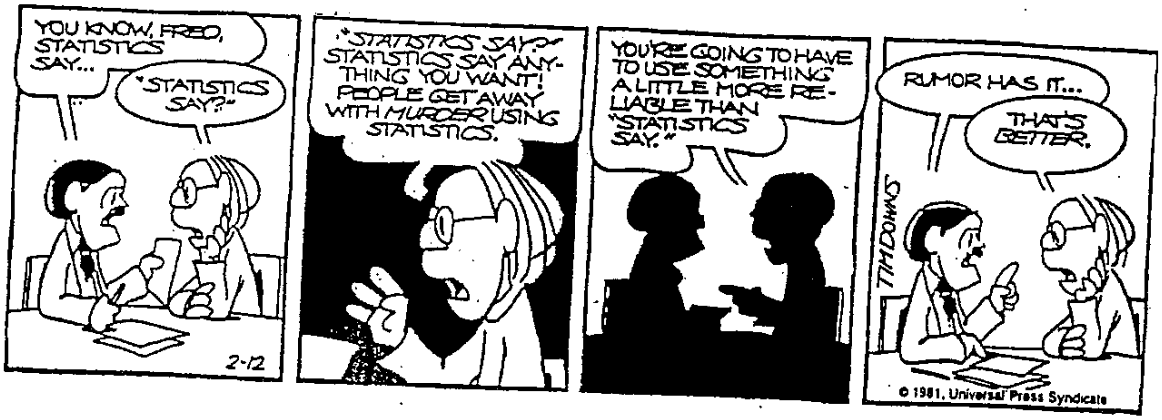
8. Do you believe that the students in your class can actually tell the difference between tap water and bottled water? Before you answer, let's perform a brief *simulation*. We'll assume that you and your classmates are guessing at which cup holds the bottled water. Then each student would have a 1-in-3 chance of identifying the correct cup. Roll your die once for each student in your class. Let rolling a 1 or a 2 represent a correct guess. Let rolling a 3 through 6 represent an incorrect guess. (Note that this assignment of numbers gives each individual a 2-in-6 chance of being correct, which is the same as a 1-in-3 chance.) Record the number of times that you get a 1 or a 2. This result simulates the number of correct identifications made by the class.
9. On a number line drawn on the board by your teacher, mark an X above the number of correct identifications in your simulation. Based on the class's simulation results, how many correct identifications would make you doubt that students were just guessing? Why?
10. Look back to your class's actual tasting results in Step 6. What do you conclude about students' abilities to distinguish tap water from bottled water?

Introduction

Do cell phones cause brain cancer? How well do SAT scores predict college success? Should arthritis sufferers take Celebrex to ease their pain, or are the risks too great? What percent of U.S. children are overweight? How strong is the evidence for global warming? These are just a few of the questions that statistics can help answer. But what is statistics? And why should you study it?


YOU KNOW, FRED, STATISTICS SAY...
"STATISTICS SAY?"
2-12
"STATISTICS SAY?" STATISTICS SAY ANYTHING YOU WANT! PEOPLE GET AWAY WITH *MURDER* USING STATISTICS.
YOU'RE GOING TO HAVE TO USE SOMETHING A LITTLE MORE RELIABLE THAN "STATISTICS SAY."
TIM DOWNS
RUMOR HAS IT...
THAT'S *BETTER*.

© 1981, Universal Press Syndicate

statistics

Statistics is the science (and art) of learning from data. Data are usually numbers, but they are not "just numbers." *Data are numbers with a context.* The number 10.5, for example, carries no information by itself. But if we hear that a friend's new baby weighed 10.5 pounds at birth, we congratulate her on the healthy size of the child. The context engages our background knowledge and allows us to make judgments. We know that a baby weighing 10.5 pounds is quite large, and that a human baby is unlikely to weigh 10.5 ounces or 10.5 kilograms. The context makes the number informative.

You can find lots of data in newspapers and magazines and on the Internet. Such data are ripe for exploration. Part I of this book focuses on *exploratory data analysis.* In Chapters 1 through 4, you'll develop tools and strategies for organizing, describing, and analyzing data.

Sometimes data provide insights about questions we have asked. More often, researchers follow a careful plan for *producing data* to answer specific questions. In Part II of the book, you'll discover how they do it. Chapter 5 shows you how to design *surveys, experiments,* and *observational studies* correctly. Such well-produced data help us get the most reliable answers to difficult questions.

Probability is the study of chance behavior. When you flip coins, roll dice, deal cards, or play the lottery, the results are uncertain. But the laws of probability can tell you how likely (or unlikely) certain outcomes are. You'll learn how to calculate probabilities in Part III of the book, in Chapters 6 through 9.

population
sample

How can we draw conclusions about a large group (***population***) of *individuals*—people, animals, or things—based on information about a much smaller group (***sample***)? This is the challenge of *statistical inference.* Inference questions ask us to test claims about or provide estimates for unknown population values. Valid inference depends on appropriate data production, skillful data analysis, and careful use of probability. In Part IV of the book, we discuss the logic behind statistical inference and some of its methods. In Chapters 10 through 15, you'll learn some common ways of testing claims and computing estimates.

This Preliminary Chapter is intended to give you a snapshot of what statistics is all about. Where do data come from? What should you do with data once you have them? How can probability help you? What conclusions can you draw? Keep reading for some answers.

In your lifetime, you will be bombarded with data and statistical information. Opinion poll results, television ratings, gas prices, unemployment rates, medical study outcomes, and standardized test scores are discussed daily in the media. People make important decisions based on such data. Statistics will help you make sense of information like this. A solid understanding of statistics will enable you to make sound decisions based on data in your everyday life.

Data Production: Where Do You Get Good Data?

You want data on a question that interests you. Maybe you want to know what causes of death are most common among young adults, or whether the math performance of American schoolchildren is getting better.

It is tempting just to draw conclusions from our own experience, making no use of more representative data. You think (without really thinking) that the students at your school are typical. We hear a lot about AIDS, so we assume it must be a leading cause of death among young people. Or we recall an unusual incident that sticks in our memory exactly because it is unusual. When an airplane crash kills several hundred people, we fear that flying is unsafe, even though data on all flights show that flying is much safer than driving. Here's an example that shows why **data beat personal experiences.**

Example P.1

Power lines and cancer
Got data?

Does living near power lines cause leukemia in children? The National Cancer Institute spent 5 years and $5 million gathering data on this question. The researchers compared 638 children who had leukemia with 620 who did not. They went into the homes and actually measured the magnetic fields in children's bedrooms, in other rooms, and at the front door. They recorded facts about power lines near the family home and also near the mother's residence when she was pregnant. Result: no connection between leukemia and exposure to magnetic fields of the kind produced by power lines was found. The editorial that accompanied the study report in the *New England Journal of Medicine* proclaimed, "It is time to stop wasting our research resources" on the question.[2]

Now consider a devastated mother whose child has leukemia and who happens to live near a power line. In the public mind, the striking story wins every time. A statistically literate person, however, knows that data are more reliable than personal experience because they systematically describe an overall picture rather than focus on a few incidents.

