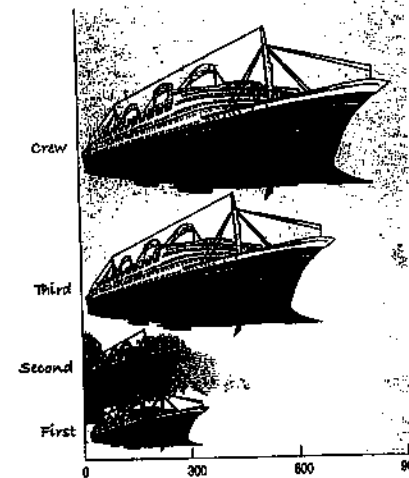
The same data as a relative frequency table.
Table 3.3

One way to put all 2201 people on the *Titanic* into piles is by ticket *Class*, counting up how many had each kind of ticket. We can organize these counts into a **frequency table**, which records the totals and the category names.

The variable ticket *Class* has only a few categories, so a frequency table is easy to read even though we have thousands of cases. A frequency table with dozens or hundreds of categories would be much harder to read. We use the names of the categories to label each row in the frequency table. For ticket *Class*, these are "First," "Second," "Third," and "Crew."

A relative frequency table is similar but gives the *percentages*, rather than the counts, of the values in each category. Both types of tables describe the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs.

The Area Principle



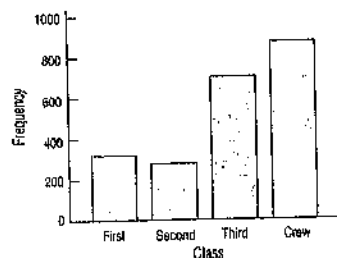
How many people were in each class on the *Titanic*? From this display it looks as though the service must have been great, since most aboard were crew members. Although the length of each ship here corresponds to the correct number, the impression is all wrong. In fact, only 40% were crew. Figure 3.2

Now that we have the frequency table, we're ready to follow the three rules of data analysis and make a picture of the data. But a bad picture can distort our understanding rather than help it. Here's a graph of the *Titanic* data. What impression do you get about who was aboard the ship?

It sure looks like most of the people on the *Titanic* were crew members, with a few passengers along for the ride. That doesn't seem right. What's wrong? The lengths of the ships do match the totals in the table. (You can check the scale at the bottom.) However, experience and psychological tests show that our eyes tend to be more impressed by the *area* than by other aspects of each ship image. So, even though the length of each ship matches up with one of the totals, it's the associated *area* in the image that we notice. Since there were about 3 times as many crew as second-class passengers, the ship depicting the number of crew is about 3 times longer than the ship depicting second-class passengers, but it occupies about 9 times the area. As you can see from the frequency table (Table 3.2), that just isn't a correct impression.

The best data displays observe a fundamental principle of graphing data called the **area principle**. The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. Violations of the area principle are a common way to lie (or, since most mistakes are unintentional, we should say *err*) with Statistics.

Bar Charts

People on the *Titanic* by Ticket Class

With the area principle satisfied, we can see the true distribution more clearly. Figure 3.3

For some reason, some computer programs give the name "bar chart" to any graph that uses bars. And others use different names according to whether the bars are horizontal or vertical. Don't be misled. "Bar chart" is the term for a display of counts of a categorical variable with bars.

Here's a chart that obeys the area principle. It's not as visually entertaining as the ships, but it does give an accurate visual impression of the distribution. The height of each bar shows the count for its category. The bars are the same width, so their heights determine their areas, and the areas are proportional to the counts in each class. Now it's easy to see that the majority of people on board were *not* crew, as the ships picture led us to believe. We can also see that there were about 3 times as many crew as second-class passengers. And there were more than twice as many third-class passengers as either first- or second-class passengers, something you may have missed in the frequency table. Bar charts make these kinds of comparisons easy and natural.

A bar chart displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison. Bar charts have small spaces between the bars to indicate that these are free-standing bars that could be rearranged into any order. The bars are lined up along a common base.

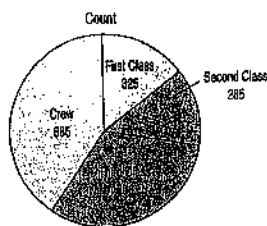
Usually they stick up like this



times they run sideways like this



Pie Charts



The number of *Titanic* passengers in each class. Figure 3.4

If we want to draw attention to the relative *proportion* of passengers falling into each of these classes, we could replace the counts with percentages in a bar chart. A better choice might be a pie chart. Pie charts show the whole group of cases as a circle. They slice the circle into pieces whose size is proportional to the fraction of the whole in each category.

