

UC Connect
Analytical Methods Notes
From
Seeing Through Statistics
Jessica M. Utts

Chapter 1

The Benefits and Risks of Using Statistics

1.1 STATISTICS

When you hear the word *statistics*, you probably either get an attack of math anxiety or think about lifeless numbers, such as the population of the city or town where you live, as measured by the latest census, or the per capita income in Japan. The goal of this book is to open a whole new world of understanding of the term *statistics*. By the time you finish reading this book, you will realize that the invention of statistical methods is one of the most important developments of modern times. These methods influence everything from life-saving medical advances to which television shows remain on the air.

The word **statistics** is actually used to mean two different things. The better known definition is that statistics are numbers measured for some purpose. A more appropriate, complete definition is the following:

Statistics is a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty.

Using this definition, you have undoubtedly used statistics in your own life. For example, if you were faced with a choice of routes to get to school or work, or to get between one classroom building and the next, how would you decide which one to take? You would probably try each of them a number of times (thus gaining information) and then choose the best one according to some criterion important to you, such as speed, fewer red lights, more interesting scenery, and so on. You might even use different criteria on different days—such as when the weather is pleasant versus when it is not. In any case, by sampling the various routes and comparing them, you would have gained and processed useful information to help you make a decision.

In this book, you will learn ways to intelligently improve your own methods for collecting and processing complex information. You will learn how to interpret information that others have collected and processed and how to make decisions when faced with uncertainty. In Case Study 1.1, we will see how one researcher followed a casual observation to a fascinating conclusion.

Case Study 1.1: Heart or Hypothalamus?

SOURCE: Salk (1973), pp. 26–29.

You can learn a lot about nature by observation. You can learn even more by conducting a carefully controlled experiment. This case study has both. It all began when psychologist Lee Salk noticed that, despite his knowledge that the hypothalamus plays an important role in emotion, it was the heart that seemed to occupy the thoughts of poets and songwriters. There were no everyday expressions or song titles such as “I love you from the bottom of my hypothalamus” or “My hypothalamus longs for you.” Yet, there was no physiological reason for suspecting that the heart should be the centre of such attention. Why had it always been the designated choice?

Salk began wondering about the role of the heart in human relationships. He also noticed that when, on 42 separate occasions, he watched a rhesus monkey at the zoo holding her baby, and she held the baby on the left side, close to her heart, on 40 of those occasions. He then observed 287 human mothers within 4 days after giving birth and noticed that 237, or 83%, held their babies on the left. Handedness did not explain it; 83% of the right-handed mothers and 78% of the left-handed mothers exhibited the left-side preference. When asked why they chose the left side, the right-handed mothers said it was so their right hand would be free. The left-handed mothers said it was because they could hold the baby better with their dominant hand. In other words, both groups were able to rationalize holding the baby on the left based on their own preferred hand. Salk wondered if the left side would be favoured when carrying something other than a newborn baby. He found a study in which shoppers were observed leaving a supermarket carrying a single bag; exactly half of the 438 adults carried the bag on the left. But when stress was involved, the results were different. Patients at a dentist’s office were asked to hold a 5-inch rubber ball while the dentist worked on their teeth. Substantially more than half held the ball on the left.

Salk speculated that “it is not in the nature of nature to provide living organisms with biological tendencies unless such tendencies have survival value.” He surmised that there must indeed be survival value to having a newborn infant placed close to the sound of its mother’s heartbeat. To test this conjecture, Salk designed a study in a baby nursery at a New York City hospital. He arranged for the nursery to have the continuous sound of a human heartbeat played over a loudspeaker. At the end of 4 days, he measured how much weight the babies had gained or lost. Later, with a new group of babies in the nursery, no sound was played. Weight gains were again measured after 4 days. The results confirmed what Salk suspected. Although they did not eat more than the control group, the infants treated to the sound of the heartbeat gained more weight (or lost less). Further, they spent much less time crying. Salk’s conclusion was that “newborn infants are soothed by the sound of the normal adult heartbeat.” Somehow, mothers intuitively know that it is important to hold their babies on the left side. What had started as a simple observation of nature led to a further understanding of an important biological response of a mother to her newborn infant.

1.2 DETECTING PATTERNS AND RELATIONSHIPS

Some differences are obvious to the naked eye, such as the fact that the average man is taller than the average woman. If we were content to know only about such obvious relationships, we would not need the power of statistical methods. But had you noticed that babies who listen to the sound of a heartbeat gain more weight? Have you ever noticed that taking aspirin helps prevent heart attacks? How about the fact that people are more likely to buy blue jeans in certain months of the year than in others? The fact that men have lower resting pulse rates than women do? The fact that listening to Mozart improves performance on the spatial reasoning questions of an IQ test? All of these are relationships that have been demonstrated in studies using proper statistical methods, yet none of them are obvious to the naked eye. Let's take the simplest of these examples—one you can test yourself—and see what's needed to properly demonstrate the relationship. Suppose you wanted to verify the claim that, on average, men have lower resting pulse rates than women do. Would it be sufficient to measure only your own pulse rate and that of friend of the opposite sex? Obviously not. Even if the pair came out in the predicted direction, the singular measurements would certainly not speak for all members of each sex. It is not easy to conduct a study properly, but it is easy to understand much of how it should be done. We will examine each of the following concepts in great detail in the remainder of this book; here we just introduce them, using the simple example of comparing male and female pulse rates.

To conduct a study properly, one must

1. Get a representative sample.
2. Get a large enough sample.
3. Decide whether the study should be an observational study or an experiment.

1. Get a representative sample. Most researchers hope to extend their results beyond just the participants in their research. Therefore, it is important that the people or objects in a study be representative of the larger group for which conclusions are to be drawn. We call those who are actually studied a **sample** and the larger group from which they were chosen a **population**. (In Chapter 4 we will learn some ways to select a proper sample.) For comparing pulse rates, it may be convenient to use the members of your class. But this sample would not be valid if there were something about your class that would relate pulse rates and sex, such as if the entire men's track team happened to be in the class. It would also be unacceptable if you wanted to extend your results to an age group much different from the distribution of ages in your class. Often researchers are constrained to using such "convenience" samples, and we will discuss the implications of this later in the book.

2. Get a large enough sample. Even experienced researchers often fail to recognize the importance of this concept. In Part 4 of this book, you will learn how to detect the problem of a sample that is too small; you will also learn that such a sample can sometimes lead to erroneous conclusions. In comparing pulse rates, collecting one pulse rate from each sex obviously does not tell us much. Is two enough?

Four? One hundred? The answer to that question depends on how much *variability* there is among pulse rates. If all men had pulse rates of 65 and all women had pulse rates of 75, it wouldn't take long before you recognized a difference. However, if men's pulse rates ranged from 50 to 80 and women's pulse rates ranged from 52 to 82, it would take many more measurements to convince you of a difference. The question of how large is "large enough" is closely tied to how diverse the measurements are likely to be within each group. The more diverse, or variable, the individuals within each group, the larger the sample needs to be to detect a real difference between the groups.

3. Decide whether the study should be an observational study or an experiment.

For comparing pulse rates, it would be sufficient to measure or "observe" both the pulse rate and the sex of the people in our sample. When we merely observe things about our sample, we are conducting an **observational study**. However, if we were interested in whether frequent use of aspirin would help prevent heart attacks (which has been suggested as a likely possibility), it would not be sufficient to simply observe whether people frequently took aspirin and then whether they had a heart attack. It could be that people who were more concerned with their health were both more likely to take aspirin and less likely to have a heart attack, or vice versa.

To be able to make a causal connection, we would have to conduct an **experiment** in which we *randomly* assigned people to one of two groups. **Random assignments** are made by doing something akin to flipping a coin to determine the group membership for each person. In one group, people would be given aspirin and, in the other, they would be given a dummy pill that looked like aspirin. So as not to influence people with our expectations, we would not tell people which one they were taking until the experiment was concluded. In Case Study 1.2, we briefly examine such an experiment; in Chapter 5 we discuss these ideas in much more detail.

Does Aspirin Prevent Heart Attacks?

In 1988, the Steering Committee of the Physicians' Health Study Research Group released the results of a 5-year experiment conducted using 22,071 male physicians between the ages of 40 and 84. The physicians had been randomly assigned to two groups. One group took an ordinary aspirin tablet every other day, whereas the other group took a "placebo," a pill designed to look just like an aspirin but with no active ingredients. Neither group knew whether they were taking the active ingredient. The

results, shown in Table 1.1, support the conclusion that taking aspirin does indeed help reduce the risk of having a heart attack. The rate of heart attacks in the group taking aspirin was only 55% of the rate of heart attacks in the placebo group, or just slightly more than half as big. Because the men were randomly assigned to the two conditions, other factors, such as amount of exercise, should have been similar for both groups. The only substantial difference in the two groups should have been whether they took the aspirin or the placebo. Therefore, we can conclude that taking aspirin caused the lower rate of heart attacks for that group. Notice that, because the participants were all male physicians, these conclusions may not apply to the general population of men. They may not apply to women at all because no women were included in the study. More recent evidence has provided even more support for this effect, however, something we will examine in more detail in an example in Chapter 25.

Table 1.1: The Effect of Aspirin on Heart Attacks

Condition	Heart Attack	No Heart Attack	Attacks per 1000
Aspirin	104	10,933	9.42
Placebo	189	10,842	17.13

1.3 DON'T BE DECEIVED BY IMPROPER USE OF STATISTICS

Let's look at some examples representative of the kinds of abuses of statistics you may see in the media. In the example below, the statistics have been taken out of their proper context.

Example 1:

When a federal air report ranked the state of New Jersey as 22nd in the nation in its release of toxic chemicals, the New Jersey Department of Environmental Protection happily took credit (Wang, 1993, p. 170). The statistic was based on a reliable source, a study by the U.S. Environmental Protection Agency. However, the ranking had been made based on total pounds released, which was 38.6 million for New Jersey. When this total was turned into pounds per square mile in the state, New Jersey, it became apparent, was one of the worst—fourth on the list. Because New Jersey is one of the smallest states by area, the figures were quite misleading until adjusted for size.

Example 2:

An article headlined “New study confirms too much pot impairs brain” read as follows:

More evidence that chronic marijuana smoking impairs mental ability: Researchers at the University of Iowa College of Medicine say a test shows those who smoke seven or more marijuana joints per week had lower math, verbal and memory scores than non-marijuana users. Scores were particularly reduced when marijuana users held a joint's smoke in their lungs for longer periods.— San Francisco Examiner, 13 March 1993, p. D-1.

This research was clearly based on an observational study because people cannot be randomly assigned to either smoke marijuana or not. The headline is misleading because it implies that there is a causal connection between smoking marijuana and brain functioning. All we can conclude from an observational study is that there is a relationship. It could be the case that people who choose to smoke marijuana are those who would score lower on the tests anyway.

Example 3:

STUDY: SMOKING MAY LOWER KIDS' IQS

Source: Davis (CA)
Enterprise, 11 Feb. 1994.

Don't Make Causal Connections from Observational Studies

Rochester, N.Y. (AP)—Second-hand smoke has little impact on the intelligence scores of young children, researchers found. But women who light up while pregnant could be dooming their babies to lower IQs, according to a study released Thursday.

Children ages 3 and 4 whose mothers smoked 10 or more cigarettes a day during pregnancy scored about 9 points lower on the intelligence tests than the offspring of non-smokers, researchers at Cornell University and the University of Rochester reported in this month's Paediatrics journal. That gap narrowed to 4 points against children of non-smokers when a wide range of interrelated factors were controlled. The study took into account second-hand smoke as well as diet, education, age, drug use, parents' IQ, quality of parental care and duration of breast feeding. “It is comparable to the effects that moderate levels of lead exposure have on children's IQ scores,” said Charles Henderson, senior research associate at Cornell's College of Human Ecology in Ithaca.

Read the article above, and then read the headline again. Notice that the headline stops short of making a causal connection between smoking during pregnancy and lower IQs in children. Reading the article, you can see that the results are based on an observational study and not an experiment—with good reason: It would clearly be unethical to randomly assign pregnant women to either smoke or not. With studies like this, the best that can be done is to try to measure and statistically adjust for other factors that might be related to both smoking behaviour and children's IQ scores. Notice that when the researchers did so, the gap in IQ between the children of smokers and non-smokers narrowed from 9 points down to 4 points. There may be even more factors that the researchers did not measure that would account for the remaining 4-point difference. Unfortunately, with an observational study, we simply cannot make causal conclusions. We will explore this particular example in more detail in Chapter 6.

1.4 SUMMARY AND CONCLUSIONS

In this chapter, we have just begun to examine both the advantages and the dangers of using statistical methods. We have seen that it is not enough to know the results of a study, survey, or experiment. We also need to know how those numbers were collected and who was asked. In the upcoming chapters, you will learn much more about how to collect and process this kind of information properly and how to detect problems in what others have done. You will learn that a relationship between two characteristics (such as smoking marijuana and lower grades) does not necessarily mean that one causes the other, and you will learn how to determine other plausible explanations. In short, you will become an educated consumer of statistical information.

Chapter 2

Reading the News

2.1 THE EDUCATED CONSUMER OF DATA

Pick up any newspaper or newsmagazine and you are almost certain to find a story containing conclusions based on data. Should you believe what you read? Not always. It depends on how the data were collected, measured, and summarized. In this chapter, we discuss seven critical components of statistical studies. We examine the kinds of questions you should ask before you believe what you read. We go into further detail about these issues in subsequent chapters. The goal in this chapter is to give you an overview of how to be a more educated consumer of the data you encounter in your everyday life.

What Are Data?

In statistical parlance, **data** is a plural word referring to a collection of numbers or other pieces of information to which meaning has been attached. For example, the numbers 1, 3, and 10 are not necessarily data, but they become so when we are told that these were the weight gains in grams of three of the infants in Salk's heartbeat study, discussed in Chapter 1. In Case Study 1.2, the data consisted of two pieces of information measured on each participant: (1) whether they took aspirin or a placebo, and (2) whether they had a heart attack.

Don't Always Believe What You Read

When you read the results of a study in the newspaper, you are rarely presented with the actual data. Someone has usually summarized the information for you, and he or she has probably already drawn conclusions and presented them to you. Don't always believe them. The meaning we can attach to data, and to the resulting conclusions, depends on how well the information was acquired and summarized.

In the remaining chapters of Part 1, we look at proper ways to obtain data. In Part 2, we turn our attention to how it should be summarized. In Part 4, we learn the power as well as the limitations of using the data collected from a sample to make conclusions about the larger population. In the rest of this chapter, we address seven features of statistical studies that you should think about when you read a news article.

You will begin to be able to think critically and make your own conclusions about what you read.

2.2 HOW TO BE A STATISTICS SLEUTH:

SEVEN CRITICAL COMPONENTS

Reading and interpreting the results of surveys or experiments is not much different from reading and interpreting the results of other events of interest, such as sports competitions or criminal investigations. If you are a sports fan, then you know what information should be included in reports of competitions, and you know when crucial information is missing. If you have ever

been involved in an event that was later reported in the newspaper, you know that missing information can lead readers to erroneous conclusions.

In this section, you are going to learn what information should be included in news reports of statistical studies. Unfortunately, crucial information is often missing. With some practice, you can learn to figure out what's missing, as well as how to interpret what's reported. You will no longer be at the mercy of someone else's conclusions. You will be able to determine them yourself.

To provide structure to our examination of news reports, let's list Seven Critical Components that determine the soundness of statistical studies. A good news report should provide you with information about all of the components that are relevant to that study.