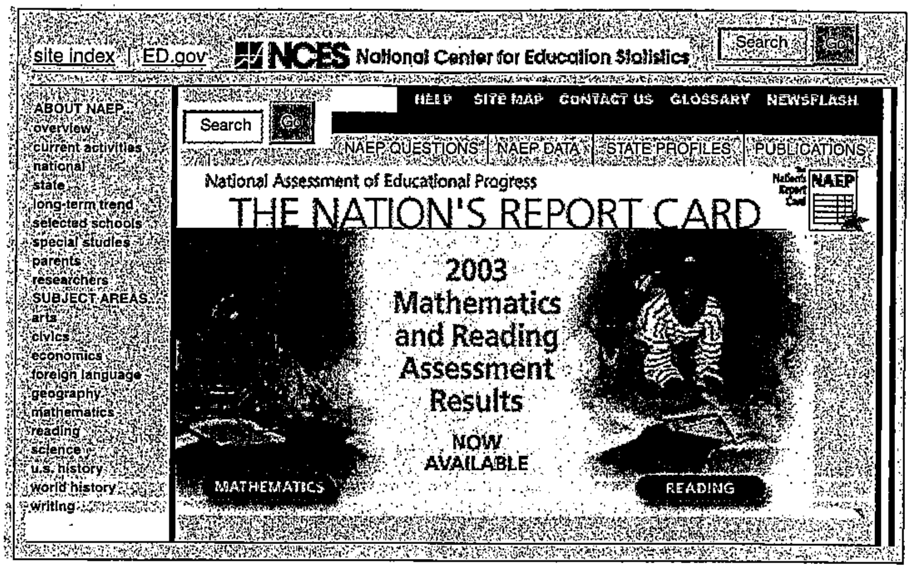
A better tactic is to head for the library or the Internet. There you will find plenty of data, not gathered specifically to answer your questions but available for your use. Recent data can be found online, but locating them can be challenging. Government statistical offices are the primary source for demographic, social, and economic data. Many nations have a single statistical office, like Statistics Canada (www.statcan.ca) or Mexico's INEGI (www.inegi.gob.mx). The United States does not have a national statistical office. More than 70 federal agencies collect data. Fortunately, you can reach most of them through the government's handy FedStats site (www.fedstats.gov).

Example P.2

Causes of death and math scores
Finding data on the Internet

If you visit the National Center for Health Statistics Web site, www.cdc.gov/nchs, you will learn that accidents are the most common cause of death for U.S. citizens aged 20 to 24, accounting for over 40% of all deaths. Homicide is next, followed by suicide. AIDS ranks seventh, behind heart disease and cancer, at 1% of all deaths. The data also show that it is dangerous to be a young man: the overall death rate for men aged 20 to 24 is three times that for women, and the death rate from homicide is more than five times higher among men.

Figure P.1 *The Web sites of government statistical offices are prime sources of data. Here is the home page of the National Assessment of Educational Progress.*



If you go to the National Center for Education Statistics Web site, www.nces.ed.gov, you will find the latest National Assessment of Educational Progress (Figure P.1), which provides full details about the math skills of schoolchildren. Math scores have slowly but steadily increased since 1990. All racial/ethnic groups, both girls and boys, and students in most states are getting better in math.

The library and the Internet are sources of ***available data.***

Available Data

Available data are data that were produced in the past for some other purpose but that may help answer a present question.

Available data are the only data used in most student reports. Because producing new data is expensive, we all use available data whenever possible. However, the clearest answers to present questions often require data produced to answer those specific questions. The main statistical designs for producing data are *surveys, experiments,* and *observational studies.*

surveys

Surveys are popular ways to gauge public opinion. The idea of a survey is pretty simple:

- Select a *sample* of people to represent a larger *population.*

- Ask the individuals in the sample some questions and record their responses.
- Use sample results to draw some conclusions about the population.

In practice, however, getting valid survey results is not so easy. As the following example shows, where the data come from is important.

Example P.3 *Having kids or not?*

Good and bad survey results

The advice columnist Ann Landers once asked her readers, "If you had it to do over again, would you have children?" A few weeks later, her column was headlined "70% OF PARENTS SAY KIDS NOT WORTH IT." Indeed, 70% of the nearly 10,000 parents who wrote in said they would not have children if they could make the choice again. Do you believe that 70% of all parents regret having children?

You shouldn't. The people who took the trouble to write Ann Landers are not representative of all parents. Their letters showed that many of them were angry at their children. All we know from these data is that there are some unhappy parents out there. A statistically designed poll, unlike Ann Landers's appeal, targets specific people chosen in a way that gives all parents the same chance to be asked. Such a poll later showed that 91% of parents would have children again.

The lesson: if you are careless about how you get your data, you may announce 70% "No" when the truth is close to 90% "Yes."

census

You may have wondered: why not survey everyone in the population (a *census*) rather than a sample? Usually, it would take too long and cost too much. Our goal in choosing a sample is a picture of the population, disturbed as little as possible by the act of gathering information. Sample surveys are one kind of ***observational study.***

In other settings, we gather data from an ***experiment.*** In doing an experiment, we don't just observe individuals or ask them questions. We actually do something to people, animals, or objects to observe the response. Experiments can answer questions such as "Does aspirin reduce the chance of a heart attack?" and "Do more college students prefer Pepsi to Coke when they taste both without knowing which they are drinking?" Experiments, like samples, provide useful data only when properly designed. The distinction between experiments and observational studies is one of the most important ideas in statistics.

Observational Study versus Experiment

In an observational study, we observe individuals and measure variables of interest but do not attempt to influence the responses.

In an experiment, we deliberately do something to individuals in order to observe their responses.

The next example illustrates the difference between an observational study and an experiment.

Example P.4 *Estrogen and heart attacks*

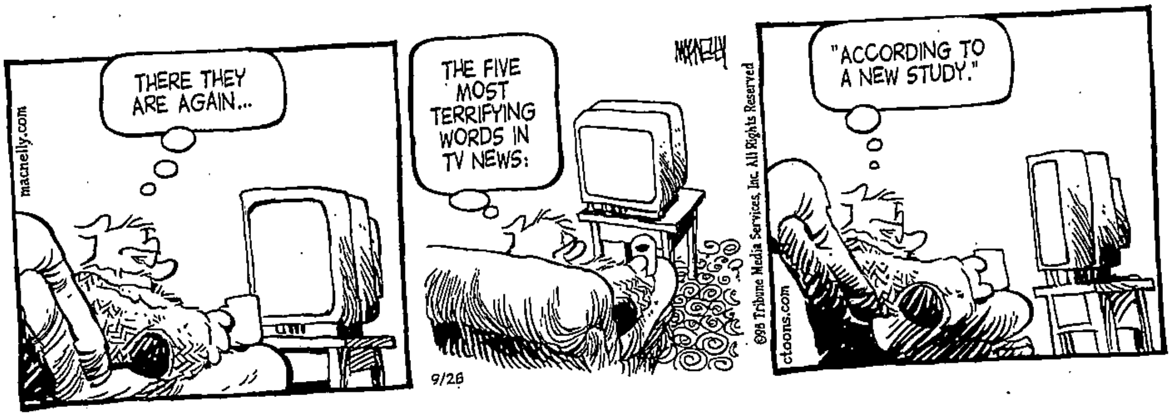
Observational study versus experiment

Should women take hormones such as estrogen after menopause, when natural production of these hormones ends? In 1992, several major medical organizations said "Yes." Women who took hormones seemed to reduce their risk of a heart attack by 35% to 50%. The risks of taking hormones appeared small compared with the benefits.

The evidence in favor of hormone replacement came from a number of studies that simply compared women who were taking hormones with others who were not. But women who chose to take hormones were typically richer and better educated, and they saw doctors more often than women who did not take hormones. These women did many things to maintain their health. It isn't surprising that they had fewer heart attacks.

Experiments were needed to get convincing data on the link between hormone replacement and heart attacks. In the experiments, women did not decide what to do. A coin toss assigned each woman to one of two groups. One group took hormone replacement pills; the other took dummy pills that looked and tasted the same as the hormone pills. All kinds of women were equally likely to get either treatment. By 2002, several experiments with women of different ages showed that hormone replacement does *not* reduce the risk of heart attacks. The National Institutes of Health, after reviewing the evidence, concluded that the earlier observational studies were wrong. Taking hormones after menopause fell quickly out of favor.[3]

Observational studies are essential sources of data about topics from the opinions of voters to the behavior of animals in the wild. But an observational study, even one based on a statistical sample, is a poor way to gauge the effect of a change. To see the response to a change, we must actually impose the change. When our goal is to understand cause and effect, experiments are the best source of convincing data.


macnelly.com
THERE THEY ARE AGAIN...
THE FIVE MOST TERRIFYING WORDS IN TV NEWS:
9/28
MACNELLY
©98 Tribune Media Services, Inc. All Rights Reserved
ctoons.com
"ACCORDING TO A NEW STUDY."