Here again it's easy to see that the crew was the largest segment of those aboard. Pie charts give a quick impression of how a whole group is partitioned into smaller groups. Because we're used to cutting up pies into 2, 4, or 8 pieces, pie charts are good for seeing if the relative frequency of a category is near $1/2$, $1/4$, or $1/8$. For example, you may be able to tell that the pink slice, representing the second-class passengers, is very close to $1/8$ of the total. It's harder to see that there were about twice as many third-class as first-class passengers. Were there more crew or more third-class passengers? Comparisons such as these are easier in a bar chart.

Children and First-Class Ticket Holders First?

We know how many tickets of each class were sold on the *Titanic*, and we know that only 32% of all those aboard the *Titanic* survived. Was there a relationship between the kind of ticket a passenger held and the passenger's chances of making it into the lifeboat? To answer this question, we need to look at the two categorical variables *Class* and *Survival* together.

Contingency Tables

When we look at two categorical variables together, we often arrange the counts in a two-way table. Because the table shows how the individuals are distributed along each variable, contingent on the value of the other variable, such a table is called a contingency table.

Here is a contingency table of those aboard the *Titanic* classified according to class of ticket and whether they survived or didn't.

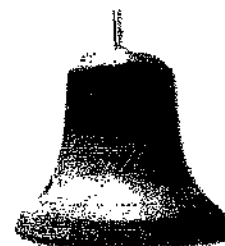
		Class				Total
		First	Second	Third	Crew	
Survival	Alive	202	118	178	212	710
	Dead	123	167	528	673	1491
	Total	325	285	706	885	2201

Contingency table of ticket *Class* and *Survival*. The bottom line of "Totals" is the same as the previous frequency table. Table 3.4

The margins of the table, both on the right and at the bottom, give totals. The bottom line of the table is just the frequency distribution of ticket *Class*. A distribution, like this, of one of the variables in a contingency table is called its marginal distribution.

The cells of the table give the counts or frequencies for every combination of values for the two variables. If you look down the column for second-class passengers to the first cell, you can see that 118 second-class passengers survived. Looking at the third-class passengers, we see that more third-class passengers (178) survived. Those 118 surviving second-class passengers were nearly half of the 285 total in second class. But the 178 third-class survivors were a much smaller fraction of the total of 706 third-class passengers.

It might be more useful to have percentages, but to do that we have to make choices. We know that 118 second-class passengers survived. We could display this number as a percentage, but as a percentage of what? The total number of passengers (118 is 5.4% of the total, 2201)? The number of second-class passengers (118 is 41.4% of the 285 second-class passengers)? The number of survivors (118 is 16.6% of the 710 survivors)? All of these are possibilities, and all are potentially useful or interesting. You'll probably wind up calculating (or letting your technology



A bell-shaped artifact from the *Titanic*.

calculate) lots of percentages. Most statistics programs offer a choice of total percent, row percent, or column percent for contingency tables. Here are the counts and all three percentages displayed as they might be by a computer package:

		Class					Total
		First	Second	Third	Crew	Total	
Survival	Alive	Count	202	118	178	212	710
	% of Row		28.5%	16.6%	25.1%	29.9%	100%
	% of Column		62.2%	41.4%	25.2%	24.0%	32.3%
	% of Table		9.18%	5.36%	8.09%	9.63%	32.3%
Survival	Dead	Count	123	167	528	673	1491
	% of Row		8.25%	11.2%	35.4%	45.1%	100%
	% of Column		37.8%	58.6%	74.8%	76.0%	67.7%
	% of Table		5.59%	7.59%	24.0%	30.6%	67.7%
Survival	Total	Count	325	285	706	885	2201
	% of Row		14.8%	12.9%	32.1%	40.2%	100%
	% of Column		100%	100%	100%	100%	100%
	% of Table		14.8%	12.9%	32.1%	40.2%	100%

Another contingency table of ticket Class. This time we see not only the counts for each combination of Class and Survival (in bold) but the percentages these counts represent. For each count, there are three choices for the percentage: by row, by column, and by table total. There's probably too much information here for this table to be useful. Table 3.5

Each cell of this table gives the count, row percent, column percent, and table percent in that order. This is an example of why contingency tables can look so confusing. There's too much information to sort through at one glance. While it's fine to consider all these choices, it's probably better to look at them one at a time. In this table each row shows what percentage of passengers were in each class.