- Component 1:** The *source* of the research and of the *funding*.
- Component 2:** The *researchers* who had *contact* with the participants.
- Component 3:** The *individuals* or objects studied and how they were *selected*.
- Component 4:** The exact nature of the *measurements* made or *questions* asked.
- Component 5:** The *setting* in which the measurements were taken.
- Component 6:** The *extraneous differences* between groups being compared.
- Component 7:** The *magnitude* of any claimed effects or differences.

Before delving into some examples, let's examine each component more closely. You will find that most of the problems with studies are easy to identify. Listing these components simply provides a framework for using your common sense.

Component 1: The source of the research and of the funding

Studies are conducted for three major reasons. First, governments and private companies need to have data in order to make wise policy decisions. Information such as unemployment rates and consumer spending patterns are measured for this reason. Second, researchers at universities and other institutes are paid to ask and answer interesting questions about the world around us. The curious questioning and experimentation of such researchers have resulted in many social, medical, and scientific advances. Much of this research is funded by government agencies, such as the National Institutes of Health. Third, companies want to convince consumers that their programs and products work better than the competition, or special-interest groups want to prove that their point of view is held by the majority.

Unfortunately, it is not always easy to discover who funded research. Many university researchers are now funded by private companies. In her book *Tainted Truth* (1994), Cynthia Crossen warns us:

Private companies, meanwhile, have found it both cheaper and more prestigious to retain academic, government, or commercial researchers than to set up in-house operations that some might suspect of fraud. Corporations, litigants, political candidates, trade associations, lobbyists, special interest groups—all can buy research to use as they like. (p. 19).

If you discover that a study was funded by an organization that would be likely to have a strong preference for a particular outcome, it is especially important to be sure that correct scientific procedures were followed. In other words, be sure the remaining components have sound explanations.

Component 2: The researchers who had contact with the participants

It is important to know who actually had contact with the participants and what message those people conveyed. Participants often give answers or behave in ways to comply with the desires of the researchers. Consider, for example, a study done at a shopping mall to compare a new brand of a certain product to an old familiar brand. Shoppers are asked to taste each brand and state their preference. It is crucial that both the person presenting the two brands and the respondents be kept entirely blind as to which is which until after the preferences have been selected. Any clues might bias the respondent to choose the old familiar brand. Or, if the interviewer is clearly eager to have them choose one brand over the other, the respondents will most likely oblige in order to please. As another example, if you discovered that a study on the prevalence of illegal drug use was conducted by sending uniformed police officers door to door, you would probably not have much faith in the results. We will discuss other ways in which researchers influence participants in Chapters 4 and 5.

Component 3: The individuals or objects studied and how they were selected

It is important to know to whom the results can be extended. For example, until recently, many medical studies included men only, so the results were of little value to women. Many studies rely on volunteers recruited through the newspaper, who are usually paid a small amount for their participation. People who would respond to such recruitment efforts may differ in relevant ways from those who would not.

Studies relying on voluntary responses are likely to be biased because only those who feel strongly about the issues are likely to respond.

Component 4: The exact nature of the *measurements* made or *questions* asked

As you will see in Chapter 3, precisely defining and measuring most of the things researchers study isn't easy. For example, if you wanted to measure whether people "eat breakfast," how would you do so? What if they just have juice? What if they work until midmorning and then eat a meal that satisfies them until dinner? You need to understand exactly what the various definitions mean when you read about someone else's measurements. In polls and surveys, the "measurements" are usually answers to specific questions. Both the wording and the ordering of the questions can influence answers. For example, a question about "street people" would probably elicit different responses than a question about "families who have no home." Ideally, you should be given the exact wording that was used in a survey or poll,

Component 5: The *setting* in which the measurements were taken

The setting in which measurements were taken includes factors such as when and where they were taken and whether respondents were contacted by phone, mail, or in person. A study can be easily biased by timing. For example, opinions on whether criminals should be locked away for life may change drastically following a highly publicized murder or kidnapping case. If a study is conducted by telephone and calls are made only in the evening, certain groups of people would be excluded, such as those who work the evening shift or who routinely eat dinner in restaurants. Where the measurements were taken can also influence the results. Questions about sensitive topics, such as sexual behaviour or income, might be more readily answered over the phone, where respondents feel more anonymous. Sometimes research is done in a laboratory or university office, and the results may not readily extend to a natural setting. For example, studies of communication between two people are sometimes done by asking them to conduct a conversation in a university office with a tape recorder present. Such conditions almost certainly produce more limited conversation than would occur in a more natural setting.

Component 6: The *extraneous differences* between groups being compared

Often, when groups exhibit a difference in some feature of interest, researchers try to attribute the difference to membership in the group. For example, if the group of people who smoke marijuana has lower test scores than the group of people who doesn't, researchers may conclude that the lower test scores are due to smoking marijuana. Often, however, other disparities in the groups can explain the observed difference just as well. For example, people who smoke marijuana may simply be the type of people who are less motivated to study and thus would score lower on tests whether they smoked or not. Reports of research should include an explanation of any such extraneous differences that might account for the results. We will explore the issue of extraneous factors, and how to control for them, in much more detail in Chapter 5.

Component 7: The *magnitude* of any claimed effects or differences

Media reports about statistical studies often fail to tell you how large the observed effects were. Without that knowledge, it is hard for you to assess whether you think the results are of any practical importance. For example, if, based on Case Study 1.2, you were told simply that taking aspirin every other day reduced the risk of heart attacks, you would not be able to determine whether it would be worthwhile to take aspirin. You should instead be told that for the men in the study, the rate was reduced from about 17 heart attacks per 1000 participants without aspirin to about 9.4 heart attacks per 1000 with aspirin. Often news reports simply report that a treatment had an effect or that a difference was observed. We will investigate this issue in great detail in Part 4 of this book.

2.3 HYPOTHETICAL EXAMPLES OF BAD REPORTS

Throughout this book, you will see numerous examples of real studies and news reports. So that you can get some practice finding problems without having to read unnecessarily long news articles, let's examine some hypothetical reports. These are admittedly more problematic than many real reports because they serve to illustrate several difficulties at once.

Read each article and see if your common sense gives you some reasons why the headline is misleading. Then proceed to read the commentary about the Seven Critical Components.

Hypothetical News Article 1: “Study Shows Psychology Majors Are Smarter Than Chemistry Majors”

A fourth-year psychology student, for her senior thesis, conducted a study to see if students in her major were smarter than those majoring in chemistry. She handed out questionnaires in five advanced psychology classes and five advanced chemistry labs. She asked the students who were in class to record their grade-point averages (GPAs) and their majors. Using the data only from those who were actually majors in these fields in each set of classes, she found that the psychology majors had an average GPA of 3.05, whereas the chemistry majors had an average GPA of only 2.91. The study was conducted last Wednesday, the day before students went home to enjoy Thanksgiving dinner.

Component 1: The source of the research and of the funding

The study was a senior thesis project conducted by a psychology major. Presumably, it was cheap to run and was paid for by the student. One could argue that she would have a reason to want the results to come out as they did, although with a properly conducted study, the motives of the experimenter should be minimized. As we shall see, there were additional problems with this study.

Component 2: The researchers who had contact with the participants

Presumably, only the student conducting the study had contact with the respondents. Crucial missing information is whether she told them the purpose of the study. Even if she did not tell them, many of the psychology majors may have known her and known what she was doing. Any clues as to desired outcomes on the part of experimenters can bias the results.

Component 3: The individuals or objects studied and how they were selected

The individuals selected are the crux of the problem here. The measurements were taken on advanced psychology and chemistry students, which would have been fine if they had been sampled correctly. However, only those who were in the psychology classes or in the chemistry labs that day were actually measured. Less conscientious students are more likely to leave early before a holiday, but a missed class is probably easier to make up than a missed lab. Therefore, perhaps a larger proportion of the students with low grade-point averages were absent from the psychology classes than from the chemistry labs. Due to the missing students, the investigator's results would overestimate the average GPA for psychology students more so than for chemistry students.

Component 4: The exact nature of the measurements made or questions asked

Students were asked to give a “self-report” of their grade-point averages. A more accurate method would have been to obtain this information from the registrar at the university. Students may not know their exact grade-point average. Also, one group may be more likely to know the exact value than the other. For example, if many of the chemistry majors were planning to apply to medical school in the near future, they may be only too aware of their grades. Further, the headline implies that GPA is a measure of intelligence. Finally, the research assumes that GPA is a standard measure. Perhaps grading is more competitive in the chemistry department.

Component 5: The setting in which the measurements were taken

Notice that the article specifies that the measurements were taken on the day before a major holiday. Unless the university consisted mainly of commuters, many students may have left early for the holiday, further aggravating the problem that the students with lower grades were more likely to be missing from the psychology classes than from the chemistry labs. Further, because students turned in their questionnaires anonymously, there was presumably no accountability for incorrect answers.

Component 6: The extraneous differences between groups being compared

It is difficult to know what differences might exist without knowing more about the particular university. For example, because psychology is such a popular major, at some universities students are required to have a certain GPA before they are admitted to the major. A university with a separate premedical major might have the best of the science students enrolled in that major instead of chemistry. Those kinds of extraneous factors would be relevant to interpreting the results of the study.

Component 7: The magnitude of any claimed effects or differences

The news report does present this information. Additional useful information would be to know how many students were included in each of the averages given, what percentage of all students in each major were represented in the sample, and how much variation there was among GPAs within each of the two groups.

Hypothetical News Article 3: “Researchers Find Drug to Cure Excessive Barking in Dogs”

Barking dogs can be a real problem, as anyone who has been kept wake at night by the barking of a neighbour’s canine companion will know. Researchers at a local university have tested a new drug that they hope will put all concerned to rest. Twenty dog owners responded to a newspaper article asking for volunteers with problem barking dogs to participate in a study. The dogs were randomly assigned to two groups. One group of dogs was given the drug, administered as a shot, and the other dogs were not. Both groups were kept overnight at the research facility and frequency of barking was observed. The researchers deliberately tried to provoke the dogs into barking by doing things like ringing the doorbell of the facility and having a mail carrier walk up to the door. The two groups were treated on separate weekends because the facility was only large enough to hold ten dogs. The researchers left a tape recorder running and measured the amount of time during which any barking was heard. The dogs who had been given the drug spent only half as much time barking as did the dogs in the control group.

Component 1: The source of the research and of the funding

We are not told why this study was conducted. Presumably it was because the researchers were interested in helping to solve a societal problem, but perhaps not. It is not uncommon for drug companies to fund research to test a new product or a new use for a current product. If that were the case, the researchers would have added incentive for the results to come out favourable to the drug. If everything were done correctly, such an incentive wouldn’t be a major factor; however, when research is funded by a private source, that information should be announced when the results are announced.

Component 2: The researchers who had contact with the participants

We are not given any information about who actually had contact with the dogs. One important question is whether the same handlers were used with both groups of dogs. If not, the difference in handlers could explain the results. Further, we are not told whether the dogs were primarily left alone or were attended most of the time. If researchers were present most of the time, their behaviour toward the dogs could have had a major impact on the amount of barking.

Component 3: The individuals or objects studied and how they were selected

We are told that the study used dogs whose owners volunteered them as problem dogs for the study. Although the report does not mention payment, it is quite common for volunteers to receive monetary compensation for their participation. The volunteers presumably lived in the area of the university. The dog owners had to be willing to be separated from their pets for the weekend. These and other factors mean that the owners and dogs who participated may differ from the general population. Further, the initial reasons for the problem behaviour may vary from one participant to the next, yet the dogs were measured together. Therefore, there is no way to ascertain if, for example, dogs who bark only because they are lonely would be helped. In any case, we cannot extend the results of this study to conclude that the drug would work similarly on all dogs or even on all problem dogs. Because the dogs were randomly assigned to the two groups—and if there were no other problems—we would be able to extend the results to all dogs similar to those who participated.

Component 4: The exact nature of the measurements made or questions asked

The researchers measured each group of dogs as a group, by listening to a tape and recording the amount of time during which there was any barking. Because dogs are quite responsive to group behaviour, one barking dog could set the whole group barking for a long time. Therefore, just one particularly obnoxious dog in the control group alone could explain the results. It would have been better to separate the dogs and measure each one individually.

Component 5: The setting in which the measurements were taken

The groups were measured on separate weekends. This creates another problem. First, the researchers knew which group was which, and may have unconsciously provoked the control group slightly more than the group receiving the drug. Further, conditions differed over the 2 weeks. Perhaps it was sunny one weekend and raining the next, or there were other subtle differences, such as more traffic one weekend than the next, small planes overhead, and so on. All of these could change the behaviour of the dogs but might go unnoticed or unreported by the experimenters. The measurements were also taken outside of the dogs’ natural environments. The dogs in the experimental group in particular would have reason to be upset because they were first given a shot and then put together with nine other dogs in the research facility. It would have been better to put them back into their natural environment because that’s where the problem barking was known to occur.

Component 6: The extraneous differences between groups being compared

The dogs were randomly assigned to the two groups, which should have minimized overall differences in size, temperament, and so on. However, differences were induced between the two groups by the way the experiment was conducted. Recall that the two groups were measured on different weekends—this could have created the difference in behaviour. Also, the treated dogs were given a shot to administer the drug, whereas the control group was given no shot. It could be that the very act of

getting a shot made the drug group lethargic. A better design would have been to administer a placebo shot—that is, a shot with an inert substance—to the control group.

Component 7: The *magnitude* of any claimed effects or differences

We are told only that the treated group barked half as much as the control group. We are not told how much time either group spent barking. If one group barked 8 hours a day but the other group only 4 hours a day, that would not be a satisfactory solution to the problem of barking dogs.

2.4 PLANNING YOUR OWN STUDY: DEFINING THE COMPONENTS IN ADVANCE

Although you may never have to design your own survey or experiment, it will help you understand how difficult it can be if we illustrate the Seven Critical Components for a very simple hypothetical study you might want to conduct. Suppose you are interested in determining which of three local supermarkets has the best prices so you can decide where to shop. Because you obviously can't record and summarize the prices for all available items, you would have to use some sort of sample. To obtain meaningful data, you would need to make many decisions. Some of the Components need to be reworded because they are being answered in advance of the study, and obviously not all of the Components are relevant for this simple example. However, by going through them for such a simple case, you can see how many ambiguities and decisions can arise when designing a study.

Component 1: The *source* of the research and of the *funding*

Presumably you would be funding the study yourself, but before you start you need to decide why you are doing the study. Are you only interested in items you routinely buy, or are you interested in comparing the stores on the multitude of possible items?

Component 2: The *researchers* who had *contact* with the participants

In this example, the question would be who is going to visit the stores and record the prices.

Will you personally visit each store and record the prices? Will you send friends to two of the stores and visit the third yourself? If you use other people, you would need to train them so there would be no ambiguities.

Component 3: The *individuals* or objects studied and how they were *selected*

In this case, the “objects studied” are items in the grocery store. The correct question is, “On what items should prices be recorded?” Do you want to use exactly the same items at all stores? What if one store offers its own brand but another only offers name brands? Do you want to choose a representative sampling of items you are likely to buy or choose from all possible items? Do you want to include non-food items? How many items should you include? How should you choose which ones to select? If you are simply trying to minimize your own shopping bill, it is probably best to list the 20 or 30 items you buy most often. However, if you are interested in sharing your results with others, you might prefer to choose a representative sample of items from a long list of possibilities.

Component 4: The *exact nature* of the *measurements* made or *questions* asked

You may think that the cost of an item in a supermarket is a well-defined measurement. But if a store is having a sale on a particular item on your list, should you use the sale price or the regular price? Should you use the price of the smallest possible size of the product? The largest? What if a store always has a sale on one brand or another of something, such as laundry soap, and you don't really care which brand you buy? Should you then record the price of the brand on sale that week? Should you record the prices listed on the shelves or actually purchase the items and see if the prices listed were accurate?