Exercises

P.1 Need a Jolt? Jamie is a hard-core computer programmer. She and all her friends prefer Jolt cola (caffeine equivalent to two cups of coffee) to either Coke or Pepsi (caffeine equivalent to less than one cup of coffee). Explain why Jamie's preference is not good evidence that most young people prefer Jolt to Coke or Pepsi.

P.2 Cell phones and brain cancer One study of cell phones and the risk of brain cancer looked at a group of 469 people who have brain cancer. The investigators matched each cancer patient with a person of the same age, sex, and race who did not have brain cancer, then asked about the use of cell phones.[4] Result: "Our data suggest that the use of handheld cellular phones is not associated with risk of brain cancer."

(a) Is this an observational study or an experiment? Justify your answer.

(b) Based on this study, would you conclude that cell phones do not increase the risk of brain cancer? Why or why not?

P.3 Learning biology with computers An educational software company wants to compare the effectiveness of its computer animation for teaching biology with that of a textbook presentation. The company gives a biology pretest to each of a group of high school juniors, then divides them into two groups. One group uses the animation, and the other studies the text. The company retests all students and compares the increase in biology test scores in the two groups.

(a) Is this an experiment or an observational study? Justify your answer.

(b) If the group using the computer animation has a much higher average increase in test scores than the group using the textbook, what conclusions, if any, could the company draw?

P.4 Survey, experiment, or observational study? What is the best way to answer each of the questions below: a survey, an experiment, or an observational study that is not a survey? Explain your choices. For each question, write a few sentences about how such a study might be carried out.

(a) Are people generally satisfied with how things are going in the country right now?

(b) Do college students learn basic accounting better in a classroom or using an online course?

(c) How long do your teachers wait on the average after they ask the class a question?

P.5 I'll drink to that! In adults, moderate use of alcohol is associated with better health. Some studies suggest that drinking wine rather than beer or spirits yields added health benefits.

(a) Explain the difference between an observational study and an experiment to compare people who drink wine with people who drink beer.

(b) Suggest some characteristics of wine drinkers that might benefit their health. In an observational study, these characteristics are mixed up with the effects of drinking wine on people's health.

P.6 Get a job! Find some information on this question: what percent of college undergraduates work part-time or full-time while they are taking classes? Start with the National Center for Education Statistics Web site, www.nces.ed.gov. Keep a detailed written record of your search.

Data Analysis: Making Sense of Data

data analysis

The first step in understanding data is to hear what the data say, to "let the statistics speak for themselves." But numbers speak clearly only when we help them speak by organizing, displaying, summarizing, and asking questions. That's *data analysis.*

Any set of data contains information about some group of ***individuals.*** The characteristics we measure on each individual are called ***variables.***

The importance of data integrity It has been accepted that global warming is a serious ecological problem. But Yale University researchers examined satellite and weather-balloon data collected since 1979 by NOAA (National Oceanic and Atmospheric Administration). They discovered that the satellites had drifted in orbit, throwing off the timing of temperature measurements. Nights looked as warm as days. Corrective action has shown that the pace of global warming over the past 30 years has actually been quite slow, a total increase of about 1 degree Fahrenheit. The lesson: always ask, "How were the data produced?"

Individuals and Variables

Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

A college's student data base, for example, includes data about every currently enrolled student. The students are the *individuals* described by the data set. For each individual, the data contain the values of *variables* such as age, gender, choice of major, and grade point average. In practice, any set of data is accompanied by background information that helps us understand it.

When you meet a new set of data, ask yourself the following *key questions:*

1. Who are the individuals described by the data? How many individuals are there?
2. What are the variables? In what units is each variable recorded? Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms.
3. Why were the data gathered? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for?
4. When, where, how, and by whom were the data produced? Where did the data come from? Are these available data or new data produced to answer current questions? Are the data from an experiment or an observational study? From a census or a sample? Who directed the data production? Can we trust the data?

Some variables, like gender and college major, simply place individuals into categories. Others, like age and grade point average (GPA), take numerical values for which we can do arithmetic. It makes sense to give an average GPA for a group of students, but it does not make sense to give an "average" gender. We can, however, count the numbers of female and male students and do arithmetic with these counts.

Categorical and Quantitative Variables

A categorical variable places an individual into one of several groups or categories.

A quantitative variable takes numerical values for which arithmetic operations such as adding and averaging make sense.

Example P.5 Education in the United States

Four key questions.

Here is a small part of a data set that describes public education in the United States:

State	Region	Population (1000s)	SAT verbal	SAT math	Percent taking	Percent no HS	Teachers' pay ($1000)
CA	PAC	35,894	499	519	54	18.9	54.3
CO	MTN	4,601	551	553	27	11.3	40.7
CT	NE	3,504	512	514	84	12.5	53.6

Answer the four key questions about these data.

1. Who? The *individuals* described are the states. There are 51 of them, the 50 states and the District of Columbia, but we give data for only 3: California (CA), Colorado (CO), and Connecticut (CT). Each row in the table describes one individual.

2. What? The rest of the columns each contain the values of one variable for all the individuals. This is the usual arrangement in data tables. Seven *variables* are recorded for each state. The second column lists which region of the country the state is in. Region is a categorical variable. The Census Bureau divides the nation into nine regions. These three are Pacific (PAC), Mountain (MTN), and New England (NE). The third column contains state populations, in thousands of people. Population is a quantitative variable. Be sure to notice that the *units* are thousands of people. California's 35,894 stands for 35,894,000 people.

The remaining five variables are the average scores of the states' high school seniors on the SAT verbal and mathematics exams, the percent of seniors who take the SAT, the percent of students who did not complete high school, and average teachers' salaries in thousands of dollars. These are all quantitative variables. Each of these variables needs more explanation before we can fully understand the data.

Beginning in March 2005, the new SAT consisted of three tests: Critical Reading, Math, and Writing.

3. Why? Some people will use these data to evaluate the quality of individual states' educational programs. Others may compare states using one or more of the variables. Future teachers might want to know how much they can expect to earn.

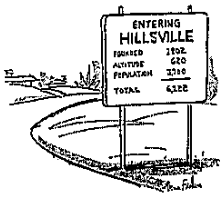
4. When, where, how, and by whom? The population data come from the Current Population Survey, conducted by the federal government. They are fairly accurate as of July 1, 2004, but don't show later changes in population. State SAT averages came from the College Board's Web site, www.collegeboard.com, and were based on a census of all test takers that year. The percent of students who did not graduate in each state was determined by the 2003 Current Population Survey. Average teacher salaries were reported in the 2003 *Statistical Abstract of the United States*, using data provided by the National Education Association for 2002. These data are estimates based on samples of teachers from each state.

A variable generally takes values that vary (hence the name "variable"!). Categorical variables sometimes have similar counts in each category and sometimes don't. For example, if you recorded values of the variable "birth month" for the students at your school, you would expect about an equal number of students in each of the categories (January, February, March, . . .). If you measured the variable "favorite type of music," however, you might see very different counts in the categories classical, gospel, rock, rap, and so on. Quantitative variables may take values that are very close together or values that are quite spread out. We call the pattern of variation of a variable its ***distribution.***

Distribution

The **distribution** of a variable tells us what values the variable takes and how often it takes these values.

exploratory data analysis



Statistical tools and ideas can help you examine data in order to describe their main features. This examination is sometimes called *exploratory data analysis.* (We prefer data analysis.) Like an explorer crossing unknown lands, we first simply describe what we see. Each example we meet will have some background information to help us, but our emphasis is on examining the data. Here are two basic strategies that help us organize our exploration of a set of data:

- Begin by examining each variable by itself. Then move on to study relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will organize our learning the same way. Chapters 1 and 2 examine single-variable data, and Chapters 3 and 4 look at relationships among variables. In both settings, we begin with graphs and then move on to numerical summaries.

Describing Categorical Variables

The values of a categorical variable are labels for the categories, such as "male" and "female." The distribution of a categorical variable lists the categories and gives either the *count* or the *percent* of individuals who fall in each category.

Example P.6 Do you wear your seat belt?