		Class					Total
		First	Second	Third	Crew	Total	
Survival	Alive	Count	202	118	178	212	710
	% of Row		28.5%	16.6%	25.1%	29.9%	100%
Survival	Dead	Count	123	167	528	673	1491
	% of Row		8.25%	11.2%	35.4%	45.1%	100%
Survival	Total	Count	325	285	706	885	2201

A contingency table of Class by Survival with only counts and row percentages. Notice how much easier this table is to read than the previous one. Of course, two other similar tables could be made for column percentages and table percentages. Table 3.6

Marginal and Conditional Distributions

Did the chance of surviving the *Titanic* sinking depend on ticket class? Does the distribution of survivors' ticket class look the same as that distribution for non-survivors? To answer the question, let's look at the table.

First we restrict our attention only to the survivors. This is like redefining the *Who* of the study. The *Who* we're interested in right now is only the survivors. Their numbers are in the first row of the contingency table. A distribution of one variable for only those individuals satisfying some condition on another variable is called a **conditional distribution**.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	202	118	178	212	710
		28.5%	16.6%	25.1%	29.9%	100%

The conditional distribution of ticket Class, conditional on having survived. Table 3.7

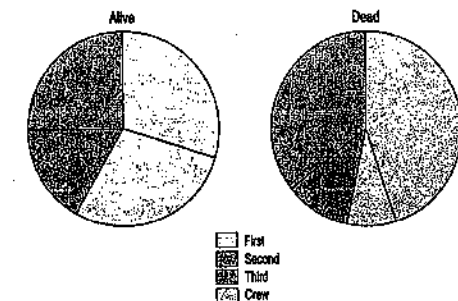
Now we do the same thing for the nonsurvivors. The numbers for the nonsurvivors are found in the following row:

		Class				
		First	Second	Third	Crew	Total
Survival	Dead	123	167	528	673	1491
		8.25%	11.2%	35.4%	45.1%	100%

The conditional distribution of ticket Class, conditional on having perished. Table 3.8

Let's compare these conditional distributions. Among survivors 28.5% held a first class ticket compared to only 8.25% among those who died. That looks like a big difference.

Pie charts of the distribution of Class for the survivors and nonsurvivors separately. Do the distributions appear to be the same? We're primarily concerned with percentages here, so pie charts are a good choice. Figure 3.5



The nonsurvivors are mostly crew and third-class passengers. The survivors, on the other hand, are more uniformly split up across all four classes. If the percentages of ticket class had been about the same across the two survival groups, we would have said that survival was independent of class. But it's not. The differences we see between the two conditional distributions suggest that survival may have depended on ticket class.

It is interesting to know that *Class* and *Survival* are associated. That's an important part of the *Titanic* story. And we know they're associated because we can see that the distribution of ticket classes differs between survivors and those who died.

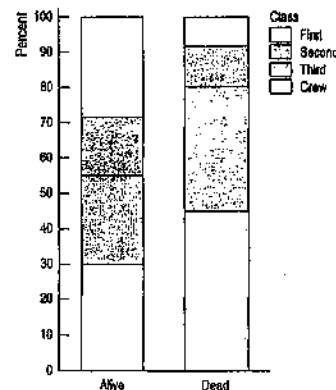
Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are not.¹ In a contingency table, when the distribution of *one* variable is the same for all categories of another, we say that the variables are independent. We'll see a way to check for independence formally later in the book. For now, we'll just compare the distributions.

Segmented Bar Charts

We could display the same information by dividing up bars rather than circles. The resulting segmented bar chart treats each bar as the "whole" and divides it proportionally into segments corresponding to the percentage in each group. We can clearly see that the distributions of ticket classes are different, indicating that survival was not independent of ticket class.

A segmented bar chart for *Class* by *Survival*. Notice that although the totals for survivors and nonsurvivors are quite different, the bars are the same height because we have converted the numbers to percentages. Compare this with the side-by-side pie charts of the same data.

Figure 3.6



¹This kind of "backwards" reasoning shows up surprisingly often in science—and in Statistics. We'll see it again.

Examining Contingency Tables STEP-BY-STEP

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer ("Fatty Fish Consumption and Risk of Prostate Cancer," *Lancet*, June 2001). Their results are summarized in this table.



We asked for a picture of a "man eating fish." This is what we got.

	Prostate Cancer	
	No	Yes
Never/seldom	110	14
Small part of diet	2420	201
Moderate part	2769	209
Large part	507	42

Table 3.9

Is there an association between fish consumption and prostate cancer?

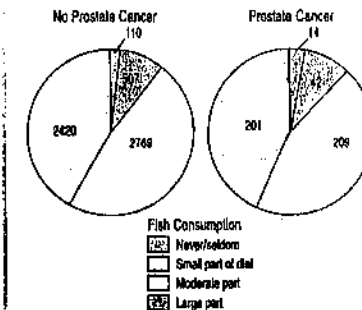
Think

Clarify: Identify the variables and report the W's. Be certain that the data are counts and that the categories do not overlap so that no individual is counted twice.