Component 5: The *setting* in which the measurements were taken

When will you conduct the study? Supermarkets in university towns may offer sale prices on items typically bought by students at certain times of the year—for example, just after students have returned from vacation. Many stores also offer sale items related to certain holidays, such as ham or turkey just before Christmas or eggs just before Easter. Should you take that kind of timing into account?

Component 6: The *extraneous differences* between groups being compared

There should be no extraneous differences related to the direct costs of the items. However, if you were conducting the study in order to minimize your shopping costs, you might ask if there are hidden costs for shopping at one store versus another. For example, do you always have to wait in line at one store and not at another, and should you therefore put a value on your time? Does one store make mistakes at the cash register more often than another? Does one store charge a higher fee to use your cash card for payment? Does it cost more to drive to one store than another?

Component 7: The *magnitude* of any claimed effects or differences

This component should enter into your decision about where to shop after you have finished the study. Even if you find that one store costs less than another, the magnitude of the difference may not convince you to shop there. You would probably want to figure out approximately how much shopping in a particular store would save you over the course of a year. You can see why knowing the amount of a difference found in a study is an important component for using that study to make future decisions.

Chapter 3

Measurements, Mistakes, And Misunderstandings

3.1 SIMPLE MEASURES DON'T EXIST

In the last chapter, we listed seven critical components that need to be considered when someone conducts a study. You saw that many decisions need to be made and many potential problems can arise when you try to use data to answer a question. One of the hardest decisions is contained in Component 4—that is, in deciding exactly what to measure or what questions to ask. In this chapter, we focus on problems with defining measurements and on the subsequent misunderstandings and mistakes that can result. When you read the results of a study, it is important that you understand exactly how the information was collected and what was measured or asked. Consider something as apparently simple as trying to measure your own height. Try it a few times and see if you get the measurement to within a quarter of an inch from one time to the next. Now imagine trying to measure something much more complex, such as the amount of fat in someone's diet or the degree of happiness in someone's life. Researchers routinely attempt to measure these kinds of factors.

3.2 IT'S ALL IN THE WORDING

You may be surprised at how much answers to questions can change based on simple changes in wording. Here is one example. Loftus and Palmer (1974; quoted in Plous, 1993, p. 32) showed college students films of an automobile accident, after which they asked them a series of questions. One group was asked the question: "About how fast were the cars going when they contacted each other?" The average response was 31.8 miles per hour. Another group was asked: "About how fast were the cars going when they collided with each other?" In that group, the average response was 40.8 miles per hour. Simply changing from the word *contacted* to the word *collided* increased the estimates of speed by 9 miles per hour, or 28%, even though the respondents had witnessed the same film.

Many pitfalls can be encountered when asking questions in a survey or experiment.

Here are some of them; each will be discussed in turn:

1. Deliberate bias
2. Unintentional bias
3. Desire to please
4. Asking the uninformed
5. Unnecessary complexity
6. Ordering of questions
7. Confidentiality and anonymity

Deliberate Bias

Sometimes, if a survey is being conducted to support a certain cause, questions are deliberately worded in a biased manner. Be careful about survey questions that begin with phrases like "Do you agree that. . . ." Most people want to be agreeable and will be inclined to answer yes unless they have strong feelings the other way. For example, suppose an anti-abortion group and a pro-choice group each wanted to conduct

a survey in which they would find the best possible agreement with their position. Here are two questions that would each produce an estimate of the proportion of people who think abortion should be completely illegal. Each question is almost certain to produce a different estimate: 1. Do you agree that abortion, the murder of innocent beings, should be outlawed? 2. Do you agree that there are circumstances under which abortion should be legal, to protect the rights of the mother?

Appropriate wording should not indicate a desired answer. For instance, a Gallup Poll conducted in June 1998 contained the question "Do you think it was a good thing or a bad thing that the atomic bomb was developed?" Notice that the question does not indicate which answer is preferable. In fact 61% of the respondents said "bad," whereas 36% said "good" and 3% were undecided.

Unintentional Bias

Sometimes questions are worded in such a way that the meaning is misinterpreted by a large percentage of the respondents. For example, if you were to ask people whether they use drugs, you would need to specify if you mean prescription drugs, illegal drugs, over-the-counter drugs, or common substances such as caffeine. If you were to ask people to recall the most important date in their life, you would need to clarify if you meant the most important calendar date or the most important social engagement with a potential partner. (It is unlikely that anyone would mistake the question as being about the shrivelled fruit, but you can see that the same word can have multiple meanings.)

Desire to Please

Most survey respondents have a desire to please the person who is asking the question. They tend to understate their responses about undesirable social habits and opinions, and vice versa. For example, in recent years estimates of the prevalence of cigarette smoking based on surveys do not match those based on cigarette sales. Either people are not being completely truthful or lots of cigarettes are ending up in the garbage.

Asking the Uninformed

People do not like to admit that they don't know what you are talking about when you ask them a question. Crossen (1994, p. 24) gives an example: "When the American Jewish Committee studied Americans' attitudes toward various ethnic groups, almost 30% of the respondents had an opinion about the fictional Wisians, rating them in social standing above a half-dozen other real groups, including Mexicans, Vietnamese, and African blacks." Political pollsters, who are interested in surveying only those who will actually vote, learned long ago that it is useless to simply ask people if they plan to vote. Most of them will say yes. Instead, they ask questions to establish a history of voting, such as "Where did you go to vote in the last election?"

Unnecessary Complexity

If questions are to be understood, they must be kept simple. A question such as "Shouldn't former drug dealers not be allowed to work in hospitals after they are released from prison?" is sure to lead to confusion. Does a yes answer mean they should or they shouldn't be allowed to work in hospitals? It would take a few readings to figure that out. Another way in which a question can be unnecessarily complex is to actually ask more than one question at once. An example would be a question such as, "Do you support the president's health care plan because it would ensure that all Americans receive health coverage?" If you agree with the idea that all Americans should receive health coverage, but disagree with the remainder of the plan, do you answer yes or no? Or what if you support the President's plan, but not for that reason?

Ordering of Questions

If one question requires respondents to think about something that they may not have otherwise considered, then the order in which questions are presented can change the results. For example, suppose a survey were to ask, "To what extent do you think teenagers today worry about peer pressure related to drinking alcohol?" and then ask, "Name the top five pressures you think face teenagers today." It is quite likely that respondents would use the idea they had just been given and name peer pressure related to drinking alcohol as one of the five choices.

Confidentiality and Anonymity

People sometimes answer questions differently based on the degree to which they believe they are anonymous. Because researchers often need to perform follow-up surveys, it is easier to try to ensure confidentiality than true anonymity. In ensuring confidentiality, the researcher promises not to release identifying information about respondents. In a truly anonymous survey, the researcher does not know the identity of the respondents. Questions on issues such as sexual behaviour and income are particularly difficult because people consider those to be private matters. A variety of techniques have been developed to help ensure confidentiality, but surveys on such issues are hard to conduct accurately.

3.3 OPEN OR CLOSED QUESTIONS:

SHOULD CHOICES BE GIVEN?

An **open question** is one in which respondents are allowed to answer in their own words, whereas a **closed question** is one in which they are given a list of alternatives from which to choose their answer. Usually the latter form offers a choice of "other," in which the respondent is allowed to fill in the blank.

Problems with Closed Questions

To show the limitation of closed questions, Schuman and Scott (22 May 1987) asked about "the most important problem facing this country today." Half of the sample, 171 people, were given this as an open question. The most common responses were

Unemployment (17%)
General economic problems (17%)
Threat of nuclear war (12%)
Foreign affairs (10%)

In other words, one of these four choices was volunteered by over half of the respondents. The other half of the sample was given this as a closed question. Following is the list of choices and the percentage of respondents who chose them:

The energy shortage (5.6%)
The quality of public schools (32.0%)
Legalized abortion (8.4%)
Pollution (14.0%)

These four choices combined were mentioned by only 2.4% of respondents in the open-question survey; yet they were selected by 60% when they were the only specific choices given. Further, respondents in this closed-question survey were given an open choice. In addition to the list of four, they were told: "If you prefer, you may name a different problem as most important." On the basis of the closed-form questionnaire, policymakers would have been seriously misled about what is important to the public. It is possible to avoid this kind of astounding discrepancy. If closed questions are preferred, they first should be presented as open questions to a test sample before the real survey is conducted. Then the most common responses should be included in the list of choices for the closed question. This kind of exercise is usually done as part of what's called a "pilot survey," in which various aspects of a study design can be tried before it's too late to change them.

Problems with Open Questions

The biggest problem with open questions is that the results can be difficult to summarize. If a survey includes thousands of respondents, it can be a major chore to categorize their responses. Another problem, found by Schuman and Scott (22 May 1987), is that the wording of the question might unintentionally exclude answers that would have been appealing had they been included in a list of choices (such as in a closed question). To test this, they asked 347 people to "name one or two of the most important national or world event(s) or change(s) during the past 50 years." The most common choices and the percentage who mentioned them were

World War II (14.1%)
Assassination of John F. Kennedy (4.6%)
The Vietnam War (10.1%)
Don't know (10.6%)
All other responses (53.7%)

The same question was then repeated in closed form to a new group of 354 people. Five choices were given: the first four choices in the preceding list plus "invention of the computer." Of the 354 respondents, the percentage of those who selected each choice was:

World War II (22.9%)
Exploration of space (15.8%)
Assassination of John F. Kennedy (11.6%)
The Vietnam War (14.1%)
Invention of the computer (29.9%)
Don't know (0.3%)
All other responses (5.4%)

The most frequent response was "invention of the computer," which had been mentioned by only 1.4% of respondents in the open question. Clearly the wording of the question led respondents to focus on "events" rather than "changes," and the invention of the computer did not readily come to mind. When it was presented as an option, however, people realized that it was indeed one of the most important events or changes during the past 50 years. In summary, there are advantages and disadvantages to both approaches. One compromise is to ask a small test sample to list the first several answers that come to mind, and then use the most common of those in a closed-question survey. These choices could be supplemented with additional answers such as "invention of the computer," which may not readily come to mind. Remember that, as the reader, you have an important role in interpreting the results. You should always be informed as to whether questions were asked in open or closed form, and if the latter, you should be told what the choices were. You should also be told whether "don't know" or "no opinion" was offered as a choice in either case.

3.4 DEFINING WHAT IS BEING MEASURED

THE UNEMPLOYED

Ask people whether they know anyone who is unemployed; they will invariably say yes. But most people don't realize that in order to be officially unemployed, and included in the unemployment statistics given by the U.S. government, you must meet very stringent criteria. These criteria exclude "discouraged workers" who are "identified as persons who, though not working or seeking work, have a current desire for a job, have looked for one within the past year [but not within the past four weeks], and are currently available for work" (U.S. Dept. of Labour, September 1992, p. 10). If you know someone who fits that definition, you would undoubtedly think of that person as unemployed. However, he or she would not be included in the official statistics. You can see that the true number of people who are not working is higher than government statistics would lead you to believe. _

This example illustrates that when you read about measurements taken by someone else, you should not automatically assume you are speaking a common language. A precise definition of what is meant by "adolescent sexuality" or "unemployment" should be provided.

Some Concepts Are Hard to Define Precisely

Sometimes it is not the language but the concept itself that is ill-defined. For example, there is still not universal agreement on what should be measured with intelligence, or IQ, tests. The tests were originated at the beginning of the 20th century in order to determine the mental level of school children. The intelligence quotient (IQ) of a child was found by dividing the child's "mental level" by his or her chronological age. The "mental level" was determined by comparing the child's performance on the test with that of a large group of "normal" children, to find the age group the individual's performance matched. Thus, if an 8-year-old child performed as well on the test as a "normal" group of 10-year-old children, he or she would have an IQ of $100 \cdot (10/8) = 125$.

IQ tests have been expanded and refined since the early days, but they continue to be surrounded by controversy. One reason is that it is very difficult to define what is meant by intelligence. It is difficult to measure something if you can't even agree on what it is you are trying to measure. If you are interested in knowing more about these tests and the surrounding controversies, you can find numerous books on the subject. Anastasi and Urbina (1997) provide a detailed discussion of a large variety of psychological tests, including IQ tests.

Measuring Attitudes and Emotions

Similar problems exist with trying to measure attitudes and emotions such as self-esteem and happiness. The most common method for trying to measure such things is to have respondents read statements and determine the extent to which they agree with the statement. For example, a test for measuring happiness might ask respondents to indicate their level of agreement, from "strongly disagree" to "strongly agree," with statements such as "I generally feel optimistic when I get up in the morning." To produce agreement on what is meant by characteristics such as "introversion," psychologists have developed standardized tests that claim to measure those attributes.

3.5 DEFINING A COMMON LANGUAGE

So that we're all speaking a common language for the rest of this book, we need to define some terms. We can perform different manipulations on different types of data, so we need a common understanding of what those types are. Other terms defined in this section are those that are well-known in everyday usage but that have a slightly different technical meaning.

Categorical versus Measurement Variables

Thus far in this book, we have seen examples of measuring opinions (such as what you think is the most important problem facing society), numerical information (such as weight gain in infants), and attributes that can be transformed into numerical information (such as IQ). To understand what we can do with these measurements, we need definitions to distinguish numerical measures from qualitative ones. Although statisticians make numerous fine distinctions among types of measurements, for our purposes it will be sufficient to distinguish between just two main types: categorical variables and measurement variables.

Categorical Variables

Categorical variables are those we can place into a category but that may not have any logical ordering. For example, you could be categorized as male or female. You could also be categorized based on what you name as the most important problem facing society. Notice that we are limited in how we can manipulate this kind of information numerically. For example, we cannot talk about the average problem facing society in the same way as we can talk about the average weight gain of infants during the first few days of life.

Measurement Variables

Measurement variables are those for which we can record a numerical value and then order respondents according to those values. For example, IQ is a measurement variable because it can be expressed as a single number. An IQ of 130 is higher than an IQ of 100. Age, height, and number of cigarettes smoked per day are other examples of measurement variables. Notice that these can be worked with numerically. Of course, not all numerical summaries will make sense even with measurement variables. For example, if one person in your family smokes 20 cigarettes a day and the remaining three members smoke none, it is accurate but misleading to say that the average number of cigarettes smoked by your family per day is 5 per person. We will learn about reasonable numerical summaries in Chapter 7.

Continuous versus Discrete Measurement Variables

Even when we can measure something with a number, we may need to distinguish further whether it can fall on a continuum. A **discrete variable** is one for which you could actually count the possible responses. For example, if we measure the number of automobile accidents on a certain stretch of highway, the answer could be zero, one, two, three, and so on. It could not be 2 1/2 or 3.8. Conversely, a **continuous variable** can be anything within a given interval. Age, for example, falls on a continuum. Something of a grey area exists between these definitions. For example, if we measure age to the nearest year, it may seem as though it should be called a discrete variable. But the real difference is conceptual. With a discrete variable you can count the possible responses without having to round off. With a continuous variable you can't. In case you are confused by this, note that long ago you probably figured out the difference between the phrases "the number of" and "the amount of." You wouldn't say, "the amount of cigarettes smoked," nor would you say, "the number of water consumed." Discrete variables are analogous to numbers of things, and continuous variables are analogous to amounts. You still need to be careful about wording, however, because we have a tendency to express continuous variables in discrete units. Although you wouldn't say, "the number of water consumed," you might say, "the number of glasses of water consumed." That's why it's the *concept* of number versus amount that you need to think about.

Validity, Reliability, Bias, and Variability

The words we define in this section are commonly used in the English language, but they also have specific definitions when applied to measurements. Although these definitions are close to the general usage of the words, to avoid confusion we will spell them out.