Describing categorical variables

Each year, the National Highway and Traffic Safety Administration (NHTSA) conducts an observational study on seat belt use. The table below shows the percent of front-seat passengers who were observed to be wearing their seat belts in each region of the United States in 1998 and 2003.[5]

Region	Percent wearing seat belts, 2003	Percent wearing seat belts, 1998
Northeast	74	66.4
Midwest	75	63.6
South	80	78.9
West	84	80.8

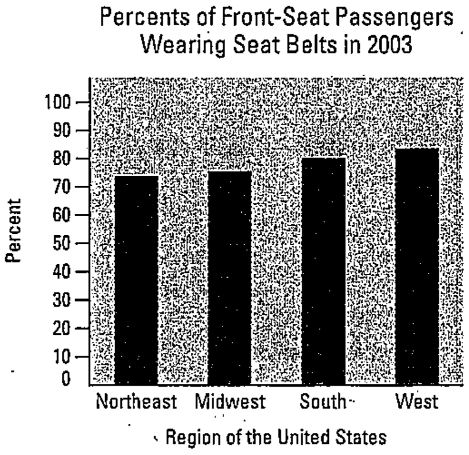
What do these data tell us about seat belt usage by front-seat passengers?

The *individuals* in this observational study are front-seat passengers. For each individual, the values of two *variables* are recorded: region (Northeast, Midwest, South, or West) and seat belt use (yes or no). Both of these variables are categorical.

bar graph

Figure P.2(a) shows a ***bar graph*** for the 2003 data. Notice that the vertical scale is measured in percents.

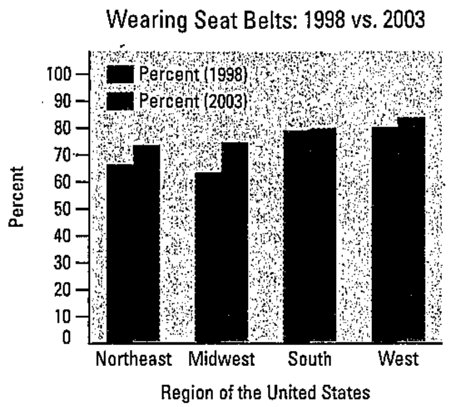
Figure P.2a (a) *A bar graph showing the percent of front-seat passengers who wore their seat belts in each of four U.S. regions in 2003.*



Front seat passengers in the South and West seem more concerned about wearing seat belts than those in the Northeast and Midwest. In all four regions, a high percent of front-seat passengers were wearing seat belts. Figure P.2(b) (on the next page) shows a ***side-by-side bar graph*** comparing seat belt usage in 1998 and 2003. Seat belt usage increased in all four regions over the five-year period.

side-by-side bar graph

Figure P.2b (b) *A side-by-side bar graph comparing the percent of front-seat passengers who wore their seat belts in the four U.S. regions in 1998 and 2003.*



Describing Quantitative Variables

dotplot

Several types of graphs can be used to display quantitative data. One of the simplest to construct is a *dotplot.*

Example P.7 *GOOOOAAAAALLLLLL!*
Describing quantitative variables

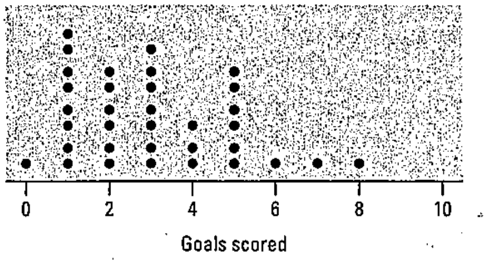
The number of goals scored by the U.S. women's soccer team in 34 games played during the 2004 season is shown below:[6]

3 0 2 7 8 2 4 3 5 1 1 4 5 3 1 1 3 3 3 2 1
2 2 2 4 3 5 6 1 5 5 1 1 5

What do these data tell us about the performance of the U.S. women's team in 2004?

A *dotplot* of the data is shown in Figure P.3. Each dot represents the goals scored in a single game. From this graph, we can see that the team scored between 0 and 8 goals per game. Most of the time, they scored between 1 and 5 goals. Their most frequent number of goals scored (the *mode*) was 1. They averaged 3.059 goals per game. (Check our calculation of the *mean* on your calculator.)

Figure P.3 *A dotplot of goals scored by the U.S. women's soccer team in 2004.*



Making a statistical graph is not an end in itself. After all, a computer or graph ing calculator can make graphs faster than we can. The purpose of a graph is tc help us understand the data. After you (or your calculator) make a graph, alway: ask, "What do I see?"

Exploring Relationships between Variables

CAUTION

Quite often in statistics, we are interested in examining the relationship between two variables. For instance, we may want to know how the percent of students taking the SAT in U.S. states is related to those states' average SAT math scores, or perhaps how seat belt usage is related to region of the country. As the next example illustrates, **many relationships between two variables are influenced by other variables lurking in the background.**

Example P.8 *On-time flights*

Describing relationships between variables

Air travelers would like their flights to arrive on time. Airlines collect data about on-time arrivals and report them to the Department of Transportation. Here are one month's data for flights from several western cities for two airlines:

	On time	Delayed
Alaska Airlines	3274	501
America West	6438	787

You can see that the percents of late flights were

$$\text{Alaska Airlines} \quad \frac{501}{3775} = 13.3\%$$

$$\text{America West} \quad \frac{787}{7225} = 10.9\%$$

It appears that America West does better.

This isn't the whole story, however. For each flight (individual), we have data on two categorical variables: the airline and whether or not the flight was late. Let's add data on a third categorical variable, departure city.[7] The following table summarizes the results.

	Alaska Airlines		America West	
Departure city	On time	Delayed	On time	Delayed
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Francisco	503	102	320	
Seattle	1841	305	201	
Total	3274	501	6438	

The "Total" row shows that the new table describes the same flights as the earlier table. Look again at the percents of late flights, first for Los Angeles:

$$\text{Alaska Airlines} \quad \frac{62}{559} = 11.1\%$$

$$\text{America West} \quad \frac{117}{811} = 14.4\%$$

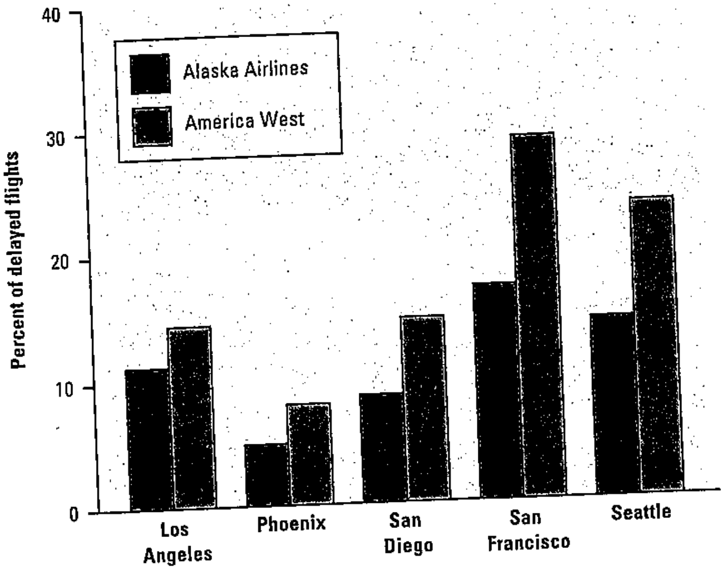
Alaska Airlines is better. The percents of late flights for Phoenix are

$$\text{Alaska Airlines} \quad \frac{12}{233} = 5.2\%$$

$$\text{America West} \quad \frac{415}{5255} = 7.9\%$$

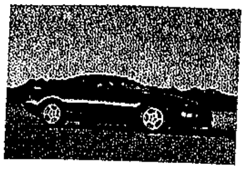
Alaska Airlines is better again. In fact, as Figure P.4 shows, Alaska Airlines has a lower percent of late flights at every one of these cities.