The individuals are 6272 Swedish men followed by medical researchers for 30 years. The variables record their fish consumption and whether or not they were diagnosed with prostate cancer. The data are reported as counts. The categories of diet do not overlap and the diagnoses do not overlap.

Show

Visualize: Make an appropriate display to see whether there is a difference in the relative proportions. Bar charts might have worked equally well.



Tell

Interpretation: Discuss the patterns in the table and displays.

There appears to be little difference between the two groups in terms of their fish consumption. Fish consumption appears to be independent of the incidence of prostate cancer.

If you can, discuss possible real-world consequences.

We see no reason for men to change their diets in an attempt to reduce their risk of prostate cancer based on this study.

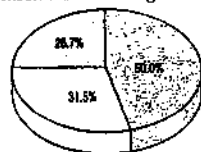
What Can Go Wrong?

- Don't violate the area principle. This is probably the most common mistake in a graphical display. It is often made in the cause of artistic presentation. Here, for example, is a display of the pie chart of the *Titanic* passengers by class:



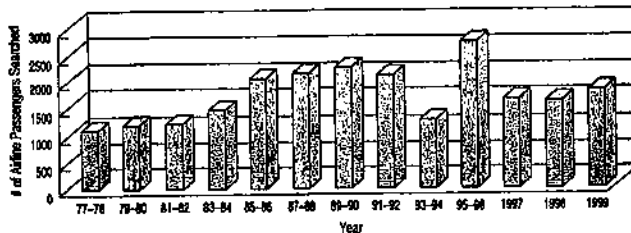
Looks pretty, doesn't it? But showing the pie on a slant violates the area principle and makes it much more difficult to compare fractions of the whole made up of each class—the principal feature that a pie chart ought to show.

Keep it honest. Here's a pie chart that displays data on the percentage of high school students who engage in specified dangerous behaviors as reported by the Centers for Disease Control. What's wrong with this plot?



Try adding up the percentages. Or look at the 50% slice. Does it look right? Then think: What are these percentages of? Is there a "whole" that has been sliced up? In a pie chart, the proportions shown by each slice of the pie must add up to 100% and each individual must fall into only one category. Of course, showing the pie on a slant makes it even harder to detect the error.

Here's another. This bar chart shows the number of airline passengers searched by security screening.



Looks like things didn't change much in the final years of the 20th century—until you read the bar labels and see that the last three bars represent single years, while all the others are for pairs of years. Of course, the false depth makes it harder to see the problem.

- Don't confuse similar-sounding percentages. These percentages sound similar but are different:
 - The percentage of those who were both in first class and survived: This would be 202/2100, or 9.18%.

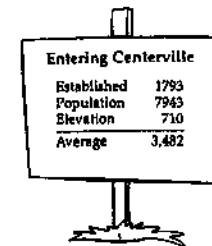
Survival	Class				Total
	First	Second	Third	Crew	
Alive	202	118	178	212	710
Dead	123	167	528	673	1491
Total	325	285	706	885	2201

- The percentage of those who survived among those who were in first class: This is 202/325, or 62.5%.
 - The percentage of those who were in first class among those who survived: This is 202/710, or 28.5%.
- In each instance, pay attention to the *Who* implicitly defined by the phrase. Often there is a restriction to a smaller group (all aboard the *Titanic*, those in first class, and those who survived, respectively) before a percentage is found. Your discussion of results must make these differences clear.

Be sure to use enough individuals. When you consider percentages, take care that they are based on a large enough number of individuals. Take care not to make a report such as this one:

We found that 66.67% of the rats improved their performance with training. The other rat died.

- Don't overstate your case. Independence is an important concept, but it is rare for two variables to be entirely independent. We don't know, for example, that fish consumption has no effect whatever on prostate cancer. All we know is that no effect was observed in that study. Other studies of other groups under other circumstances could find different results.



Simpson's Paradox

- Don't use unfair or silly averages. Sometimes averages can be misleading. Sometimes they just don't make sense at all. Be careful when averaging different variables that the quantities you're averaging are comparable. The Centerville sign says it all.

When using averages of proportions across several different groups, it's important to make sure that the groups really are comparable.