Validity

When you talk about something being *valid*, you generally mean that it makes sense to you; it is sound and defensible. The same can be said for a measurement. A **valid measurement** is one that actually measures what it claims to measure. Thus, if you tried to measure happiness with an IQ test, you would not get a valid measure of happiness.

A more realistic example would be trying to determine the selling price of a home. Getting a valid measurement of the actual sales price of a home is tricky because the purchase often involves bargaining on what items are to be left behind by the old owners, what repairs will be made before the house is sold, and so on. These items can change the recorded sales price by thousands of dollars. If we were to define the "selling price" as the price recorded in public records, it may not actually reflect the price the buyer and seller had agreed was the true worth of the home.

To determine whether a measurement is valid, you need to know exactly what was measured. For example, many readers, once they are informed of the definition, do not think the unemployment figures provided by the U.S. government are a valid measure of unemployment, as the term is generally understood. Remember that the figures do not include "discouraged workers." However, the government statistics are a valid measure of the percentage of the "civilian labour force" that is currently "unemployed," according to the precise definitions supplied by the Bureau of Labour Statistics. The problem is that most people do not understand exactly what the government has measured.

Reliability

When we say something or someone is *reliable*, we mean that that thing or person can be depended upon time after time. A reliable car is one that will start every time and get us where we are going without worry. A reliable friend is one who is always there for us, not one who is sometimes too busy to bother with us. Similarly, a **reliable measurement** is one that will give you or anyone else approximately the same result time after time, when taken on the same object or individual. For example, a reliable way to define the selling price of a home would be the officially recorded amount. This may not be valid, but it would give us a consistent figure without any ambiguity.

Reliability is a useful concept in psychological and aptitude testing. An IQ test is obviously not much use if it measures the same person's IQ to be 80 one time and 130 the next. Whether we agree that the test is measuring what we really mean by "intelligence" (that is, whether it is really valid), it should at least be reliable enough to give us approximately the same number each time. Commonly used IQ tests are fairly reliable: About two-thirds of the time, taking the test a second time gives a reading within 2 or 3 points of the first test, and, most of the time, it gives a reading within about 5 points.

The most reliable measurements are physical ones taken with a precise measuring instrument. For example, it is much easier to get a reliable measurement of height than of happiness, assuming you have an accurate tape measure.

However, you should be cautious of measurements given with greater precision than you think the measuring tool would be capable of providing. The degree of precision probably exceeds the reliability of the measurement. For example, if your friend measures the width of a swimming pool with a ruler and reports that it is 15.771 feet wide, which is 15' 9 1/4", you should be suspicious. It would be very difficult to measure a distance that large reliably with a 12-inch ruler. A second measuring attempt would undoubtedly give a different number.

Bias

A systematic prejudice in one direction is called a *bias*. Similarly, a measurement that is systematically off the mark in the same direction is called a **biased measurement**. If you were trying to weigh yourself with a scale that was not satisfactorily adjusted at the factory, and was always a few pounds under, you would get a biased view of your own weight. When we used the term earlier in discussing the wording of questions, we noted that either intentional or unintentional bias could enter into the responses of a poorly worded survey question. Notice that a biased measurement differs from an unreliable measurement because it is consistently off the mark in the same direction.

Variability

If someone has *variable* moods, we mean that that person has unpredictable swings in mood. When we say the weather is quite variable, we mean it changes without any consistent pattern. Most measurements are prone to some degree of **variability**. By that, we mean that they include unpredictable errors or discrepancies that are not readily explained. If you tried to measure your height as instructed at the beginning of this chapter, you probably found some unexplainable variability from one time to the next. If you tried to measure the length of a table by laying a ruler end to end, you would undoubtedly get a slightly different answer each time.

Unlike the other terms we have defined, which are used to characterize a single measurement, variability is a concept used when we talk about two or more measurements in relation to each other. Sometimes two measurements vary because the measuring device produces unreliable results—for example, when we try to measure a large distance with a small ruler. Other times variability results from changes in the system being measured. For example, even with a very precise measuring device your recorded blood pressure will differ from one moment to the next.

Natural Variability

A related concept that of **natural variability**, is crucial to understanding modern statistical methods. When we measure the same quantity across several individuals, such as the weight gain of newborn babies, we are bound to get some variability. Although some of this may be due to our measuring instrument, most of it is simply due to the fact that everyone is different. Variability is simply inherent in nature. Babies all gain weight at their own pace. If we want to compare the weight gain of a group of babies who have consistently listened to a heartbeat to the weight gain of a group of babies who have not, we first need to know how much variability to expect due to natural causes.

We encountered the idea of natural variability when we discussed comparing resting pulse rates of men and women in Chapter 1. If there were no variability within each sex, it would be easy to detect a difference between males and females. The more variability there is within each group, the more difficult it is to detect a difference between groups.

In Part 4, we will learn how to sort out differences due to natural variability from differences due to features we can define, measure, and possibly manipulate, such as amount of salt consumed or time spent exercising. In this way, we can study the effects of diet or medical treatments on disease, of advertising campaigns on consumer choices, of exercise on weight loss, and so on. This one basic idea, comparing natural variability to the variability induced by different behaviours, interventions, or group memberships, forms the heart of modern statistics. It has allowed Salk to conclude that heartbeats are soothing to infants and the medical community to conclude that aspirin helps prevent heart attacks. We will see numerous other conclusions based on this idea throughout this book.

Chapter 4

How to Get a Good Sample

4.1 COMMON RESEARCH STRATEGIES

In Chapters 1 to 3, we discussed scientific studies in general, without differentiating them by type. In this chapter and the next, we are going to look at proper ways to conduct specific types of studies. When you read the results of a scientific study, the first thing you need to do is determine which research strategy was used. You can then see whether or not the study used the proper methods for that strategy. In this chapter and the next, you will learn about potential difficulties and outright disasters that can befall each type of study, as well as some principles for executing them correctly. First, let's examine the common types of research strategies.

Sample Surveys

You are probably quite familiar with sample surveys, at least in the form of political and opinion polls. In a **sample survey**, a subgroup of a large population is questioned on a set of topics. The results from the subgroup are used as if they were

representative of the larger population, which they will be if the sample was chosen correctly. There is no intervention or manipulation of the respondents in this type of research, they are simply asked to answer some questions. We examine sample surveys in more depth later in this chapter.

Experiments

An **experiment** measures the effect of manipulating the environment in some way. For example, the manipulation may include receiving a drug or medical treatment, going through a training program, agreeing to a special diet, and so on. Most experiments on humans use volunteers because you can't force someone to accept a manipulation. You then measure the result of the feature being manipulated, called the **explanatory variable**, on an outcome, called the **outcome variable**. Examples of outcome variables are cholesterol level (after taking a new drug), amount learned (after a new training program), or weight loss (after a special diet).

As an example, recall Case Study 1.2, an experiment that investigated the relationship between aspirin and heart attacks. The explanatory variable, manipulated by the researchers, was whether a participant took aspirin or a placebo. The variable was then used to help explain the outcome variable, which was whether a participant had a heart attack or not. Notice that the explanatory and outcome variables are both categorical in this case, with two categories each (aspirin/placebo and heart attack/no heart attack).

Experiments are important because, unlike most other studies, they often allow us to determine cause and effect. The participants in an experiment are usually randomly assigned to either receive the manipulation or take part in a control group. The purpose of the random assignment is to make the two groups approximately equal in all respects except for the explanatory variable, which is purposely manipulated. Differences in the outcome variable between the groups, if large enough to rule out natural chance variability, can then be attributed to the manipulation of the explanatory variable.

For example, suppose we flip a coin to assign each of a number of new babies into one of two groups. Without any intervention, we should expect both groups to gain about the same amount of weight, on average. If we then expose one group to the sound of a heartbeat and that group gains significantly more weight than the other group, we can be reasonably certain that the weight gain was due to the sound of the heartbeat.

Observational Studies

As we noted in Chapter 1, an **observational study** resembles an experiment except that the manipulation occurs naturally rather than being imposed by the experimenter. For example, we can observe what happens to people's weight when they quit smoking, but we can't experimentally manipulate them to quit smoking. We must rely on naturally occurring events.

This reliance on naturally occurring events leads to problems with establishing a causal connection because we can't arrange to have a similar control group. For instance, people who quit smoking may do so because they are on a "health kick" that also includes better eating habits, a change in coffee consumption, and so on. In this case, if we were to observe a weight loss (or gain) after cessation of smoking, we would not know if it were caused by the changes in diet or the lack of cigarettes.

In an observational study, you cannot assume that the explanatory variable of interest to the researchers is the only one that may be responsible for any observed differences in the outcome variable.

A special type of observational study is frequently used in medical research.

Called a **case-control study**, it is an attempt to include an appropriate control group.

In Chapter 5, we will explore the details of how these and other observational studies are conducted, and in Chapter 6, we will cover some examples in depth. Observational studies do have one advantage over experiments. Researchers are not required to induce artificial behaviour. Participants are simply observed doing what they would do naturally; therefore, the results can be more readily extended to the real world.

Meta-Analyses

A **meta-analysis** is a quantitative review of a collection of studies all done on a similar topic. Combining information from various researchers may result in the emergence of patterns or effects that weren't conclusively available from the individual studies.

It is becoming quite common for the results of meta-analyses to appear in newspapers and magazines. For example, the top headline in the November 24, 1993, *San Jose Mercury News* was, "Why mammogram advice keeps changing: S.F. study contradicts cancer society's finding." The article explained that, in addition to the results of new research in San Francisco, "A recent analysis of eight international studies did not find any clear benefit to women getting routine mammograms while in their 40s."

When you see wording indicating that many studies were analysed together, the report is undoubtedly referring to a meta-analysis. In this case, the eight studies in question had been conducted over a 30-year period with 500,000 women. Unfortunately, that information was missing from the newspaper article. Missing information is one of the problems with trying to evaluate a news article based on meta-analysis. In Chapter 24, we will examine meta-analyses and the role they play in science.

Case Studies

A **case study** is an in-depth examination of one or a small number of individuals. The researcher observes and interviews that individual and others who know about the topic of interest. For example, to study a purported psychic healer, a researcher might observe her at work, interview her about techniques, and interview clients who had been treated by the healer. We do not cover case studies in this book because they are descriptive and do not require statistical methods. We will issue one warning, though. Be careful not to assume you can extend the findings of a case study to any person or situation other than the one studied. In fact, case studies may be used to investigate situations precisely because they are rare and unrepresentative.

4.2 DEFINING A COMMON LANGUAGE

In the remainder of this chapter, we explore the methods used in sample surveys. To make our discussion of sampling methods clear, let's establish a common language. As we have seen before, statisticians borrow words from everyday language and attach specialized meaning to them. The first thing you need to know is that researchers sometimes speak synonymously of the individuals being measured and the measurements themselves. You can usually figure this out from the context. The relevant definitions cover both meanings.

- A **unit** is a single individual or object to be measured.
- The **population** (or **universe**) is the entire collection of units about which we would like information or the entire collection of measurements we would have if we could measure the whole population.
- The **sample** is the collection of units we actually measure or the collection of measurements we actually obtain.
- The **sampling frame** is a list of units from which the sample is chosen. Ideally, it includes the whole population.
- In a **sample survey**, measurements are taken on a subset, or sample, of units from the population.
- A **census** is a survey in which the entire population is measured.

4.3 THE BEAUTY OF SAMPLING

Here is some information that may astound you. If you use commonly accepted methods to sample 1500 adults from an entire population of millions of adults, you can almost certainly gauge, to within 3%, the percentage of the entire population who have a certain trait or opinion. (There is nothing magical about 1500 and 3%, as you will soon see.) Even more amazing is the fact that this result doesn't depend on how big the population is; it depends only on how many are in the sample. Our sample of 1500 would do equally well at estimating, to within 3%, the percentage of a population of 10 billion. Of course, you have to use a proper sampling method—but we address that later.

You can see why researchers are content to rely on public opinion polls rather than trying to ask everyone for their opinion. It is much cheaper to ask 1500 people than several million, especially when you can get an answer that is almost as accurate. It also takes less time to conduct a sample survey than a census, and because fewer interviewers are needed, there is better quality control.

Accuracy of a Sample Survey: Margin of Error

Most sample surveys are used to estimate the proportion or percentage of people who have a certain trait or opinion. For example, the Nielsen ratings, used to determine the percentage of American television sets tuned to a particular show, are based on a sample of a few thousand households. Newspapers and magazines routinely conduct surveys of a few thousand people to determine public opinion on current topics of interest. As we have said, these surveys, if properly conducted, are amazingly accurate. The measure of accuracy is a number called the **margin of error**. The sample proportion differs from the population proportion by more than the margin of error less than 5% of the time, or in fewer than 1 in 20 surveys. To express results in terms of percentages instead of proportions, simply multiply everything by 100.

As a general rule, the amount by which the proportion obtained from the sample will differ from the true population proportion rarely exceeds 1 divided by the square root of the number in the sample. This is expressed by the simple formula $\frac{1}{\sqrt{n}}$, where the letter n represents the number of people in the sample.

For example, with a sample of 1600 people, we usually get an estimate that is accurate to within $1/40 = 0.025 = 2.5\%$ of the truth because the square root of 1600 is 40. You might see results such as, “Fifty-five percent of respondents support the president’s economic plan. The margin of error for this survey is plus or minus 2.5 percentage points.” This means that it is almost certain that between 52.5% and 57.5% of the entire population support the plan. In other words, add and subtract the margin of error to the sample value, and the resulting interval almost surely covers the true population value. If you were to follow this method every time you read the results of a properly conducted survey, the interval would only miss covering the truth about 1 in 20 times.

Other Advantages of Sample Surveys

When a Census Isn’t Possible

Suppose you needed a laboratory test to see if your blood had too high a concentration of a certain substance. Would you prefer that the lab measure the entire population of your blood, or would you prefer to give a sample? Similarly, suppose a manufacturer of firecrackers wanted to know what percentage of their products were duds. They would not make much of a profit if they tested them all, but they could get a reasonable estimate of the desired percentage by testing a properly selected sample. As these examples illustrate, there are situations where measurements destroy the units being tested and thus a census is not feasible.

Speed

Another advantage of a sample survey over a census is the amount of time it takes to conduct a sample survey. For example, it takes several years to successfully plan and execute a census of the entire population of the United States. Getting monthly unemployment rates would be impossible with a census; the results would be quite out of date by the time they were released. It is much faster to collect a sample than a census if the population is large.

Accuracy

A final advantage of a sample survey is that you can devote your resources to getting the most accurate information possible from the sample you have selected. It is easier to train a small group of interviewers than a large one, and it is easier to track down a small group of no respondents than the larger one that would inevitably result from trying to conduct a census.

4.4 SIMPLE RANDOM SAMPLING

The ability of a relatively small sample to accurately reflect the opinions of a huge population does not happen haphazardly. It works only if proper sampling methods are used. Everyone in the population must have a specified chance of making it into the sample. Methods with this characteristic are called **probability sampling plans**.

The simplest way of accomplishing this goal is to use a **simple random sample**. With a simple random sample, every conceivable group of people of the required size has the same chance of being the selected sample. To actually produce a simple random sample, you need only two things. First, you need a list of the units in the population. Second, you need a source of **random numbers**. Random numbers can be found in tables designed for that purpose, called “tables of random digits,” or they can be generated by computers and calculators. If the population isn’t too large, physical methods can be used, as illustrated in the next hypothetical example.