Figure P.4 *Comparing the percents of delayed flights for two airlines at five airports.*



How can it happen that Alaska Airlines wins at every city but America West wins when we combine all the cities? Look at the data: America West flies most often from sunny Phoenix, where there are few delays. Alaska Airlines flies most often from Seattle, where fog and rain cause frequent delays. What city we fly from has a major influence on the chance of a delay, so including the city data reverses our conclusion. (We'll see other examples like this one in Chapter 4 when we examine *Simpson's paradox*.) The message is worth repeating: many relationships between two variables (like airline and whether the flight was late) are influenced by other variables lurking in the background (like departure city).

Exercises



P.7 Cool car colors Here are data on the most popular car colors for vehicles made in North America during the 2003 model year.[8]

Color	Percent of vehicles
Silver	20.1
White	18.4
Black	11.6
Medium/dark gray	11.5
Light brown	8.8
Medium/dark blue	8.5
Medium red	6.9

(a) Display these data in a bar graph. Be sure to label your axes and title your graph.

(b) Describe what you see in a few sentences. What percent of vehicles had other colors?

P.8 Comparing car colors Favorite vehicle colors may differ among types of vehicle. Here are data on the most popular colors in 2003 for luxury cars and SUVs, trucks, and vans. The entry "—" means "less than 1%."

Color	Luxury car percent	SUV/truck/van percent
Black	10.9	11.6
Light brown	—	6.3
Medium/dark blue	3.8	9.3
Medium/dark gray	23.3	8.8
Medium/dark green	—	7.0
Medium red	3.9	6.2
White	30.4	22.3
Silver	18.8	17.0

(a) Make a side-by-side bar graph to compare colors by vehicle type.

(b) Write a few sentences describing what you see.

P.9 U.S. women's soccer scores In Example P.7 (page 16), we examined the number of goals scored by the U.S. women's soccer team in games during the 2004 season. Here are data on the goal differential for those same games, computed as U.S. score minus opponent's score.

3 0 2 7 8 2 4 1 4 1 −2 3 4 3 0 1 2 2 3 2 0
1 1 1 1 3 5 6 1 4 5 0 −2 5

(a) Make a dotplot of these data.

(b) Describe what you see in a few sentences.

P.10 Olympic gold! Olympic athletes like Michael Phelps, Natalie Coughlin, Amanda Beard, and Paul Hamm captured public attention by winning gold medals in the 2004 (a)

Summer Olympic Games in Athens, Greece. Table P.1 displays the total number of gold medals won by a sample of countries in the 2004 Summer Olympics.

Table P.1 Gold medals won by selected countries in the 2004 Summer Olympics

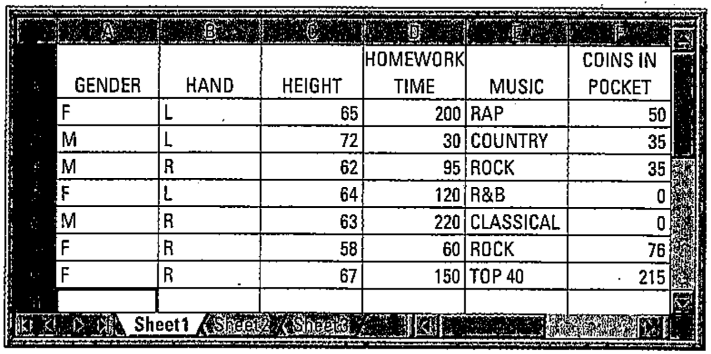
Country	Gold medals	Country	Gold medals
Sri Lanka	0	Netherlands	4
Qatar	0	India	0
Vietnam	0	Georgia	2
Great Britain	9	Kyrgyzstan	0
Norway	5	Costa Rica	0
Romania	8	Brazil	4
Switzerland	1	Uzbekistan	2
Armenia	0	Thailand	3
Kuwait	0	Denmark	2
Bahamas	0	Latvia	0
Kenya	1	Czech Republic	1
Trinidad and Tobago	0	Hungary	8
Greece	6	Sweden	4
Mozambique	0	Uruguay	0
Kazakhstan	1	United States	35

Source: BBC Olympics Web site, news.bbc.co.uk/sport1/hi/olympics_2004.

Make a dotplot to display these data. Describe the distribution of number of gold medals won.

(b) Overall, 202 countries participated in the 2004 Summer Olympics, of which 57 won at least one gold medal. Do you believe that the sample of countries listed in the table is representative of this larger population? Why or why not?

P.11 A class survey Here is a small part of the data set that describes the students in an AP Statistics class. The data come from anonymous responses to a questionnaire on the first day of class.



A	B	C	D	E	F
GENDER	HAND	HEIGHT	HOMEWORK TIME	MUSIC	COINS IN POCKET
F	L	65	200	RAP	50
M	L	72	30	COUNTRY	35
M	R	62	95	ROCK	35
F	L	64	120	R&B	0
M	R	63	220	CLASSICAL	0
F	R	58	60	ROCK	76
F	R	67	150	TOP 40	215

Answer the key questions (who, what, why, when, where, how, and by whom) for these data. For each variable, tell whether it is categorical or quantitative. Try to identify the units of measurement for any quantitative variables.

P.12 Medical study variables Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?

(a) Gender (female or male)

(b) Age (years)

(c) Race (Asian, black, white, or other)

(d) Smoker (yes or no)

(e) Systolic blood pressure (millimeters of mercury)

(f) Level of calcium in the blood (micrograms per milliliter)

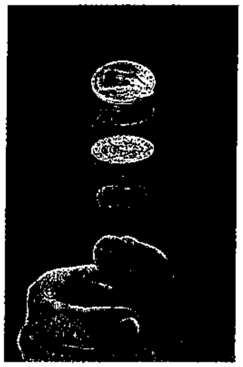
Probability: What Are the Chances?

probability

Consider tossing a single coin. The result is a matter of chance. It can't be predicted in advance, because the result will vary if you toss the coin repeatedly. But there is still a regular pattern in the results, a pattern that becomes clear only after many tosses. This remarkable fact is the basis for the idea of ***probability.***

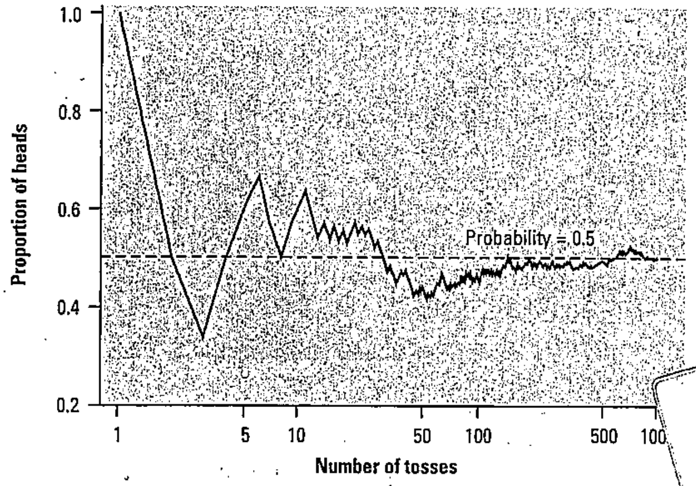
Example P.9 *Coin tossing*

Probability: what happens in the long run



When you toss a coin, there are only two possible outcomes, heads or tails. Figure P.5 shows the results of tossing a coin 1000 times. For each number of tosses from 1 to 1000, we have plotted the proportion of those tosses that gave a head. The first toss was a head, so the proportion of heads starts at 1. The second toss was a tail, reducing the proportion

Figure P.5 *The behavior of the proportion of coin tosses that give a head, from 1 to 1000 tosses of a coin. In the long run, the proportion of heads approaches 0.5, the probability of a head.*



of heads to 0.5 after two tosses. The next three tosses gave a tail followed by two heads, so the proportion of heads after five tosses is 3/5, or 0.6.

The proportion of tosses that produce heads is quite variable at first, but it settles down as we make more and more tosses. Eventually this proportion gets close to 0.5 and stays there. We say that 0.5 is the *probability* of a head. The probability 0.5 appears as a horizontal line on the graph.