It's easy to make up an example showing that averaging across very different values or groups can give absurd results. Here's how that might work. Suppose there are two pilots, Moe and Jill. Moe argues that he's the better pilot of the two, since he managed to land 83% of his last 120 flights on time compared with Jill's 78%. But let's look at the data a little more closely. Here are the results for each of their last 120 flights, broken down by the time of day they flew:

Pilot	Time of Day		
	Day	Night	Overall
Moe	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
Jill	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

On-time flights by Time of Day and Pilot. Look at the percentages within each Time of Day category. Who has a better on-time record during the day? At night? Who is better overall? Table 3.10

One famous example of Simpson's paradox arose during an investigation of admission rates for men and women at the University of California at Berkeley's graduate schools. As reported in an article in *Science*, about 45% of male applicants were admitted, but only about 30% of female applicants got in. It looked like a clear case of discrimination. However, when the data were broken down by school (Engineering, Law, Medicine, etc.) it turned out that within each school, the women were admitted at nearly the same or, in some cases, much higher rates than the men. How could this be? Women applied in large numbers to schools with very low admission rates (Law and Medicine, for example, admitted fewer than 10%). Men tended to apply to Engineering and Science. Those schools have admission rates above 50%. When the average was taken, the women had a much lower overall rate, but the average didn't really make sense.

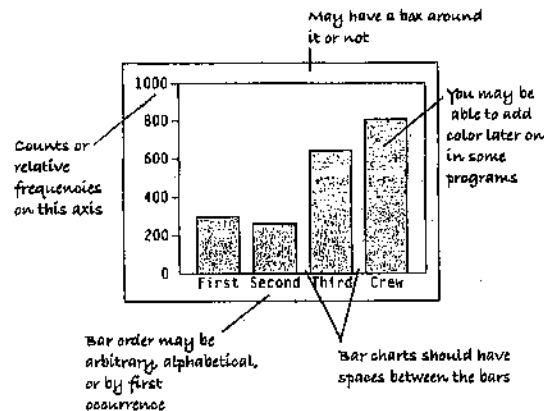
Look at the day and nighttime flights separately. For day flights, Jill had a 95% on-time rate, and Moe only a 90% rate. At night, Jill was on time 75% of the time, and Moe only 50%. So Moe is better "overall," but Jill is better both during the day and at night. How can this be?

What's going on here is a problem known as Simpson's paradox, named for the statistician who discovered it in the 1960s. It comes up rarely in real life, but there have been several well-publicized cases of it. As we can see from the pilot example, the problem is *unfair averaging* over different groups. Jill has mostly night flights, which are more difficult, so her *overall average* is heavily influenced by her nighttime average. Moe, on the other hand, benefits from flying mostly during the day, with its higher on-time percentage. With their very different patterns of flying conditions, taking an overall average is misleading. It's not a fair comparison.

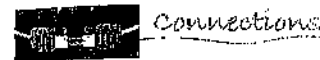
The moral of Simpson's paradox is to be careful when you average across different levels of a second variable. It's always better to compare percentages or other averages *within* each level of the other variable. The overall average may be misleading.

Displaying Categorical Data with a Computer

Although every package makes a slightly different bar chart, they all have similar features:



Sometimes the count or a percentage is printed above or on top of each bar to give some additional information. You may find that your statistics package sorts category names in annoying orders by default. For example, many packages sort categories alphabetically or by the order the categories are seen in the data set. Often, neither of these is the best choice.



Connections

All of the methods of this chapter work with *categorical variables*. You must know the *Who* of the data to know who is counted in each bar and the *What* of the variable to know where the categories come from.

Key Concepts

Frequency table	A frequency table lists the categories in a categorical variable and gives the counts or percentage of observations of each category.
Distribution	The distribution of a variable gives <ul style="list-style-type: none"> • the possible values of the variable and • the relative frequency of each value.
Area principle	In a statistical display, each data value should be represented by the same amount of area.
Bar chart	Bar charts show a bar representing the count of each category in a categorical variable.
Pie chart	Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.
Contingency table	A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once, to reveal possible patterns in one variable that may be contingent on the category of the other.
Marginal distribution	In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table.
Conditional distribution	The distribution of a variable restricting the <i>Who</i> to consider only a smaller group of individuals is called a conditional distribution.
Independence	Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other. We'll show how to check for independence in a later chapter.
Simpson's paradox	When averages are taken across different groups, they can appear to be contradictory. This is known as Simpson's paradox.

Skills

When you complete this lesson you should:

Think

- Be able to identify that a variable is categorical and choose an appropriate display for it.
- Understand how to examine the association between categorical variables by comparing conditional and marginal percentages.

Show

- Be able to summarize the distribution of a categorical variable with a frequency table.
- Be able to display the distribution of a categorical variable with a bar chart or pie chart.
- Know how to make and examine a contingency table.
- Know how to make and examine displays of the conditional distributions of one variable for two or more groups.