HOW TO SAMPLE FROM YOUR CLASS

Suppose you are taking a class with 200 students and are unhappy with the teaching method. To substantiate that a problem exists so that you can complain to higher powers, you decide to collect a simple random sample of 25 students and ask them for their opinions. Notice that a sample of this size would have a margin of error of about 20% because $\frac{1}{\sqrt{25}} = \frac{1}{5} = 0.20$. Thus, the percentage of those 25 people who were dissatisfied would almost surely be within 20% of the percentage of the entire class who were dissatisfied. If 60% of the sample said they were dissatisfied, you could tell the higher powers that somewhere between 40% and 80% of the entire class was probably dissatisfied. Although that’s not a very precise statement, it is certainly enough to show major dissatisfaction. To collect your sample, you would proceed as follows:

STEP 1: Obtain a list of the students in the class, numbered from 1 to 200.

STEP 2: Obtain 25 random numbers between 1 and 200. One simple way to do this would be to write each of the numbers from 1 to 200 on equally sized slips of paper, put them in a bag, mix them very well, and draw out 25 or use a calculator or computer to generate 25 random numbers between 1 and 200.

31 141 35 69 100 182 61 116 191 161 129 120 150 15 84 194 135 101 44 163 152 39 99 110 36

STEP 3: Locate and interview the people on your list whose numbers were selected. Notice that it is important to try to locate the actual 25 people resulting from this process. If you tried to phone someone only once and gave up when you could not reach that person, you would bias your results toward people who were home more often. If you collected your sample correctly, as described, you would have legitimate data to present to the higher powers.

4.5 OTHER SAMPLING METHODS

By now you may be asking yourself how polling organizations could possibly get a numbered list of all voters or of all adults in the country. In truth, they don't. Instead, they rely on more complicated sampling methods. Here we describe a few other sampling methods, all of which are good substitutes for simple random sampling in most situations. In fact, they often have advantages over simple random sampling.

Stratified Random Sampling

Sometimes the population of units falls into natural groups, called **strata**. For example, public opinion pollsters often take separate samples from each region of the country so they can spot regional differences as well as measure national trends. Political pollsters may sample separately from each political party to compare opinions by party.

A **stratified random sample** is collected by first dividing the population of units into groups (strata) and then taking a simple random sample from each. For example, the strata might be regions of the country or political parties. You can often recognize this type of sampling when you read the results of a survey because the results will be listed separately for each of the strata. Stratified sampling has other advantages besides the fact that results are available separately by strata. One is that different interviewers may work best with different people. For example, people from separate regions of the country (South, Northeast, and so on) may feel more comfortable with interviewers from the same region. It may also be more convenient to stratify before sampling. If we were interested in opinions of college students across the country, it would probably be easier to train interviewers at each college rather than to send the same interviewer to all campuses.

So far we have been focusing on the collection of categorical variables, such as opinions or traits people might have. Surveys are also used to collect measurement variables, such as age at first intercourse or number of cigarettes smoked per day. We are often interested in the population average for such measurements. The accuracy with which we can estimate the average depends on the natural variability among the measurements. The less variable they are, the more precisely we can assess the population average on the basis of the sample values. For instance, if everyone in a relatively large sample reports that his or her age at first intercourse was between 16 years 3 months and 16 years 4 months, then we can be relatively sure that the average age in the population is close to that. However, if reported ages range from 13 years to 25 years, then we cannot pinpoint the average age for the population nearly as accurately.

Stratified sampling can help to solve the problem of great natural variability. Suppose we could figure out how to stratify in a way that allowed little natural variability in the answers within each strata. We could then get an accurate estimate for each stratum and combine estimates to get a much more precise answer for the group than if we measured everyone together. For example, if we wanted to estimate the average weight gain of newborn babies during the first four days of life, we could do so more accurately by dividing the babies into groups based on their initial birth weight. Very heavy newborns actually tend to

lose weight during the first few days, whereas very light ones tend to gain more weight.

Stratified sampling is sometimes used instead of simple random sampling for the following reasons:

1. We can find individual estimates for each stratum.
2. If the variable measured gives more consistent values within each of the strata than within the whole population, we can get more accurate estimates of the population values.
3. If strata are geographically separated, it may be cheaper to sample them separately.
4. We may want to use different interviewers within each of the strata.

Cluster Sampling

Cluster sampling is often confused with stratified sampling, but it is actually a radically different concept and can be much easier to accomplish. The population units are again divided into groups, called **clusters**, but rather than sampling within each group, we select a random sample of clusters and measure only those clusters. One obvious advantage of cluster sampling is that you need only a list of clusters, instead of a list of all individual units.

For example, suppose we wanted to sample students living in the dormitories at a college. If the college had 30 dorms and each dorm had 6 floors, we could consider the 180 floors to be 180 clusters of units. We could then randomly select the desired

number of floors and measure everyone on those floors. Doing so would probably be much cheaper and more convenient than obtaining a simple random sample of all dormitory residents.

If cluster sampling is used, the analysis must proceed differently because similarities may exist among the members of the clusters, and these must be taken into account.

Systematic Sampling

Suppose you had a list of 5000 names and telephone numbers from which you wanted to select a sample of 100. That means you would want to select 1 of every 50 people on the list. The first idea that might occur to you is to simply choose every 50th name on the list. If you did so, you would be using a **systematic sampling plan**. With this plan, you divide the list into as many consecutive segments as you need, randomly choose a starting point in the first segment, then sample at that same point in each segment. In our example, you would randomly choose a starting point in the first 50 names, then sample every 50th name after that. When you were finished, you would have selected one person from each of 100 segments, equally spaced throughout the list.

Systematic sampling is often a good alternative to random sampling. In a few instances, however, it can lead to a biased sample, and common sense must be used to avoid those. As an example, suppose you were doing a survey of potential noise problems in a high-rise college dormitory. Further, suppose a list of residents was provided, arranged by room number, with 20 rooms per floor and two people per room. If you were to take a systematic sample of, say, every 40th person on the list, you would get people who lived in the same location on every floor—and thus a biased sampling of opinions about noise problems.

4.6 DIFFICULTIES AND DISASTERS IN SAMPLING

In theory, designing a good sampling plan is easy and straightforward. However, the real world rarely cooperates with well-designed plans, and trying to collect a proper sample is no exception. Difficulties that can occur in practice need to be considered when you evaluate a study. If a proper sampling plan is never implemented, the conclusions can be misleading and inaccurate.

Difficulties

1. Using the wrong sampling frame
2. Not reaching the individuals selected
3. Getting no response or getting a volunteer response

Disasters

4. Getting a volunteer sample
5. Using a convenience or haphazard sample

Difficulties in Sampling

Following are some problems that can occur even when a sampling plan has been well designed.

Using the Wrong Sampling Frame

Remember that the sampling frame is the list of the population of units from which the sample is drawn. Sometimes a sampling frame will either include unwanted units or exclude desired units. For example, using a list of registered voters to predict election outcomes includes those who are not likely to vote as well as those who are likely to do so. Using a telephone directory to survey the general population excludes those who move often, those with unlisted home numbers (such as many physicians and teachers), and those who cannot afford a telephone.

Common sense can often lead to a solution for this problem. In the example of registered voters, interviewers may try to first ascertain the voting history of the person contacted by asking where he or she votes and then continuing the interview only if the person knows the answer. Instead of using a telephone directory, surveys use random digit dialling. This solution still excludes those without phones but not those who didn't happen to be in the last printed directory.

Not Reaching the Individuals Selected

Even if a proper sample of units is selected, the units may not be reached. For example, *Consumer Reports* magazine mails a lengthy survey to its subscribers to obtain information on the reliability of various products. If you were to receive such a survey, and you had a close friend who had been having trouble with a highly rated automobile, you may very well decide to pass the questionnaire on to your friend to answer. That way, he would get to register his complaints about the car, but

Consumer

Reports would not have reached the intended recipient.

Telephone surveys tend to reach a disproportionate number of women because they are more likely to answer the phone. To try to counter that problem, researchers sometimes ask to speak to the oldest adult male at home. Surveys are also likely to have trouble contacting people who work long hours and are rarely home or those who tend to travel extensively. In recent years, news organizations have been pressured to produce surveys of public opinion quickly. When a controversial story breaks, people want to know how others feel about it. This pressure results in what *Wall Street Journal* reporter Cynthia Crossen calls “quickie polls.” As she notes, these are “most likely to be wrong because questions are hastily drawn and poorly pretested, and it is almost impossible to get a random sample in one night” (Crossen, 1994, p. 102). Even with the computer randomly generating phone numbers for the sample, many people are not likely to be home that night—and they may have different opinions from those who are likely to be home. Most responsible reports about polls include information about the dates during which they were conducted. If a poll was done in one night, beware! It is important that once a sample has been selected, those individuals are the ones who are actually measured. It is better to put resources into getting a smaller sample than to get one that has been biased because the survey takers moved on to the next person on the list when a selected individual was initially unavailable.

Getting No Response or a Volunteer Response

Even the best surveys are not able to contact everyone on their list, and not everyone contacted will respond. The General Social Survey (GSS), run by the prestigious National Opinion Research Centre (NORC) at the University of Chicago, noted in its September 1993 *GSS News*:

In 1993 the GSS achieved its highest response rate ever, 82.4%. This is five percentage points higher than our average over the last four years. Given the long length of the GSS (90 minutes), the high response rates of the GSS are testimony to the extraordinary skill and dedication of the NORC field staff.

Beyond having a dedicated staff, not much can be done about getting everyone in the sample to respond. Response rates should simply be reported in research summaries. As a reader, remember that the lower the response rate, the less the results can be generalized to the population as a whole. Responding to a survey (or not) is voluntary, and those who respond are likely to have stronger opinions than those who do not. With mail surveys, it may be possible to compare those who respond immediately with those who need a second prodding, and in telephone surveys you could compare those who are home on the first try with those who require numerous call-backs.

If those groups differ on the measurement of interest, then those who were never reached are probably different as well. In a mail survey, it is best not to rely solely on “volunteer response.” In other words, don’t just accept that those who did not respond the first time can’t be cajoled into it. Often, sending a reminder with a brightly coloured stamp or following up with a personal phone call will produce the desired effect. Surveys that simply use those who respond voluntarily are sure to be biased in favour of those with strong opinions or with time on their hands.

Disasters in Sampling

A few sampling methods are so bad that they don’t even warrant a further look at the study or its results.

Getting a Volunteer Sample

Although relying on volunteer *responses* presents somewhat of a difficulty in determining the extent to which surveys can be generalized, relying on a volunteer *sample* is a complete waste of time. If a magazine or television station runs a survey and asks any readers or viewers who are interested to respond, the results reflect only the opinions of those who decide to volunteer. As noted earlier, those who have a strong opinion about the question are more likely to respond than those who do not. Thus, the responding group is simply not representative of any larger group. Most media outlets now acknowledge that such polls are “unscientific” when they report the results, but most readers are not likely to understand how misleading the results can be. The next example illustrates the contradiction that can result between a scientific poll and one relying solely on a volunteer sample.

Using a Convenience or Haphazard Sample

Another worthless sampling technique for surveys is to use the most convenient group available or to decide on the spot who to sample. Again, the group does not represent any larger population.

You have seen the proper way to collect a sample and have been warned about the many difficulties and dangers inherent in the process. We finish the chapter with a famous example that helped researchers learn some of these pitfalls.

Chapter 5

Experiments and Observational Studies

5.1 DEFINING A COMMON LANGUAGE

In this chapter, we focus on studies that attempt to detect relationships between variables. Connections others have found, which we examine in this chapter and the next, include a relationship between baldness and heart attacks in men, between smoking during pregnancy and subsequent lower IQ in the child, between listening to Mozart and scoring higher on an IQ test, and between handedness and age at death. We will see that some of these connections are supported by properly conducted studies, whereas other connections are not as solid.

Explanatory Variables, Response Variables, and Treatments

In most studies, we imagine that if there is a causal relationship, it occurs in a particular direction. For example, if we found that left-handed people die at a younger age than right-handed people, we could envision reasons why their handedness might be responsible for the earlier death, such as accidents resulting from living in a right-handed world. It would be more difficult to argue that they were left-handed because they were going to die at an earlier age.

Explanatory Variables versus Response Variables

We define an **explanatory variable** to be one that attempts to explain or is purported to cause (at least partially) differences in a **response variable** (sometimes called an **outcome variable**). In the previous example, handedness would be the explanatory variable and age at death the response variable. In the Salk experiment described in Chapter 1, whether the baby listened to a heartbeat was the explanatory variable and weight gain was the response variable. In a study comparing chemotherapy to surgery for cancer, the medical treatment is the explanatory variable and surviving (usually measured as surviving for five years) or not surviving is the response variable.

Many studies have more than one explanatory variable, but usually there is only one response variable.

We can usually distinguish which is which, but occasionally we simply examine relationships in which there is no conceivable causal connection. An example is the apparent relationship between baldness and heart attacks. Because the level of baldness was measured at the time of the heart attack, the heart attack could not have caused the baldness. It would be farfetched to assume that baldness results in such stress that men are led to have heart attacks. Instead, a third variable may be causing both the baldness and the heart attack. In such cases, we simply refer to the variables generically and do not assign one to be the explanatory variable and one to be the response variable.

Treatments

Sometimes the explanatory variable takes the form of a manipulation applied by the experimenter, such as when Salk played the sound of a heartbeat for some of the babies. A **treatment** is one or a combination of explanatory variables assigned by the experimenter. Notice that the term incorporates a collection of conditions. In Salk's experiment there were two conditions: Some babies received the heartbeat treatment and others received the silent treatment.

Experiment versus Observational Study

Ideally, if we were trying to ascertain the connection between the explanatory and response variables, we would keep everything constant except the explanatory variable. We would then manipulate the explanatory variable and notice what happened to the response variable as a consequence. We rarely reach this ideal, but we can come closer with an experiment than with an observational study.

In an **experiment**, we create differences in the explanatory variable and then examine the results. In an **observational study** we observe differences in the explanatory variable and then notice whether these are related to differences in the response variable.

For example, suppose we wanted to detect the effects of the explanatory variable "smoking during pregnancy" on the response variable "child's IQ at 4 years of age." In an experiment, we would randomly assign half of the mothers to smoke during pregnancy and the other half to not smoke. In an observational study, we would merely record smoking behaviour. This example demonstrates why we can't always perform an experiment.

Two reasons why we must sometimes use an observational study instead of an experiment:

1. It is unethical or impossible to assign people to receive a specific treatment.
2. Certain explanatory variables, such as handedness, are inherent traits and cannot be randomly assigned.

Confounders and Interactions

Confounding Variables

A **confounding variable** is one whose effect on the response variable cannot be separated from the effect of the explanatory variable. For instance, if we notice that women who smoke during pregnancy have children with lower IQs than the children of women who don't smoke, it could be because women who smoke also have poor nutrition. In that case, nutrition would be confounded with smoking in terms of its influence on subsequent IQ of the child. Confounding variables are a bigger problem in observational studies than in experiments. In fact, one of the major advantages of an experiment over an observational study is that, in an experiment, the researcher attempts to control for confounding variables.

Interactions between Variables

When you read the results of a study, you should also be aware that there may be **interactions** between explanatory variables. An interaction occurs when the effect of one explanatory variable on the response variable depends on what's happening with another explanatory variable. For example, if smoking during pregnancy reduces IQ when the mother does not exercise, but raises or does not influence IQ when the mother does exercise, then we would say that smoking interacts with exercise to produce an effect on IQ. Notice that if two variables do interact, it is important that the results be given separately for each combination. To simply say that smoking lowers IQ when, in fact, it only did so for those who didn't exercise would be a misleading conclusion.

Experimental Units, Subjects, and Volunteers

In addition to humans, it is common for studies to be performed on plants, animals, machine parts, and so on. To have a generic term for this conglomeration of possibilities, we define **experimental units** to be the smallest basic objects to which we can assign different treatments in an experiment and **observational units** to be the objects or people measured in any study.