Example P.9 illustrates the big idea of probability: **chance behavior is unpredictable in the short run but has a regular and predictable pattern in the long run.** Casinos rely on this fact to make money every day of the year. We can use probability rules to analyze games of chance, like roulette, blackjack, and Texas hold 'em.

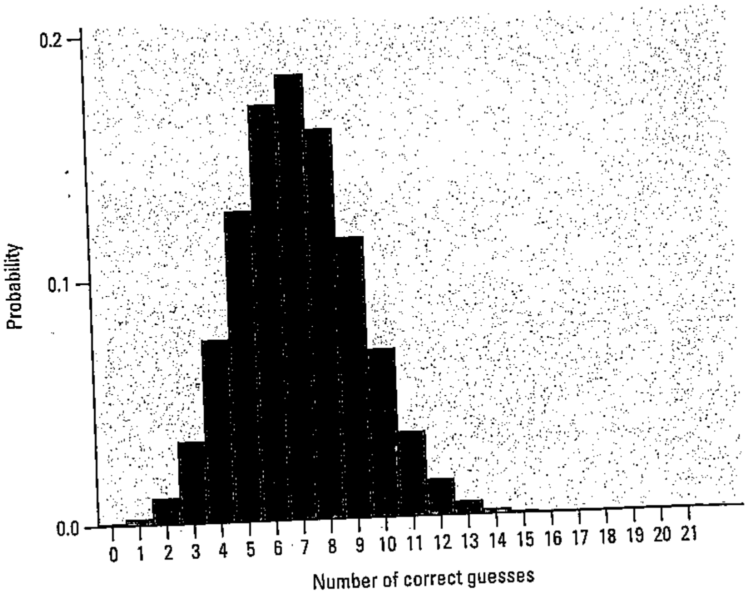
Probability plays an even more important role in the study of *variation.* If we toss a coin 30 times, will we get exactly 15 heads? Perhaps. Could we get as few as 11 heads? More than 24 heads? Probability tells us that there's about a 10% chance of getting 11 or fewer heads and less than a 1-in-1000 chance of getting more than 24 heads. If we toss our coin 30 times over and over and over again, the number of heads we obtain will vary. Probability quantifies the pattern of chance variation.

Example P.10 *Water, water everywhere*

Using probability to measure "how likely"

How can probability help us determine whether students can distinguish bottled water from tap water? Let's return to the Activity (page 4). Suppose that in Mr. Bullard's class, 13 out of 21 students made correct identifications. If we assume that the students in his

Figure P.6 *Graph showing the probability for each possible number of correct guesses in Mr. Bullard's class.*



class *cannot* tell bottled water from tap water, then each one is basically guessing, with a 1-in-3 chance of being correct. So we'd expect about one-third of his 21 students, that is, about 7 students, to guess correctly. How likely is it that as many as 13 of his 21 students would guess correctly?

Figure P.6 is a graph of the probability values for the number of correct guesses in Mr. Bullard's class. As you can see from the graph, the chance of guessing 13 or more correctly is quite small. In fact, the actual probability of doing so is 0.0068.

So what do we conclude? Either Mr. Bullard's students are guessing, and they have incredibly good luck, or the students are not guessing. Since the students have less than a 1% chance of getting so many right "just by chance," we feel pretty sure that they are not guessing. It seems that they can detect the difference in taste between tap and bottled water.

As the previous example shows, probability allows us to decide whether an observed outcome is too unlikely to be due to chance variation. Too many students were able to identify which of their three cups contained a different type of water for us to believe that they were guessing. In effect, we tested the claim that the students were guessing. This is our first encounter with ***statistical inference.*** Notice the important role that probability played in leading us to a conclusion.

statistical inference

Statistical Inference: Drawing Conclusions from Data

How prevalent is cheating on tests? Representatives from the Gallup Organization were determined to find out. They conducted an Internet survey of 1200 students, aged 13 to 17, between January 23 and February 10, 2003. The question they posed was "Have you, yourself, ever cheated on a test or exam?" Forty-eight percent of those surveyed said "Yes." If Gallup had asked the same question of *all* 13- to 17-year-old students, would exactly 48% have answered "Yes"?

Gallup is trying to estimate the unknown percent of students in this age group who would say they have cheated on a test. (Notice that we didn't say the percent of students who actually *had* cheated on a test!) Their best estimate, given the survey results, would be 48%. But the folks at Gallup know that samples vary. If they had selected a different sample of 1200 students to respond to the survey, then they would probably have gotten a different estimate. *Variation is everywhere!*

Fortunately, probability provides a description of how the sample results will vary in relation to the true population percent. Based on the sampling method that Gallup used, we can say that their estimate of 48% is very likely to be within 3% of the true population percent. That is, we can be quite confident that between 45% and 51% of *all* teenage students would say that they have cheated on a test.

Statistical inference allows us to use the results of properly designed experiments, sample surveys, and other observational studies to draw conclusions that go beyond the data themselves. Whether we are testing a claim, as in the bottled versus tap water Activity, or computing an estimate, as in the Gallup survey, we rely on probability to help us answer research questions with a known degree of confidence. Unfortunately, we cannot be *certain* that our conclusions are correct. The following example shows you why.

Example P.11 *Do mammograms help?*

Experiments and inference

Most women who reach middle age have regular mammograms to detect breast cancer. Do mammograms really reduce the risk of dying of breast cancer? To seek answers, doctors rely on "randomized clinical trials" that compare different ways of screening for breast cancer. We will see later that data from randomized comparative experiments are the gold standard. The conclusion from 13 such trials is that mammograms reduce the risk of death in women aged 50 to 64 years by 26%.[9]

On average, then, women who have regular mammograms are less likely to die of breast cancer. Of course, the results are different for different women. Some women who have mammograms every year die of breast cancer, and some who never have mammograms live to 100 and die when they crash their motorcycles. In spite of this individual variation, the results of the 13 clinical trials provide convincing evidence that women who have mammograms are less likely to die from breast cancer. That's because probability tells us that the large difference in death rates between women who had regular mammograms and those who didn't was unlikely to have occurred by chance. Can we be *sure* that mammograms reduce risk on the average? No, we can't be sure. **Because variation is everywhere, we cannot be certain about our conclusions.** However, statistics helps us better understand variation so that we can make reasonable conclusions.

Statistics gives us a language for talking about uncertainty that is used and understood by statistically literate people everywhere. In the case of mammograms, the doctors use that language to tell us that "mammography reduces the risk of dying of breast cancer by 26% (95% confidence interval, 17% to 34%)." According to Arthur Nielsen, head of the country's largest market research firm, that 26% is "a shorthand for a range that describes our actual knowledge of the underlying condition."[10] The range is 17% to 34%, and we are 95% confident that the true percent lies in that range. You will soon learn how to understand this language.

We can't escape variation and uncertainty. Learning statistics enables us to deal more effectively with these realities.

Statistical Thinking and You

The purpose of this book is to give you a working knowledge of the ideas and tools of practical statistics. Because data always come from a real-world context, doing statistics means more than just manipulating data. *The Practice of Statistics* is full of data, and each set of data has some brief background to help you understand what the data say. Examples and exercises usually express some brief understanding gained from the data. In practice, you would know much more about the background of the data you work with and about the questions you hope the data will answer. No textbook can be fully realistic. But it is important to form the habit of asking, "What do the data tell me?" rather than just concentrating on making graphs and doing calculations. This book tries to encourage good habits.

Still, statistics involves lots of calculating and graphing. The text presents the techniques you need, but you should use a calculator or computer software to automate calculations and graphs as much as possible.

Ideas and judgment can't (at least yet!) be automated. They guide you in telling the computer what to do and in interpreting its output. This book tries to explain the most important ideas of statistics, not just teach methods.

You learn statistics by doing statistical problems. This book offers four types of exercises, arranged to help you learn. Short problem sets appear after each major idea. These are straightforward exercises that help you solidify the main points before going on. The Section Exercises at the end of each numbered section help you combine all the ideas of the section. Chapter Review Exercises look back over the entire chapter. Finally, the Part Review Exercises provide challenging, cumulative problems like you might find on a final exam. At each step you are given less advance knowledge of exactly what statistical ideas and skills the problems will require, so each step requires more understanding.