Tell

- Be able to describe the distribution of a categorical variable in terms of its possible values and relative frequencies.
- Know how to describe any anomalies or extraordinary features revealed by the display of a variable.
- Be able to describe and discuss patterns found in a contingency table and associated displays of conditional distributions.

Exercises

1. Find a bar graph of categorical data from a newspaper or magazine.
 - a) Is the graph clearly labeled?
 - b) Does it violate the area principle?
 - c) Does the accompanying article tell the W's of the variable?
 - d) Do you think the article correctly interprets the data? Explain.
2. Find a pie chart of categorical data from a newspaper or magazine.
 - a) Is the graph clearly labeled?
 - b) Does it violate the area principle?
 - c) Does the accompanying article tell the W's of the variable?
 - d) Do you think the article correctly interprets the data? Explain.
3. Find a frequency table of categorical data from a newspaper or magazine.
 - a) Is it clearly labeled?
 - b) Does it display percentages or counts?
 - c) Does the accompanying article tell the W's of the variable?
 - d) Do you think the article correctly interprets the data? Explain.
4. Tables in the news II. Find a contingency table of categorical data from a newspaper or magazine.
 - a) Is it clearly labeled?
 - b) Does it display percentages or counts?
 - c) Does the accompanying article tell the W's of the variables?
 - d) Do you think the article correctly interprets the data? Explain.

5. The Centers for Disease Control lists causes of death in the United States during 1999.

Cause of death	Percent
Heart disease	30.3
Cancer	23.0
Circulatory diseases and stroke	8.4
Respiratory diseases	7.9
Accidents	4.1

- a) Is it reasonable to conclude that heart or respiratory diseases were the cause of approximately 38% of U.S. deaths in 1999?
 - b) What percent of deaths were from causes not listed here?
 - c) Create an appropriate display for these data.
6. In a December 2000 report, the U.S. Census Bureau listed the levels of educational attainment for Americans over 65. Create an appropriate display for these data, and write a sentence or two that might appear in a newspaper article about the report.

Education level	Count (thousands)
No high school diploma	9,945
HS graduate, but no college	11,701
Some college, no degree	4,481
2-year degree	1,390
4-year degree	3,133
Master's degree	1,213
Ph.D. or professional degree	757

7. A May 2001 Gallup Poll found that many Americans believe in ghosts and other supernatural phenomena. The poll was based on telephone responses from 1012 randomly selected adults. The table shows the percentages of people who expressed belief in various phenomena.

Phenomenon	Percent expressing belief
Psychic healing	54
ESP	50
Ghosts	38
Astrology	28
Channeling	15

- a) Is it reasonable to conclude that 66% of those polled expressed belief in either ghosts or astrology?
- b) Can you tell what percent of people did not believe in any of these phenomena? Explain.
- c) Create an appropriate display for these data.

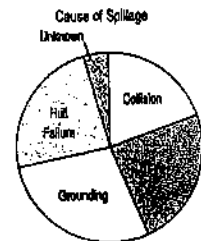
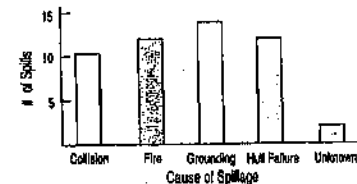
8. A study by the U.S. Bureau of Alcohol, Tobacco, and Firearms (BATF) (USA Today, 22 June 2000) surveyed 1530 investigations by the U.S. BATF into illegal gun trafficking from July 1996 through December 1998. The study reports the portion of cases that were the result of each of five gun trafficking violations:

- 46% Straw purchase (legal gun buyer acting on behalf of an illegal buyer)
- 21% Unlicensed sellers
- 14% Gun shows and flea markets
- 14% Stolen from federally licensed dealers
- 10% Stolen from residences

- a) State the W's for this study to the extent the story gives them.
- b) What do you notice about the percentages listed? What does that probably mean?
- c) Make a bar chart to display these results, and label it correctly.
- d) Write a brief report on what they say about illegal gun trafficking. Note any possible problems with the data.
- e) The study also noted that although corrupt licensed dealers accounted for 133 of the 1,530 investigations, they were linked to 40,365 of the 84,128 firearms involved in those investigations. How does this new information about the *Who* of the study affect your conclusions?

9. Oil spills. To improve the structural design of oil tankers with the objective of reducing the likelihood of an oil spill and decreasing the amount of outflow in the event of a hull puncture, a study (*Marine Technology*, Jan.

1995) reported the spillage amount and cause of puncture for 50 recent major oil spills from tankers and carriers. Here are displays. Write a brief report interpreting what the displays show. Is a pie chart an appropriate display for these data? Why or why not?