The terms **participants** or **subjects** are commonly used when the observational units are people. In most cases, the participants in studies are **volunteers**. Sometimes they are passive volunteers, such as when all patients treated at a particular medical facility are asked to sign a consent form agreeing to participate in a study. Often researchers recruit volunteers through the newspaper. For example, a weekly newspaper in a small town in California recently ran an article with the headline, "Volunteers sought for silicone study" (*Winters (CA) Express*, 16 December 1993, p. 8). The article explained that researchers at a local medical school were "seeking 100 women with silicone breast implants and 100 without who are willing to fast overnight and then give a blood sample." The article also explained who was doing the research and its purpose. Notice that by recruiting volunteers for studies, the results cannot necessarily be extended to the larger population. For example, if volunteers are enticed to participate by receiving a small payment or free medical care, as is often the case, those who respond are more likely to be from lower socioeconomic backgrounds. Common sense should enable you to figure out if this is likely to be a problem, but researchers should always report the source of their participants so you can judge this for yourself.

5.2 DESIGNING A GOOD EXPERIMENT

Designing a flawless experiment is extremely difficult, and carrying one out is probably impossible. Nonetheless, there are ideals to strive for, and in this section, we investigate those first. We then explore some of the pitfalls that are still quite prevalent in research today.

Randomization: The Crucial Element

Experiments are supposed to reduce the effects of confounding variables and other sources of bias that are necessarily present in observational studies. They do so by using a simple principle called **randomization**. We discussed the idea of randomization in Chapter 4, when we described how to choose a sample for a survey. There, we were concerned that everyone in the population had a specified probability of making it into the sample. In experimentation, we are concerned that each of the experimental units (people, animals, and so on) has a specified probability of receiving any of the potential treatments. For example,

Salk should have ensured that each group of babies available for study had an equal chance of being assigned to hear the heartbeat or to go into the silent nursery. Otherwise, he could have chosen the babies who looked healthier to begin with to receive the heartbeat treatment.

Randomizing the Order of Treatments

In some experiments, all treatments are applied to each unit. In that case, randomization should be used to determine the *order* in which they are applied. For example, suppose an experiment is conducted to determine the extent to which drinking alcohol or smoking marijuana impairs driving ability. Because drivers are all so different, it makes sense to test the same drivers under all three conditions (alcohol, marijuana, and sober) rather than using different drivers for each condition. But if everyone were tested under alcohol, then marijuana, then sober, by the time they were traveling the course for the second and third times, their performance would improve just from having learned something about the course. A better method would be to randomly assign some drivers to each of the possible orderings so the learning effect would average out over the three treatments. Notice that it is important that the assignments be made *randomly*. If we let the experimenter decide which drivers to assign to which ordering, assignments could be made to give an unfair advantage to one of the treatments.

Randomizing the Type of Treatments

Randomly assigning the treatments to the experimental units also helps protect against hidden or unknown biases. For example, suppose that in the experiment in Case Study 1.2, approximately the first 11,000 physicians who enrolled were given aspirin and the remainder were given placebos. It could be that the healthier, more energetic physicians enrolled first, thus giving aspirin an unfair advantage. In statistics, “random” is not synonymous with “haphazard”—despite what your thesaurus might say. Random assignments may not be possible or ethical under some circumstances. But in situations where randomization is feasible, it is usually not difficult to accomplish. It can easily be done with a table of random digits, a computer, or even—if done carefully—by physical means such as flipping a coin or drawing numbers from a hat. The important feature, ensured by proper randomization, is that the chances of being assigned to each condition are the same for each participant.

Control Groups, Placebos, and Blinding

Control Groups

To determine whether a drug, heartbeat sound, meditation technique, and so on, has an effect, we need to know what would have happened to the response variable if the treatment had not been applied. To find that out, experimenters create **control groups**, which are treated identically in all respects, except that they don’t receive the active treatment.

Placebos

A special kind of control group is usually used in studies of the effectiveness of drugs. A substantial body of research shows that people respond not only to active drugs but also to **placebos**. A placebo looks like the real drug but has no active ingredients. Placebos can be amazingly effective; studies have shown that they can help up to 62% of headache sufferers, 58% of those suffering from seasickness, and 39% of those with postoperative wound pain.

Because the **placebo effect** is so strong, drug research is generally conducted by randomly assigning half of the volunteers to receive the drug and the other half to receive a placebo, without telling them which they are receiving. The placebo looks just like the real thing, so the participants will not be able to distinguish between it and the actual drug and thus will not be influenced by belief biases.

Blinding

The patient isn’t the only one who can be affected by knowing whether he or she has received an active drug. If the researcher who is measuring the reaction of the patients were to know which group was which, the researcher might take the measurements in a biased fashion.

To avoid these biases, good experiments use **double-blind** procedures. A double-blind experiment is one in which neither the participant nor the researcher taking the measurements knows who had which treatment. A **single-blind** experiment is one in which only the participants do not know which treatment they have been assigned. An experiment would also be called single-blind if the participants knew the treatments but the researcher were kept blind.

Although double-blind experiments are preferable, they are not always possible. For example, in testing the effect of meditation on blood pressure, the subjects would obviously know if they were in the meditation group or the control group. In this case, the only possible experiment would be a single-blind study in which the person taking the blood pressure measurement would not know who was in which group.

Pairing and Blocking

It is sometimes easier and more efficient to have each person in a study serve as his or her own control. That way, differences that inherently exist among individuals don’t obscure the treatment effects. We encountered this idea when we discussed how to compare driving ability when under the influence of alcohol and marijuana and when sober.

Sometimes, instead of using the same individual for the treatments, researchers will match people on traits that are likely to be related to the outcome, such as age, IQ, or weight. They then randomly assign each of the treatments to one member of each matched pair or grouping. For example, in a study comparing chemotherapy to surgery to treat cancer, patients might be matched by sex, age, and level of severity of the illness. One from each pair would then be randomly chosen to receive the chemotherapy and the other to receive surgery. (Of course, such a study would only be ethically feasible if there were no prior knowledge that one treatment was superior to the other. Patients in such cases are always required to sign an informed consent.)

Matched-Pair Designs

Experimental designs that use either two matched individuals or the same individual to receive each of two treatments are called **matched-pair designs**. The important feature of these designs is that randomization is used to assign the order of the two treatments. Of course, it is still important to try to conduct the experiment in a double-blind fashion so that neither the participant nor the researcher knows which order was used.

Block Designs

An extension of the matched-pair design to three or more treatments is called a **block design**. The method discussed for comparing drivers under three conditions was a block design. Each driver is called a **block**. This somewhat peculiar terminology results from the fact that these ideas were first used in agricultural experiments, where the experimental units were plots of land that had been subdivided into “blocks.” In the social sciences, designs such as these, in which the same participants are measured repeatedly, are referred to as **repeated-measures designs**.

Case Study 5.1

Quitting Smoking with Nicotine Patches

There is no longer any doubt that smoking cigarettes is hazardous to your health and to those around you. Yet, for someone addicted to smoking, quitting is no simple matter. One promising technique for helping people to quit smoking is to apply a patch to the skin that dispenses nicotine into the blood. These “nicotine patches” have become one of the most frequently prescribed medications in the United States. To test the effectiveness of these patches on the cessation of smoking, Dr. Richard Hurt and his colleagues (Hurt et al., 23 February 1994) recruited 240 smokers at Mayo Clinics in Rochester, Minnesota; Jacksonville, Florida; and Scottsdale, Arizona. Volunteers were required to be between the ages of 20 and 65, have an expired carbon monoxide level of 10 ppm or greater (showing that they were indeed smokers), be in good health, have a history of smoking at least 20 cigarettes per day for the past year, and be motivated to quit. Volunteers were randomly assigned to receive either 22-mg nicotine patches or placebo patches for 8 weeks. They were also provided with an intervention program recommended by the National Cancer Institute, in which they received counselling before, during, and for many months after the 8-week period of wearing the patches. After the 8-week period of patch use, almost half (46%) of the nicotine group had quit smoking, whereas only one-fifth (20%) of the placebo group had. Having quit was defined as “self-reported abstinence (not even a puff) since the last visit and an expired air carbon monoxide level of 8 ppm or less” (p. 596). After a year, rates in both groups had declined, but the group that had received the nicotine patch still had a higher percentage who had successfully quit than did the placebo group: 27.5% versus 14.2%. The study was double-blind, so neither the participants nor the nurses taking the measurements knew who had received the active nicotine patches. The study was funded by a grant from Lederle Laboratories and was published in the *Journal of the American Medical Association*. _

5.3 DIFFICULTIES AND DISASTERS IN EXPERIMENTS

We have already introduced some of the problems that can be encountered with experiments, such as biases introduced by lack of randomization. However, many of the complications that result from poorly conducted experiments can be negated with proper planning and execution.

Here are some potential complications:

1. Confounding variables
2. Interacting variables
3. Placebo, Hawthorne, and experimenter effects
4. Ecological validity and generalizability

Confounding Variables

The Problem

Variables that are connected with the explanatory variable can distort the results of an experiment because they—and not the explanatory variable—may be the agent actually causing a change in the response variable.

The Solution

Randomization is the solution. If experimental units are randomly assigned to treatments, then the effects of the confounding variables should apply equally to each treatment. Thus, observed differences between treatments should not be attributable to the confounding variables.

NICOTINE PATCH THERAPY

The nicotine patch therapy in Case Study 5.1 was more effective when there were no other smokers in the participant's home. Suppose the researchers had assigned the first 120 volunteers to the placebo group and the last 120 to the nicotine group. Further, suppose that those with no other smokers at home were more eager to volunteer. Then the treatment would have been confounded with whether there were other smokers at home. The observed results showing that the active patches were more effective than the placebo patches could have merely represented a difference between those with other smokers at home and those without. By using randomization, approximately equal numbers in each group should have come from homes with other smokers. Thus, any impact of that variable would be spread equally across the two groups. _

Interacting Variables

The Problem

Sometimes a second variable interacts with the explanatory variable, but the results are reported without taking that interaction into account. The reader is then misled into thinking the treatment works equally well, no matter what the condition is for the second variable.

The Solution

Researchers should measure and report variables that may interact with the main explanatory variable(s).

OTHER SMOKERS AT HOME

In the study we discussed in Case Study 5.1, there was an interaction between the treatment and whether there were other smokers at home. The researchers measured and reported this interaction. After the 8-week patch therapy, the proportion of the nicotine group who had quit smoking was only 31% if there were other smokers at home, whereas it was 58% if there were not. In the placebo group, the proportions who had quit were the same whether there were other smokers at home or not. Therefore, it would be misleading to merely report that 46% of the nicotine recipients had quit, without also providing the information about the interaction.

Placebo, Hawthorne, and Experimenter Effects

The Problem

We have already discussed the strong effect that a placebo can have on experimental outcomes because the power of suggestion is somehow able to affect the result. A related idea is that participants in an experiment respond differently than they otherwise would, just because they *are* in the experiment. This is called the "Hawthorne effect" because it was first detected in 1924 during a study of factory workers at the Hawthorne, Illinois, plant of the Western Electric Company. (The phrase was not actually coined until much later; see French, 1953.)

Related to these effects are numerous ways in which the experimenter can bias the results. These include recording the data erroneously to match the desired outcome, treating subjects differently based on which condition they are receiving, or subtly letting the subjects know the desired outcome.

The Solution

As we have seen, most of these problems can be overcome by using double-blind designs and by including a placebo group or a control group that receives identical handling except for the active part of the treatment. Other problems, such as incorrect data recording, should be addressed by allowing data to be automatically entered into a computer as it is collected, if possible. Depending on the experiment, there may still be subtle ways in which experimenter effects can sneak into the results. You should be aware of these possibilities when you read the results of a study.

DULL RATS

In an experiment designed to test whether the expectations of the experimenter could really influence the results, Rosenthal and Fode (1963) deliberately conned

12 experimenters. They gave each one five rats that had been taught to run a maze. They told six of the experimenters that the rats had been bred to do well (that is, that they were "maze bright") and told the other six that their rats were "maze dull" and should not be expected to do well. Sure enough, the experimenters who had been told they had bright rats found learning rates far superior to those found by the experimenters who had been told they had dull rats. Hundreds of other studies have since confirmed the "experimenter effect."

Ecological Validity and Generalizability

The Problem

Suppose you wanted to compare three assertiveness training methods to see which was more effective in teaching people how to say no to unwanted requests on their time. Would it be realistic to give them the training, then measure the results by asking them to role-play in situations in which they would have to say no? Probably not because everyone involved would know it was only role-playing. The usual social pressures to say yes would not be as striking. This is an example of an experiment with little **ecological validity**. In other words, the variables have been removed from their natural setting and are measured in the laboratory or in some other artificial setting. Thus, the results do not accurately reflect the impact of the variables in the real world or in everyday life. A related problem is one we have already mentioned namely, if volunteers are used for a study, can the results be generalized to any larger group?

The Solution

There are no ideal solutions to these problems, other than trying to design experiments that can be performed in a natural setting with a random sample from the population of interest. In most experimental work, these idealistic solutions are impossible. A partial solution is to measure variables for which the volunteers might differ from the general population, such as income, age, or health, and then try to determine the extent to which those variables would make the results less general than desired. In any case, when you read the results of a study, you should question its ecological validity and its generalizability.

REAL SMOKERS WITH A DESIRE TO QUIT

The researchers in Case Study 5.1 did many things to help ensure ecological validity and generalizability. First, they used a standard intervention program available from and recommended by the National Cancer Institute instead of inventing their own so that other physicians could follow the same program. Next, they used participants at three different locations around the country, rather than in one community only, and they involved a wide range of ages (20 to 65). They included individuals who lived in households with other smokers as well as those who did not. Finally, they recorded numerous other variables (sex, race, education, marital status, psychological health, and so on) and checked to make sure these were not related to the response variable or the patch assignment.

Exercise Yourself to Sleep

SOURCE: King et al., 1 January 1997, pp. 32–37.

According to the *UC Davis Health Journal* (November–December 1997, p. 8), older adults constitute only 12% of the population but receive almost 40% of the sedatives prescribed. The purpose of this randomized experiment was to see if regular exercise could help reduce sleep difficulties in older adults. The 43 subjects were sedentary volunteers between the ages of 50 and 76, with moderate sleep problems but no heart disease. They were randomly assigned to either participate in a moderate community based exercise program four times a week for 16 weeks, or continue to be sedentary. For ethical reasons, the control group was admitted to the program when the experiment was complete. The results were striking. The exercise group fell asleep an average of 11 minutes faster and slept an average of 42 minutes longer than the control group. Note that this could not be a double-blind experiment because subjects obviously knew whether they were exercising. Because sleep patterns were self-reported, there could have been a tendency to err in reporting, in the direction desired by the experimenters. However, this experiment is an example of a well-designed experiment, given the practical constraints, and, as the authors conclude, it does allow the finding that “older adults with moderate sleep complaints can improve self-rated sleep quality by initiating a regular moderate intensity exercise program” (p. 32)

5.4 DESIGNING A GOOD OBSERVATIONAL STUDY

In trying to establish causal links, observational studies start with a distinct disadvantage compared to experiments: The researchers observe, but cannot control, the explanatory variables. However, these researchers do have the advantage that they are more likely to measure participants in their natural setting. Before looking at some complications that can arise, let's look at an example of a well-designed observational study.