Each chapter ends with a Chapter Review that includes a detailed list of specific things you should now be able to do. Go through that list, and be sure you can say "I can do that" to each item. Then try some chapter exercises.

The basic principle of learning is persistence. The main ideas of statistics, like the main ideas of any important subject, took a long time to discover and take some time to master. Once you put it all together—data analysis, data production, probability, and inference—statistics will help you answer important questions for yourself and for those around you.

Exercises

P.13 TV viewing habits You are preparing to study the television-viewing habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student. Give the units of measurement for the quantitative variables.

P.14 Roll the dice What is the probability of getting a "6" if you roll a fair six-sided die? Explain carefully what your answer means.

P.15 Tap water or bottled water, I Refer to Example P.10 (page 22). Which of the following results would provide more convincing evidence that Mr. Bullard's class could tell bottled water from tap water: 12 out of 21 correct identifications or 14 out of 21 correct identifications? Explain your answer.

P.16 Tap water or bottled water, II Refer to Example P.10 (page 22). Estimate the probability of getting 11 or more correct answers if the students were simply guessing. What would you conclude about whether Mr. Bullard's students could distinguish bottled water from tap water?

P.17 Spinning pennies Hold a penny upright on its edge under your forefinger on a hard surface, then snap it with your other forefinger so that it spins for some time before falling.

Is the coin equally likely to land heads or tails? Spin the coin a total of 20 times, recor whether it lands heads or tails each time.

(a) Make a graph like the one in Figure P.5 (page 21) that shows the proportion of heads after each toss.

(b) Based on your results, estimate the proportion of all spins of the coin that would be heads.

(c) What would you conclude about whether the coin lands heads half the time? Justify your answer.

(d) IN CLASS: Pool your results with those of your classmates. Would you change the conclusion you made in (c)? Why or why not?

P.18 Abstinence or not? An August 2004 Gallup Poll asked 439 teens aged 13 to 17 whether they thought young people should abstain from sex until marriage. 56% said "Yes."

(a) If Gallup had asked *all* teens aged 13 to 17 this question, would exactly 56% have said "Yes"? Explain.

(b) In this sample, 48% of the boys and 64% of the girls said "Yes." Are you convinced that a higher percent of girls than boys aged 13 to 17 feel this way? Why or why not?

CASE CLOSED!

Can magnets help reduce pain?

At the end of each chapter, you will be asked to use what you have learned to resolve the Case Study that opened the chapter. Just like in a court proceeding, you can exclaim "Case Closed!" when you have finished.

Start by reviewing the information in the magnets and pain relief Case Study (page 3). Then answer each of the following questions in complete sentences. Be sure to communicate clearly enough for any of your classmates to understand what you are saying.

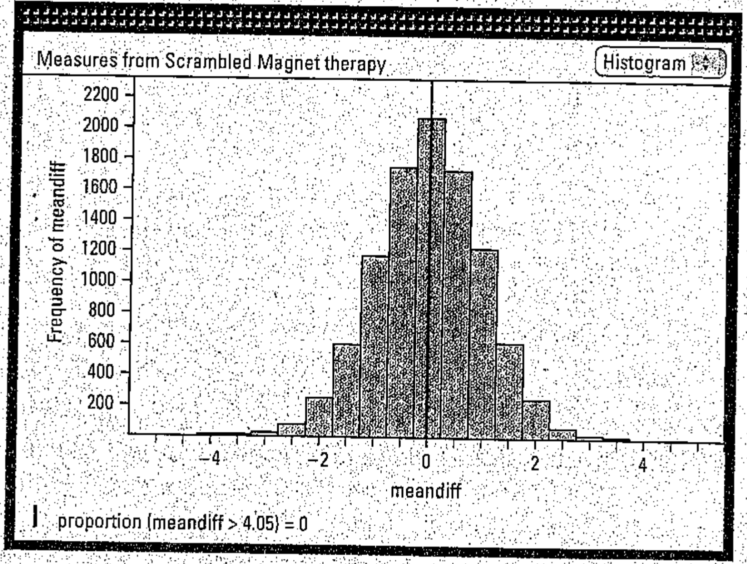
1. Data analysis
   a. Answer the key questions: who, what, why, when, where, how, and by whom?
   b. Construct separate dotplots of the pain ratings for the individuals in the active- and inactive-magnet groups. Draw your plots one above the other using the same scale.
   c. Describe what you see in your graphs.
   d. Calculate the *mean* (average) pain rating for each group. Now calculate the difference between the two means.

2. Producing data
   a. Were these available data or new data produced to answer a current question?
   b. Is the design of the study an experiment, a survey, or an observational study that is not a survey? Justify your answer.
   c. Why did researchers let chance decide (by picking from sealed envelopes) whether each patient received an active or inactive magnet?
   d. Would it matter if the patients or doctors knew which type of magnet they had? Explain.

Suppose that the active magnets don't really reduce pain. Then each patient should report the same final pain level whether he or she is assigned to the active- or inactive-magnet group. If the active and inactive magnets are equally effective, we should not observe a very large difference in the mean pain ratings of the two groups. Is the difference you calculated in Question 1(d) large or small? Before you answer, take a look at Figure P.7.

Figure P.7 displays the results of a *simulation* performed using Fathom statistical software. The computer reassigned the patients into the active-

Figure P.7 *Graph from Fathom statistical software displaying the difference in average pain score for the two groups in the magnets and pain study for 10,000 trials of a computer simulation.*



proportion (meandiff > 4.05) = 0

(continued)

and inactive-magnet groups 10,000 times, keeping each patient's final pain score the same as in the actual experiment. Each time, it computed the difference between the mean pain scores reported by the two groups. The graph displays the values of these 10,000 differences.

3. Probability
   a. Use the graph to estimate what percent of the time the difference in the groups' mean pain ratings is greater than 0. Explain your method.
   b. Based on the graph, how likely is it that the difference in mean pain ratings is greater than the one observed in this study (4.05) if the active magnets don't relieve pain?
4. Inference
   a. What would you estimate is the difference in mean pain relief when using active versus inactive magnets? Why?
   b. If you were testing the claim that the active magnets did not help reduce pain any better than the inactive magnets, what would you conclude? Explain.

Chapter Review

Summary

Statistics is the art and science of collecting, organizing, describing, analyzing, and drawing conclusions from data. When used properly, the tools of statistics can help us answer important questions about the world around us. This chapter gave you an overview of what statistics is all about: *data production, data analysis, probability,* and *statistical inference.*

Some people make decisions based on personal experiences. Statisticians make decisions based on data. **Data production** helps us answer specific questions with an **experiment** or an **observational study**. Experiments are distinguished from observational studies such as **surveys** by doing something intentionally to the individuals involved. A survey selects a **sample** from the **population** of all individuals about which we desire information. We base conclusions about the population on data about the sample.

A data set contains information on a number of **individuals**. Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables**. A variable describes some characteristic of an individual, such as a person's height, gender, or salary. Some variables are **categorical** and others are **quantitative**. A categorical variable places each individual into a category, like male or female. A quantitative variable has numerical values that measure some characteristic of each individual, like height in centimeters or annual salary in dollars. Remember to ask the key questions—who, what, why, when, where, how, and by whom?—about any data set.

The **distribution** of a variable describes what values the variable takes and how often it takes these values. To describe a distribution, begin with a graph. You can use **bar graphs** to display categorical variables. A **dotplot** is a simple graph you can use to show the distributions of quantitative variables. When examining any graph, ask yourself "What do I see?"

Exploratory data analysis uses graphs and numerical summaries to describe the variables in a data set and the relations among them. The conclusions of an exploratory analysis may not generalize beyond the specific data studied.

Probability is the language of chance. Chance behavior is unpredictable in the short run but follows a predictable pattern over many repetitions. When we're dealing with chance behavior, the rules of probability help us determine the likelihood of particular outcomes.