10. 25 countries won medals in the 2002 Winter Olympics. The table lists them, along with the total number of medals each won:

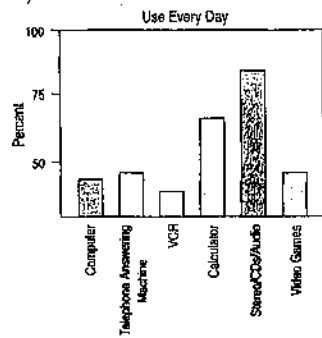
Country	Medals	Country	Medals
Germany	35	Croatia	4
USA	34	Korea	4
Norway	24	Bulgaria	3
Canada	17	Estonia	3
Austria	16	Great Britain	3
Russia	16	Australia	2
Italy	12	Czech Republic	2
France	11	Japan	2
Switzerland	11	Poland	2
China	8	Spain	2
Netherlands	8	Belarus	1
Finland	7	Slovenia	1
Sweden	6		

- a) Try to make a display of these data. What problems do you encounter?
- b) Can you find a way to organize the data so that the graph is more successful?

11. The Gallup Organization, in conjunction with CNN, USA Today, and the National Science Foundation, conducted a national survey of 744 children in grades 7 through 12—mostly comprised of students in the 'teenage' years of 13 to 17. Telephone interviews were conducted from March 20–27, 1997 from Gallup interviewing centers throughout the country. The focus of the survey was on students' familiarity with and use of modern technology, with special attention given to use of computers and the Internet. The teenagers were asked if they used each of the following technologies on a daily basis and if the technology was critically important to own. For each question, the percentage of those responding yes is given. Gallup dubbed the difference between the two percentages the "Importance Gap." Here are the results:

Technology	Use daily	Critically important to own	Importance gap
Computer	44%	77%	33
Telephone answering machine	46%	62%	16
VCR	39%	51%	12
Calculator	67%	71%	4
Stereo/CDs/audio	85%	69%	-16
Video games	46%	18%	-28

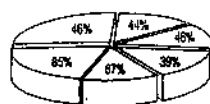
Consider the following graphical display showing the percentages of teens that use each of the technologies on a daily basis.



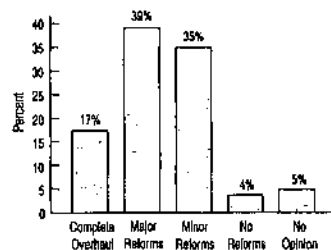
a) How much larger is the proportion of teens who use a calculator daily than the corresponding proportion for answering machines?

- b) Is that the impression given by the display? Explain.
 c) How would you improve this display?
 d) Make an appropriate display for the Importance Gap. (Hint: Make the y-axis of your chart span the range from -30 to +35.)
 e) Write a few sentences describing what you have learned about teens' attitudes toward technology.

12. Here's a display of the percentages of students who use the various technologies daily. List the errors in this display.



13. In the wake of the Enron Corporation scandal, the Gallup Organization asked 1001 American adults what kind of changes, if any, are needed in the way major corporations are audited. Here is a display of the results.



- a) Make a pie chart of the same data.
 b) Which chart works better to summarize the data? Why?
 c) Summarize the findings of the poll in a few sentences that might appear in a newspaper article.

14. A survey of autos parked in student and staff lots at a large university classified the brands by country of origin, as seen in the table.

Origin	Driver	
	Student	Staff
American	107	105
European	33	12
Asian	55	47

- a) What percent of all the cars surveyed were foreign?
 b) What percent of the American cars were owned by students?
 c) What percent of the students owned American cars?
 d) What is the marginal distribution of origin?
 e) What are the conditional distributions of origin by driver classification?
 f) Do you think that origin of the car is independent of the type of driver? Explain.

15. Prior to graduation a high school class of 2000 was surveyed about their plans. The table below displays the results for white and minority students. (The "Minority" group included African-American, Asian, Hispanic, and Native American students.)

Plans	White	Minority
4-year college	198	44
2-year college	36	6
Military	4	1
Employment	14	3
Other	16	3

- a) What percent of the graduates are white?
 b) What percent of the graduates are planning to attend a 2-year college?
 c) What percent of the graduates are white and planning to attend a 2-year college?
 d) What percent of the white graduates are planning to attend a 2-year college?
 e) What percent of the graduates planning to attend a 2-year college are white?
 f) Create a graph comparing the plans of white and minority students.
 g) Do you see any important differences in the post-graduation plans of white and minority students? Write a brief summary of what these data show, including comparisons of conditional distributions.