Baldness and Heart Attacks

On March 8, 1993, *Newsweek* announced, “A really bad hair day: Researchers link baldness and heart attacks” (p. 62). The article reported that “men with typical male pattern baldness . . . are anywhere from 30 to 300 percent more likely to suffer a heart attack than men with little or no hair loss at all.” Pattern baldness is the type affecting the crown or vertex and is not the same as a receding hairline; it affects approximately one-third of middle-aged men. The report was based on an observational study conducted by researchers at Boston University School of Medicine, in which they compared 665 men who had been admitted to the hospital with their first heart attack to 772 men in the same age group (21- to 54-years-old) who had been admitted to the same hospitals for other reasons. Thirty-five hospitals were involved, all in eastern Massachusetts and Rhode Island. The full results were reported by Lesko, Rosenberg, and Shapiro (23 February 1993).

The study found that the percentage of men who showed some degree of pattern baldness was substantially higher for those who had had a heart attack (42%) than for those who had not (34%). Further, when they used sophisticated statistical tests to ask the question in the reverse direction, they found an increased risk of heart attack for men with any degree of pattern baldness. The analysis methods included adjustments for age and other heart attack risk factors. The increase in risk was more

severe with increasing severity of baldness, after adjusting for age and other risk factors. The authors of the study speculated that there may be a third variable, perhaps a male hormone that both increases the risk of heart attacks and leads to a propensity for baldness. With an observational study such as this, scientists can establish a connection, and they can then look for causal mechanisms in future work. _

Types of Observational Studies

Case-Control Studies

Some terms are used specifically for observational studies. Case Study 5.3 is an example of a **case-control study**. In such a study, “cases” who have a particular attribute or condition are compared with “controls” who do not. In this example, those who had been admitted to the hospital with a heart attack were the cases, and those who had been admitted for other reasons were the controls. The cases and controls are compared to see how they differ on the variable of interest, which in Case Study 5.3 was the degree of baldness. Sometimes cases are matched with controls on an individual basis. This type of design is similar to a matched-pair experimental design. The analysis proceeds by first comparing the pair, then summarizing over all pairs. Unlike a matched-pair experiment, the researcher does not randomly assign treatments within pairs but is restricted to how they occur naturally. For example, to identify whether left-handed people die at a younger age, researchers might match each left-handed case with a right-handed sibling as a control, and compare their ages at death. Handedness could obviously not be randomly assigned to the two individuals, so confounding factors might be responsible for any observed differences.

Retrospective or Prospective Studies

Observational studies are also classified according to whether they are **retrospective**, in which participants are asked to recall past events, or **prospective**, in which participants are followed into the future, and events are recorded. The latter is a better procedure because people often do not remember past events accurately.

Advantages of Case-Control Studies

Case-control studies have become increasingly popular in medical research, and with good reason. Much more efficient than experiments, they do not suffer from the ethical considerations inherent in the random assignment of potentially harmful or beneficial treatments. The purpose of a case-control study is to find out whether one or more explanatory variables are related to a certain disease. For instance, in an example given later in this book, researchers were interested in whether owning a pet bird is related to incidence of lung cancer. A case-control study begins with the identification of a suitable number of cases, or people who have just been diagnosed with the disease of interest. Researchers then identify a group of controls, who are as similar as possible to the cases, except that they don’t have the disease. To achieve this similarity, researchers often use patients hospitalized for other causes as the controls.

Efficiency

The case-control design has some clear advantages over experiments as well as other observational studies. Case-control studies are very efficient in terms of time, money, and inclusion of enough people with the disease. Imagine trying to design an experiment to find out whether a relationship exists between owning a bird and getting lung cancer. You would randomly assign people to either own a bird or not and then wait to see how many in each group contracted lung cancer. The problem is that you would have to wait a long time, and even then, you would have very few cases of lung cancer in either group. In the end, you may not have enough cases for a valid comparison.

A case-control study, in contrast, would identify a large group of people who had just been diagnosed with lung cancer, and would then ask them whether they had owned a pet bird. A similar control group would be identified and asked the same question. A comparison would then be made between the proportion of cases (lung cancer patients) who had birds and the proportion of controls who had birds.

Reducing Potential Confounding Variables

Another advantage of case-control studies over other observational studies is that the controls are chosen to try to reduce potential confounding variables. For example, in Case Study 5.3, suppose it were the case that bald men were simply less healthy than other men and were therefore more likely to get sick in some way. An observational study that recorded only whether someone had baldness and whether they had had a heart attack would not be able to control for that fact. By using other hospitalized patients as the controls, the researchers were able to at least partially account for general health as a potential confounding factor. You can see that careful thought is needed to choose controls that reduce potential confounding factors and do not introduce new ones. For example, suppose we wanted to know if heavy exercise induced heart attacks and, as cases, we used people as they were admitted to the hospital with a heart attack. We would certainly not want to use other newly admitted patients as controls. People who were sick enough to enter the hospital (for anything other than sudden emergencies) would

probably not have been recently engaging in heavy exercise. When you read the results of a case-control study, you should pay attention to how the controls were selected.

5.5 DIFFICULTIES AND DISASTERS IN OBSERVATIONAL STUDIES

As with other types of research we have discussed, when you read the results of observational studies you need to watch for procedures that could negate the results of the study.

Here are some complications that can arise:

1. Confounding variables and the implications of causation
2. Extending the results inappropriately
3. Using the past as a source of data

Confounding Variables and the Implications of Causation

The Problem

Don't be fooled into thinking that a link between two variables established by an observational study implies that one causes the other. There is simply no way to separate out all potential confounding factors if randomization has not been used.

The Solution

A partial solution is achieved if researchers measure all the potential confounding variables they can imagine and include those in the analysis to see whether they are related to the response variable. Another partial solution can be achieved in case control studies by choosing the controls to be as similar as possible to the cases. The other part of the solution is up to the reader: Don't be fooled into thinking a causal relationship exists.

SMOKING DURING PREGNANCY

In Chapter 1, we introduced a study showing that women who smoked during pregnancy had children whose IQs at age 4 were lower than those of similar women who had not smoked. The difference was as high as 9 points before accounting for confounding variables, such as diet and education, but was reduced to just over 4 points after accounting for those factors. However, other confounding variables could exist that are different for mothers who smoke and that were not measured and analysed, such as amount of exercise the mother got during pregnancy. Therefore, we should not conclude that smoking during pregnancy necessarily caused the children to have lower IQs.

Extending the Results Inappropriately

The Problem

Many observational studies use convenience samples, which are generally not representative of any population. Results should be considered with that in mind. Case control studies often use only hospitalized patients, for example.

The Solution

If possible, researchers should use an entire segment of the population rather than just a convenient sample. In studying the relationship between smoking during pregnancy and child's IQ, for example, the researchers included most of the women in a particular county in upstate New York who were pregnant during the right time period. Had they relied solely on volunteers recruited through the media, their results would not be as extendable.

BALDNESS AND HEART AT TACKS

The observational study relating baldness and heart attacks only used men who were hospitalized for some reason. Although that may make sense in terms of providing a more similar control group, you should consider whether the results should be extended to all men.

Using the Past as a Source of Data

The Problem

Retrospective observational studies can be particularly unreliable because they ask people to recall past behaviour. Some medical studies, in which the response variable is whether someone has died, can be even worse because they rely on the memories of relatives and friends rather than on the actual participants. Retrospective studies also suffer from the fact that variables that confounded things in the past may no longer be similar to those that would currently be confounding variables, and researchers may not think to measure them.

The Solution

If at all possible, prospective studies should be used. That's not always possible. For example, researchers who first considered the potential causes of AIDS or Toxic Shock Syndrome had to start with those who were inflicted and try to find common factors from their pasts. If possible, retrospective studies should use authoritative sources such as medical records, rather than relying on memory.

DO LEFT-HANDERS DIE YOUNG?

A few years ago, a highly publicized study pronounced that left-handed people did not live as long as right-handed people (Coren and Halpern, 1991). In one part of the study, the researchers had sent letters to next of kin for a random sample of recently deceased individuals, asking which hand the deceased had used for writing, drawing, and throwing a ball. They found that the average age of death for those who had been left-handed was 66, whereas for those who had been right-handed, it was 75. What they failed to take into account was that in the early part of the 20th century, many children were forced to write with their right hands, even if their natural inclination was to be left-handed. Therefore, people who died in their 70s and 80s during the time of this study were more likely to be right-handed than those who died in their 50s and 60s. The confounding factor of how long ago one learned to write was not taken into account. A better study would be a prospective one, following current left- and right-handers to see which group survived longer.

Chapter 7

Summarizing and Displaying Measurement Data

7.1 TURNING DATA INTO INFORMATION

Looking at a long list of numbers is about as informative as looking at a scrambled set of letters. To get information out of data, the data have to be organized and summarized. As an example, suppose you were told that you received a score of 80 on an examination and that the scores in the class were as follows:

75, 95, 60, 93, 85, 84, 76, 92, 62, 83, 80, 90, 64, 75, 79, 32, 78, 64, 98, 73, 88, 61, 82, 68, 79, 78, 80, 55

How useful would that list of numbers be to you, at first glance? Do you have any idea where you are relative to the rest of the class? The first thought that may occur to you is to put the numbers into increasing order so you could see where your score was relative to the others. Doing that, you find:

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

How you can see that you are somewhat above the middle, but this list still isn't easy to assimilate into a useful picture. It would help if we could summarize the numbers.

The Mean, the Median, and Outliers

There are three kinds of useful information about a set of data, and each can be measured and expressed in a variety of ways. One useful concept is the idea of the "centre" of the data. What's a typical or average value? For the test scores just given, the numerical **average**, or **mean**, is 76.04. As another measure, consider that there were 28 values in that test score set, so the **median**, with half of the scores above and half of the scores below it, is 78.5, halfway between 78 and 79.

You can see that the median is somewhat higher than the mean in this case. That's because a very low score, 32, pulled down the mean. It didn't pull down the median because, as long as that very low score was 78 or less, its effect on the median would be the same. If one or two scores are far removed from the rest of the data, they are called **outliers**. There are no hard and fast rules for determining what qualifies as an outlier, but in this case most people would agree that the score of 32 is an outlier.

Mode

Another measure of "centre," called the **mode**, is occasionally useful. The mode is simply the most common value in the list. For the exam scores, no single mode exists because each of the scores 64, 75, 8, 79 and 80 occurs twice. The mode is most useful for discrete or categorical data with a relatively small number of possible values. For example, if you measured the class standing of all the students in your statistics class and coded them with 1 = freshman, 2 = sophomore, and so on, it would probably be more useful to know the mode (most common class standing) than to know the mean or the median.

Variability

The second kind of useful information contained in a set of data is the **variability**. How spread out are the values? Are they all close together? Are most of them together, but a few of them outliers? Knowing that the mean is about 76, your test score of 80 is still hard to interpret. It would obviously have a different connotation for you if the scores ranged from 72 to 80 rather than from 32 to 98.

Range

The simplest measure of variability is the **range**, which is just as it sounds. In this case, the scores went from 32 to 98, for a range of 66 points. We look at a more complicated measure of variability, the standard deviation, later.

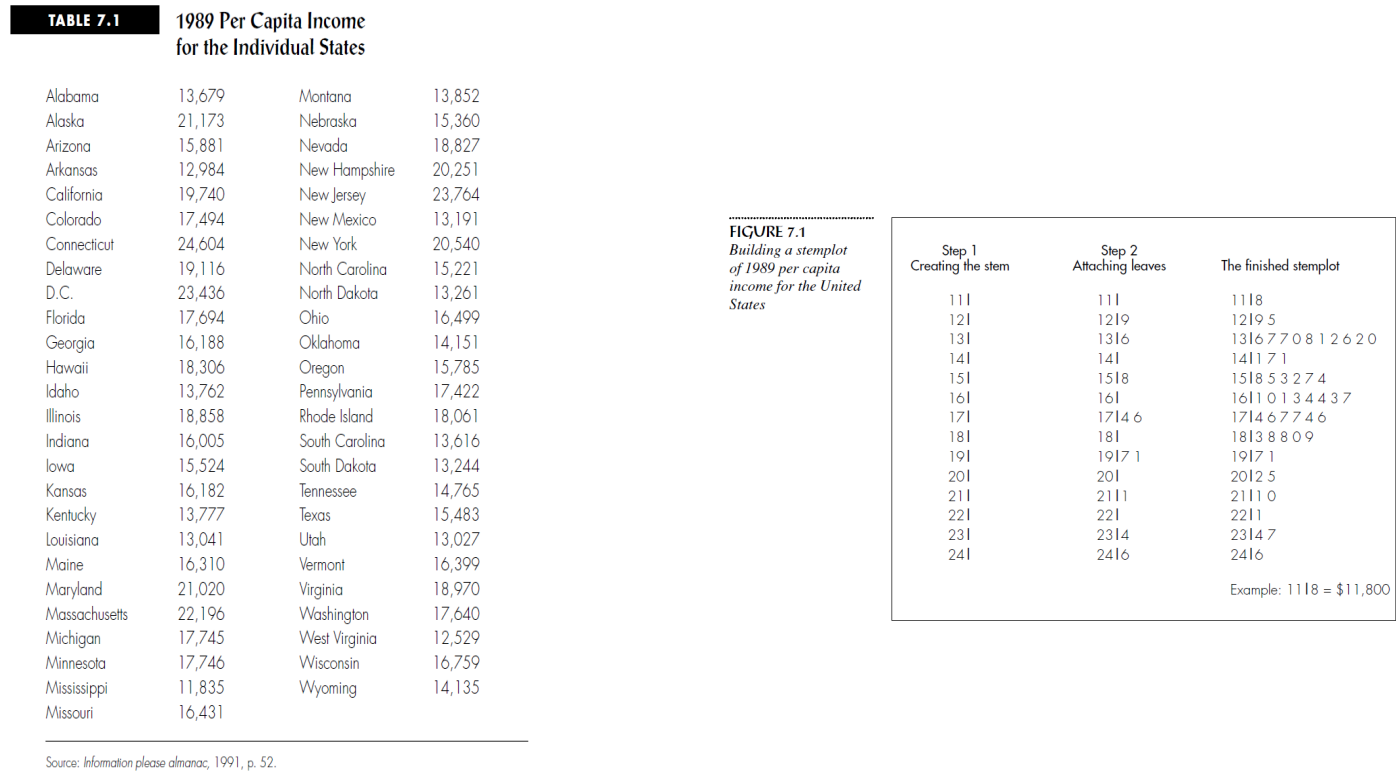
Shape

The third kind of useful information is the **shape**, which can be derived from a certain kind of picture of the data. We can answer questions such as: Are most of the values clumped in the middle with values tailing off at each end? Are there two distinct groupings? Are most of the values clumped together at one end with a few very high or low values? You can see that your score of 80 would have different meanings depending on how the other students' scores grouped together. For example, if half of the remaining students had scores of 50 and the other half scores of 100, then even though your score of 80 was "above average," it wouldn't look so good. Next we focus on how to look at the shape of the data.

7.2 THE TOTAL PICTURE: STEMPLOTS AND HISTOGRAMS

About Stemplots

A **stemplot** is a quick and easy way to put a list of numbers into order while getting a picture of their shape. The easiest way to describe a stemplot is to construct one. We could use the test scores we've been using, but instead let's turn to some real data, where each number has an identity. Table 7.1 lists per capita income for each of the 50 states and the District of Columbia in 1989. Scanning the list gives us some information, but it would be easier to get the big picture if it were in some sort of numerical order. We could simply list the states by value instead of alphabetically, but that would not give us a picture of the shape.



Before reading any further, look at the right-most part of Figure 7.1 so you can see what a completed stemplot looks like. Each of the digits extending to the right represents one data point. The first thing you see is 11|8. That represents a per capita income of \$11,800, which is a truncated version of the per capita income of Mississippi, \$11,835, the lowest per capita income in the batch. (To truncate a number, simply drop off the digits at the end. This is slightly different from rounding off.) When more than one state shares the number to the left, a number is given on the right for each of those states. For example, the second item in the stemplot is 12|9 5. This represents two states: Arkansas, \$12,984, and West Virginia, \$12,529. For one more example, notice the bottom of the picture, 24|6. That represents \$24,600, which is the truncated value for the \$24,604 per capita income of Connecticut.