Statistical inference produces answers to specific questions, along with a statement of how confident we can be that the answer is correct. The conclusions of statistical inference are usually intended to apply beyond the individuals actually studied. Successful statistical inference requires production of data intended to answer the specific questions posed.

What You Should Have Learned

Here is a review list of the most important skills you should have acquired from your study of this chapter.

A. Where Do Data Come From?

1. Explain why we should not draw conclusions based on personal experiences.
2. Recognize whether a study is an experiment, a survey, or an observational study that is not a survey.
3. Determine the best method for producing data to answer a specific question: experiment, survey, or other observational study.
4. Locate available data on the Internet to help you answer a question of interest.

B. Dealing with Data

1. Identify the individuals and variables in a set of data.
2. Classify each variable as categorical or quantitative. Identify the units in which each quantitative variable is measured.
3. Answer the key questions—who, what, why, when, where, how, and by whom?—about a given set of data.

C. Describing Distributions

1. Make a bar graph of the distribution of a categorical variable. Interpret bar graphs.

2. Make a dotplot of the distribution of a quantitative variable. Describe what you see.
3. Given a relationship between two variables, identify variables lurking in the background that might affect the relationship.

D. Probability

1. Interpret probability as what happens in the long run.
2. Use simulations to determine how likely an outcome is to occur.

E. Statistical Inference

1. Use the results of simulations and probability calculations to draw conclusions that go beyond the data.
2. Give reasons why conclusions cannot be certain in a given setting.

Web Links

These sites are excellent sources for available data:

U.S. Census Bureau Home Page www.census.gov

Data and Story Library lib.stat.cmu.edu/DASL/

Chapter Review Exercises

P.19 TV violence A typical hour of prime-time television shows three to five violent acts. Linking family interviews and police records shows a clear association between time spent watching TV as a child and later aggressive behavior.[11]

(a) Explain why this is an observational study rather than an experiment.

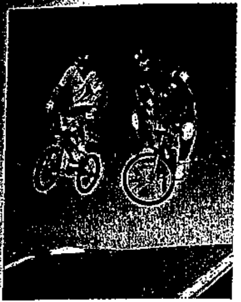
(b) Suggest several other variables describing a child's home life that may be related to how much TV he or she watches. Explain why these variables make it difficult to conclude that more TV *causes* aggressive behavior.

P.20 How safe are teen drivers? Find some information to help answer this question. Start with the National Highway and Traffic Safety Administration Web site, www.nhtsa.gov. Keep a detailed written record of your search.

P.21 Give it some gas! Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 2004 model motor vehicles:

Make and Model	Vehicle type	Transmission type	Number of cylinders	City MPG	Highway MPG
Acura NSX	Two-seater	Automatic	6	17	24
BMW 330I	Compact	Manual	6	20	30
Cadillac Seville	Midsize	Automatic	8	18	26
Ford F150 2WD	Standard pickup truck	Automatic	6	16	19

Answer the key questions (who, what, why, when, where, how, and by whom?) for these data. Visit the government's fuel economy Web site www.fueleconomy.gov for more information about how these data were produced. For each variable, tell whether it is categorical or quantitative. Be sure to identify the units of measurement for any quantitative variables.



P.22 Wearing bicycle helmets According to the 2003 Youth Risk Behavior Survey, 85.9% of high school students reported rarely or never wearing bicycle helmets. The table below shows additional results from this survey, broken down by gender and grade in school.

	Rarely or never wore bicycle helmets		
Grade	**Female (%)**	**Male (%)**	**Total (%)**
9	80.3	86.4	83.9
10	85.9	88.1	87.1
11	86.8	87.6	87.3
12	86.1	87.5	86.9

(a) Make a bar graph to show the percent of students in each grade who said they rarely or never wore bicycle helmets. Write a few sentences describing what you see.

(b) Now make a side-by-side bar graph to compare the percents of males and females at each grade level who said they rarely or never wore bicycle helmets. Describe what you see in a few sentences.

P.23 Three of a kind You read in a book on poker that the probability of being dealt three of a kind in a five-card poker hand is 1/50. Explain in simple language what this means.

P.24 Baseball and steroids Late in 2004, baseball superstar Barry Bonds admitted using creams and ointments that contained steroids. Bonds said he didn't know that these substances contained steroids. A Gallup Poll asked a random sample of U.S. adults whether they thought Bonds was telling the truth: 42% said "probably not" and 33% said "definitely not."

(a) Why did Gallup survey a random sample of U.S. adults rather than a sample of people attending a Major League Baseball game?

(b) If Gallup had surveyed all U.S. adults instead of a sample, about what percent of the responses would be "probably not"? "Definitely not"? Explain.

(c) Can we conclude based on these results that Barry Bonds is lying? Why or why not?

P.25 Magnets and pain, I Refer to Case Closed! (page 26). Suppose the difference in the mean pain scores of the active and inactive groups had been 2.5 instead of 4.05. What conclusion would you draw about whether magnets help relieve pain in postpolio patients? Explain.

P.26 Magnets and pain, II Refer to the chapter-opening Case Study (page 3). The researchers decided to analyze the patients' final pain ratings. It also makes sense to examine the *difference* between patients' initial pain ratings and their final pain ratings in the active and inactive groups. Here are the data:

Active: 10, 6, 1, 10, 6, 8, 5, 5, 6, 8, 7, 8, 7, 6, 4, 4, 7, 10, 6, 10, 6, 5, 5, 1, 0, 0, 0, 0, 1
Inactive: 4, 3, 5, 2, 1, 4, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1

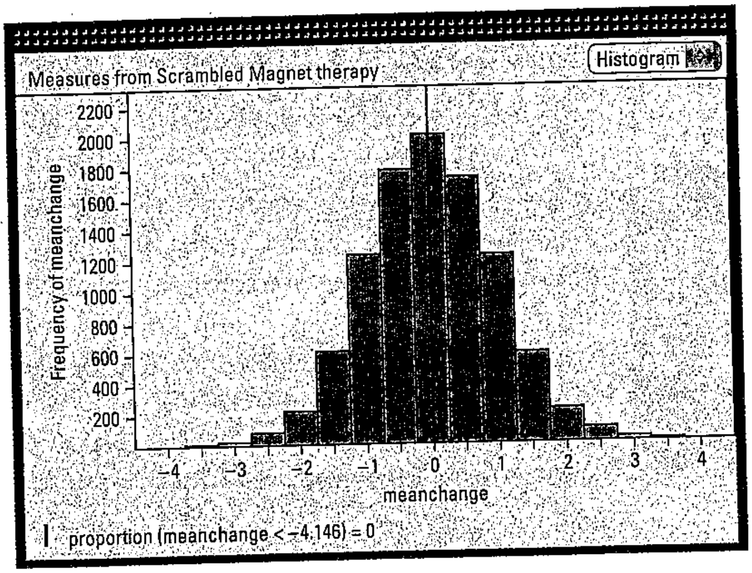
(a) Construct a dotplot for the active group's data. Describe what you see.

(b) Now make a dotplot for the inactive group's data immediately beneath using the same scale as the graph you made in (a). Write a few sentences comparing the changes in pain ratings for patients in the active and inactive groups.

(c) Calculate the mean (average) change in pain rating for each group.

(d) Figure P.8 shows the results of 10,000 repetitions of a computer simulation. As in Case Closed! (page 26), the computer redistributed the patients into the active- and inactive-magnet groups 10,000 times. Each time, it computed the difference between the mean "decrease in pain" scores reported by the two groups. The graph displays the values of these 10,000 differences. If you were testing the claim that the active magnets did not help reduce pain any better than the inactive magnets, what would you conclude? Explain.

Figure P.8 *Graph from Fathom statistical software displaying the difference in average decrease in pain for the two groups in the magnets and pain study for 10,000 trials of a computer simulation.*



proportion (meanchange < −4.146) = 0

P.27 Are you driving a gas guzzler? Table P.2 displays the highway gas mileage for 30 model year 2004 midsize cars.