16. The table below compares what Ithaca High School students did after graduation in 1959, 1970, and 1980.

What Graduates Did	1959	1970	1980
Continuing education	197	388	320
Employed	103	137	98
In the military	20	18	18
Other	13	58	45

- a) What percent of all these graduates joined the military?
 b) What percent of these students graduated in 1970?

- c) What percent of the 1970 graduates joined the military?
 d) Of the students in these surveys who joined the military, what percent graduated in 1970?
 e) What is the marginal distribution of postgraduation activities?
 f) What is the conditional distribution of postgraduation activities among the class of 1959?
 g) Does this study present any evidence that post-graduation plans have changed over this 21-year period? Write a brief description of these data. Include an appropriate graph.

17. Statistics Canada provides, on its Web site, the following data on the Canadian population (in thousands). (Zeros indicate counts below 500.)

- a) What percent of Canadian citizens speak only English?
 b) What percent of Canadian citizens speak French?
 c) What percent of Quebec residents speak French?
 d) What percent of French-speaking Canadians live in Quebec?
 e) Do you think that language knowledge and province of residence are independent for Canadians? Explain.

Province	English Only	French Only	Both	Neither	Total
Newfoundland	525	0	21	1	547
Prince Edward Island	118	0	15	0	133
Nova Scotia	813	1	84	1	900
New Brunswick	418	73	238	0	730
Quebec	359	3,952	2,661	74	7,045
Ontario	9,116	47	1,235	245	10,643
Manitoba	984	1	103	12	1,100
Saskatchewan	921	0	51	5	977
Alberta	2,455	2	179	34	2,669
British Columbia	3,342	2	249	97	3,690
Yukon Territory	27	0	3	0	31
Northwest Territories	36	0	3	1	39
Nunavut	20	0	1	4	25
Total	19,134	4,078	4,843	474	28,529

18. Taiacos. A study by the University of Texas Southwestern Medical Center examined 626 people to see if there was an increased risk of contracting hepatitis C associated with

having a tattoo. If the subject had a tattoo, researchers asked whether it had been done in a commercial tattoo parlor or elsewhere. Write a brief description of the association between tattooing and hepatitis C, including an appropriate graphical display.

	Tattoo done in commercial parlor	Tattoo done elsewhere	No tattoo
Has hepatitis C	17	8	18
No hepatitis C	35	53	495

19. Just how accurate are the weather forecasts we hear every day? The table below compares the daily forecast with a city's actual weather for a year.

		Actual Weather	
		Rain	No rain
Forecast	Rain	27	63
	No rain	7	268

- a) On what percent of days did it actually rain?
 b) On what percent of days was rain predicted?
 c) What percent of the time was the forecast correct?
 d) Do you see evidence of an association between the type of weather and the ability of forecasters to make an accurate prediction? Write a brief explanation, including an appropriate graph.
20. **Federal Prisoners.** The table below shows the number of federal prison inmates serving sentences for various types of offenses in 1990 and 1998. Counts given are in thousands of prisoners.

Type of Offense	Year	
	1990	1998
Violent (murder, robbery, etc.)	10	13
Property (burglary, fraud, etc.)	8	9
Drugs	30	63
Public order (immigration, weapons, etc.)	9	22
Other	1	2

- a) Write a brief description of these data, in the proper context, highlighting any important changes you see in the prison population.

- b) Do these data indicate that there was an increase in drug use in the United States during the 1990's? Explain.

21. In July 1991 and again in April 2001 the Gallup Poll asked random samples of 1015 adults about their opinions on working parents. The table summarizes responses to this question:

"Considering the needs of both parents and children, which of the following do you see as the ideal family in today's society?"

Based upon these results, do you think there was a change in people's attitudes during the 10 years between these polls? Explain.

	Year	
	1991	2001
Both work full time	142	131
One works full time, other part time	274	244
One works, other works at home	152	173
One works, other stays home for kids	396	416
No opinion	51	51

22. In 2000 the *Journal of the American Medical Association (JAMA)* published a study that examined pregnancies that resulted in the birth of twins. Births were classified as preterm with intervention (induced labor or cesarean), preterm without procedures, or term/post-term. Researchers also classified the pregnancies by the level of prenatal medical care the mother received (inadequate, adequate, or intensive). The data, from the years 1995-97, are summarized in the table below. Figures are in thousands of births. (*JAMA* 284 [2000]:335-341)

TWIN BIRTHS 1995-1997 (IN THOUSANDS)

Level of Prenatal Care	Preterm (induced or cesarean)	Preterm (without procedures)	Term or postterm	Total
	Intensive	18	15	
Adequate	46	43	65	154
Inadequate	12	13	38	63
Total	76	71	131	278

- a) What percent of these mothers did not receive adequate medical care during their pregnancies?