Stemplots are sometimes called **stem-and-leaf plots** or **stem-and-leaf diagrams**. Only two steps are needed to create a stemplot.

In this example, the per capita incomes range from a low of \$11,835 (for Mississippi) to a high of \$24,604 (for Connecticut). The goal is to use the first few digits in each number as the stem, in such a way that about 6 to 15 numbers are lined up vertically. In this case, we use the first two digits, ranging from 11, representing \$11,000, to 24, representing \$24,000, resulting in 14 stems.

The second step is to attach a **leaf** to represent each data point. The next digit in the number is used as the leaf, and the remaining digits are simply dropped. For example, for Alabama, the first number in our list, the per capita income is \$13,679. The first two digits have been used in the stem, so we use the third digit, 6, as the leaf. We attach it to the stem value of 13. The middle part of Figure 7.1 shows the picture after leaves have been attached only for the states of Alabama through Florida. The finished picture, on the right, has the leaves attached for the per capita incomes of all 51 states. Sometimes an additional step is taken and the leaves are ordered numerically on each branch.

Now that we have a picture of the per capita income data, it is easy to make several observations. First, there is a wide range of incomes, with Connecticut, the highest, being more than twice that of Mississippi, the lowest. Second, there appear to be two clusters of values, one in the range of \$13,000 and another in the \$16,000 to \$17,000 range. Third, more states are at the lower end than at the upper end. Although the range from \$20,000 and above includes over one-third of the stem values, there are only eight states, or fewer than one-sixth the total number of states, in that range. Finally, we can see that there are no extreme outliers; that is, no states are very far above or below the others.

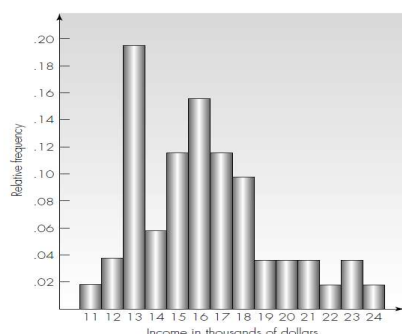
The solution to this problem is to reuse each of the digits 5, 6, and 7. Because you need to have equally spaced intervals, you could use each of the digits two or five times. If you use them each twice, the first listed would receive leaves from 0 to 4, and the second would receive leaves from 5 to 9. Thus, each stem value would encompass a range of five beats per minute of pulse. If you use each digit five times, each stem value would receive leaves of two possible values. The first stem for each digit would receive 0 and 1, the second would receive 2 and 3, and so on. Notice that if you tried to use the initial pulse digits three or four times each, you could not evenly divide the leaves among them because there are always ten possible values for leaves. Figure 7.2 shows two possible stemplots for the same hypothetical pulse data. Stemplot *a* shows the digits 5, 6, and 7 used twice; stemplot *b* shows them used five times.

514	514
517 8 9	517
610 2 3 3 4 4	518 9
615 5 5 6 7 7 8 9	610
710 0 1 2 4	612 3 3
715 8	614 4 5 5 5
a	616 7 7
	618 9
	710 0 1
	712
	714 5
	718
	b

Histograms are pictures related to stemplots. For very large data sets, a histogram is more feasible than a stemplot because it doesn't list every data value. To create a histogram, simply divide the range of the data into intervals in much the same way as we did when creating a stemplot. But instead of listing each individual value, simply count how many fall into each part of the range. Draw a bar whose height is equal to the count for each part of the range. Or, equivalently, make the height equal to the *proportion* of the total count that falls in that interval. Figure 7.3 shows a histogram for the data from Table 7.1. Notice that the heights of the bars are represented as proportions. For example, there are six values in the per capita income range from

\$17,000 to \$17,999, so that represents $6/51 = 0.12$, or 12% of the data. The heights of all of the bars must sum to 1, or 100%. If you wanted to know what proportion of the data fell into a certain interval or range, you would simply sum the heights of the bars for that range. Also notice that if you were to turn the histogram on its side, it would look very much like a stemplot except that the labels would differ slightly.

FIGURE 7.3
Histogram of per
capita income data



Defining a Common Language about Shape

Symmetric Data Sets

Scientists often talk about the “shape” of data; what they really mean is the shape of the stemplot or histogram resulting from the data. A **symmetric** data set is one in which, if you were to draw a line through the centre, the picture on one side would be a mirror image of the picture on the other side. A special case, which will be discussed in detail in Chapter 8, is a **bell-shaped** data set, in which the picture is not only symmetric, but also shaped like a bell. The stemplots in Figure 7.2, displaying pulse rates, are approximately symmetric and bell-shaped.

Unimodal or Bimodal

Recall that the mode is the most common value in a set of data. If there is a single prominent peak in a histogram or stemplot, as in Figure 7.2, the shape is called **unimodal**, meaning “one mode.” If there are two prominent peaks, the shape is called **bimodal**, meaning two modes. Figures 7.1 and 7.3 illustrate that the per capita incomes in the United States are bimodal because, as noted earlier, there appear to be two clusters of incomes.

Skewed Data Sets

In common language, something that is skewed is off-centre in some way. In statistics, a **skewed** data set is one that is basically unimodal but is substantially off from being bell-shaped. If it is **skewed to the right**, the higher values are more spread out than the lower values, as in the per capita income data. If it is **skewed to the left**, then the lower values are more spread out and the higher ones tend to be clumped. This terminology results from the fact that, before computers were used, shape pictures were always hand drawn using the horizontal orientation in Figure 7.3. Notice that a picture that is skewed to the right, like Figure 7.3, extends further to the right of the highest peak (the tallest bar) than to the left.

7.3 FIVE USEFUL NUMBERS: A SUMMARY

A **five-number summary** is a useful way to summarize a long list of numbers. As the name implies, this is a set of five numbers that provide a good summary of the entire list. Figure 7.7 shows what the five useful numbers are and the order in which they are usually displayed.

The lowest and highest values are self-explanatory. The median, which we discussed earlier, is the number such that half of the values are above it and half are below it. If there are an odd number of values in the data set, the median is simply the middle value in the ordered list. If there is an even number of values, the median is the average of the middle two values. For example, the median of the list 70, 75, 85, 86, 87 is 85 because it is the middle value. If the list had an additional value of 90 in it, the median would be 85.5, the average of the middle two numbers, 85 and 86.

The median can be found quickly from a stemplot, especially if the leaves have been ordered. Using Figure 7.1, convince yourself that the median per capita income is the 26th value ($51 = 25 + 1 + 25$) from either end, which is \$16,300. (Be careful to count the values in numerical order when you get to the stem value 16, and not in the order in which they appear.)

The **quartiles** are simply the medians of the two halves of the ordered list. The **lower quartile**—because it’s halfway into the first half—is one quarter of the way from the bottom. Similarly, the **upper quartile** is one quarter of the way down from the top. Complicated algorithms exist for finding exact quartiles. We can get close enough by simply finding the median first, then finding the medians of all the numbers below it and all the numbers above it. For the per capita income data, the lower quartile

would be the median of the 25 values below the median of \$16,300, or \$13,800. The upper quartile would be the median of the upper 25 values, \$18,800. Notice that these are the 13th from the bottom and the top of each half because $25 = 12 + 1 + 12$. The five-number summary for the per capita income data is thus:

\$16,300	
\$13,800	\$18,800
\$11,800	\$24,600

These five numbers provide a useful summary of the entire set of 51 numbers. We can get some idea of the middle, the spread, and whether or not the values are clumped at one end or the other. Because a much larger gap exists between \$18,800 and \$24,600 than between \$11,800 and \$13,800, we know that the values are probably more clumped at the lower end and more spread out at the upper end. In other words, this confirms that the data are skewed to the right.

One final note about using stemplots in this way: Remember that we dropped the last two digits on the per capita incomes when we created the stemplots. If we wanted to find the exact median for the data, we would simply figure out which value corresponded to \$16,300 on the stemplot. Scanning the list in Figure 7.1, we see that it is \$16,310, the per capita income for Maine. Thus, the exact median is \$16,310. Similarly, we could find the exact values for the other four numbers in the five-number summary. However, the truncated values from the stemplot are generally close enough to give us the picture we need.

7.4 BOXPLOTS

A visually appealing and useful way to present a five-number summary is through a *boxplot*, sometimes called a *box and whisker plot*. This simple picture also allows easy comparison of the centre and spread of data collected for two or more groups.

HOW MUCH DO STATISTICS STUDENTS SLEEP?

During spring semester, 1998, 190 students in a statistics class at a large university were asked to answer a series of questions in class one day, including how many hours they had slept the night before (a Tuesday night). A five-number summary for the reported number of hours of sleep is

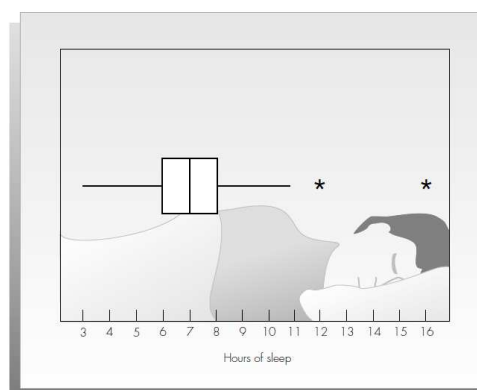
7	
6	8
3	16

Two individuals reported that they slept 16 hours; the maximum for the remaining 188 students was 12 hours.

Creating a Boxplot

The boxplot for the hours of sleep is presented in Figure 7.8, and illustrates how a boxplot is constructed. The ends of the “box” represent the lower and upper quartiles. The line in the middle of the box represents the median. The width of the box is the distance between the lower and upper quartiles, so the box has a range that covers half the data values. This distance is called the **interquartile range** because it’s the distance between the lower and upper quartiles. The lines or “whiskers” extend from the box until they either reach the minimum (on the left) and maximum (on the right) or until they have already extended a specified length, whichever comes first. The specified length is 1.5 times the width of the box. The asterisks indicate data values that are beyond that specified length and are considered to be outliers. In Figure 7.8, notice that the box extends from 6 hours to 8 hours (the lower and upper quartiles), with the median marked at 7 hours. The interquartile range is thus 2 hours (from 6 to 8 hours), so the whiskers will extend a

FIGURE 7.8
Boxplot for hours
of sleep



maximum of $1.5 \times 2 = 3$ hours beyond the box. The left whisker does exactly that, coincidentally reaching the minimum value of 3 hours of sleep. The right whisker extends to 11 hours ($8 + 3$), but that leaves outliers at 12 hours and 16 hours, which are marked by asterisks. The asterisk at 16 hours actually represents two equal data values, but there is no way to indicate that on the boxplot.

Interpreting Boxplots

Notice that boxplots essentially divide the data into fourths. The lowest fourth of the data values is contained in the range of values below the start of the box, the next fourth is contained in the first part of the box (between the lower quartile and the median), the next fourth is in the upper part of the box, and the final fourth is between the box and the upper end of the picture. Outliers are also easily identified. In the boxplot in Figure 7.8, we can see that one-fourth of the students slept between 3 and 6 hours the previous night, one-fourth slept between 6 and 7 hours, one-fourth slept between 7 and 8 hours, and the final fourth slept between 8 and 16 hours. We can thus immediately see that the data are skewed to the right because the final fourth covers an 8-hour period whereas the lowest fourth covers only a 3-hour period. As the next example illustrates, boxplots are particularly useful for comparing two or more groups on the same measurement. Although almost the same information is contained in five-number summaries, the visual display makes similarities and differences much more obvious.

7.5 TRADITIONAL MEASURES: MEAN, VARIANCE, AND STANDARD DEVIATION

The five-number summary has come into use relatively recently. Traditionally, only two numbers have been used to describe a set of numbers: the **mean**, representing the centre, and the **standard deviation**, representing the spread or variability in the values. Sometimes the **variance** is given instead of the standard deviation. The standard deviation is simply the square root of the variance, so once you have one you can easily compute the other.

The mean and standard deviation are most useful for symmetric sets of data with no outliers. However, they are very commonly quoted, so it is important to understand what they represent, including their uses and their limitations.

The Mean

As we discussed earlier, the **mean** is the numerical average of a set of numbers. In other words, we add up the values and divide by the number of values. The mean can be distorted by one or more outliers and is thus most useful when there are no extreme values in the data. For example, suppose you are a student taking four classes, and the number of students in each is, respectively, 20, 25, 35, and 200. What is your typical class size? Notice that the median is 30 students. The mean, however, is $280/4$ or 70 students. The mean is severely affected by the one large class size of 200 students.

As another example, refer to Figure 7.5, which displays claims paid by an insurance company for collision damages. The majority of claims were for amounts less than \$4000, and the median claim is only \$3500. But because there were a few very high claims, the mean amount paid is \$5178.

Because the mean can be distorted by outliers, data involving incomes or prices are usually summarized by using the median. For example, the median price of a house in a given area, instead of the mean price, is routinely quoted in the newspaper. That's because one house that sold for several million dollars would substantially distort the mean but would have little effect on the median.

Again, the mean is most useful for symmetric data sets with no outliers. As another example, notice that the British male heights in Figure 7.4 fit that description. In such cases, the mean and median should be about equal. For Figure 7.4, the mean height is 1732.5 millimetres (about 68.25 inches) and the median height is 1725 millimetres (about 68 inches).

The Standard Deviation and Variance

It is not easy to compute the **standard deviation** of a set of numbers, but most calculators and computer programs such as *Excel* now handle that task for you. It is more important to know how to interpret the standard deviation, which is a useful measure of how spread out the numbers are.

Consider the following two sets of numbers, both with a mean of 100:

NUMBERS	MEAN	STANDARD DEVIATION
100, 100, 100, 100, 100	100	0
90, 90, 100, 110, 110	100	10

The first set of numbers has no spread or variability to it at all. It has a standard deviation of 0. The second set has some spread to it; on average, the numbers are about 10 points away from the mean, except for the number that is exactly at the mean. That set has a standard deviation of 10.

Computing the Standard Deviation

Here are the steps necessary to compute the standard deviation:

1. Find the mean.
2. Find the deviation of each value from the mean = value – mean.
3. Square the deviations.
4. Sum the squared deviations.
5. Divide the sum by the number of values – 1 ($n-1$), resulting in the variance.
6. Take the square root of the variance. The result is the standard deviation.

Let's try this for the set of values 90, 90, 100, 110, 110. The number of values (n) is 5

1. The mean is 100.
2. The deviations are –10, –10, 0, 10, 10.
3. The squared deviations are 100, 100, 0, 100, 100.
4. The sum of the squared deviations is 400.
5. $n-1 = 5-1 = 4$, so the variance is $400/4 = 100$.
6. The standard deviation is the square root of 100, or 10.

Although it may seem more logical in step 5 to divide by the number of values, rather than by the number of values minus 1, there is a technical reason for subtracting 1. The reason is beyond the level of this discussion but concerns statistical bias, as discussed in Chapter 3.

The easiest interpretation is to recognize that *the standard deviation is roughly the average distance of the observed values from their mean*. Where the data have a bell shape, the standard deviation is quite useful indeed. For example, the Stanford-Binet IQ test is designed to have a mean of 100 and a standard deviation of 16. If we were to produce a histogram of IQs for a large, representative group, we would find it to be approximately bell-shaped. Its centre would be at 100. If we were to determine how far each person's IQ fell from 100, we would find an average distance, on one side or the other, of about 16 points. For shapes other than bell shapes, the standard deviation is useful as an intermediate tool for more advanced statistical procedures; it is not very useful on its own, however.