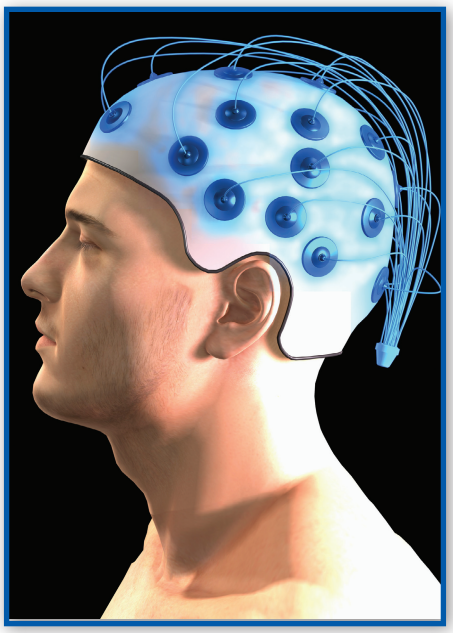
Measuring

8



MedicalRF.com/Corbis

CASE STUDY Are people with larger brains more intelligent? People have investigated this question throughout history. To answer it, we must **measure** "intelligence." This requires us to reduce the vague idea to a number that can go up or down. The first step is to say what we mean by intelligence. Does a vast knowledge of many subjects constitute intelligence? How about the ability to solve difficult puzzles or do complicated mathematical calculations? Or is it some combination of all of these?

Once we decide what intelligence is, we must actually produce the numbers. Should we use the score on a written test? Perhaps a formula that also includes grades in school would be better. Not only is it hard to say exactly what "intelligence" is, but it's hard to attach a number to measure whatever we say it is. And in the end, can we even trust the number we produce?

By the end of this chapter, you will have learned principles that will help you understand the process of measurement and determine whether you can trust the resulting numbers.

Measurement basics

Statistics deals with data, and the data may or may not be numbers. For example, planning the production of data through a sample or an experiment does not, by itself, produce numbers. Once we have our sample respondents or our experimental subjects, we must still *measure* whatever characteristics interest us. First, think broadly: Are we trying to measure the right things? Are we overlooking some outcomes that are important, even though they may be hard to measure?

EXAMPLE 1 But what about the patients?

Clinical trials tend to measure things that are easy to measure: blood pressure, tumor size, virus concentration in the blood. They often don't directly measure what matters most to patients—does the treatment really improve their lives? One study found that only 5% of trials published between 1980 and 1997 measured the effect of treatments on patients' emotional well-being or their ability to function in social settings.

Once we have decided what properties we want to measure, we can think about how to do the measurements.

Measurement

We **measure** a property of a person or thing when we assign a value to represent the property.

We often use an **instrument** to make a measurement. We may have a choice of the **units** we use to record the measurements.

The result of measurement is a numerical **variable** that takes different values for people or things that differ in whatever we are measuring.

What are your units?

Not paying attention to units of measurement can get you into trouble. In 1999, the *Mars Climate Orbiter* burned up in the Martian atmosphere. It was supposed to be 93 miles (150 kilometers) above the planet but was, in fact, only 35 miles (57 kilometers) up. It seems that Lockheed Martin, which built the *Orbiter*, specified important measurements in English units (pounds, miles). The National Aeronautics and Space Administration team, who flew the spacecraft, thought the numbers were in metric system units (kilograms, kilometers). There went $125 million.

EXAMPLE 2 Length, college readiness, highway safety

To measure the length of a bed, you can use a tape measure as the *instrument*. You can choose either inches or centimeters as the *unit of measurement*. If you choose centimeters, your *variable* is the length of the bed in centimeters.

To measure a student's readiness for college, you might ask the student to take the SAT Reasoning exam. The exam is the *instrument*. The *variable* is the student's score in points, somewhere between 400 and 1600 if you combine the Evidence-Based Reading and Writing and Mathematics sections of the SAT. "Points" are the *units of measurement*, but these are determined by a complicated scoring system described at the SAT website (www.collegeboard.com).

How can you measure the safety of traveling on the highway? You might decide to use the number of people who die in motor vehicle accidents in a year as a *variable* to measure highway safety. The government's Fatality Analysis Reporting System collects data on all fatal traffic crashes. The *unit of measurement* is the number of people who died, and the Fatality Analysis Reporting System serves as our measuring *instrument*.

Here are some questions you should ask about the variables in any statistical study:

1. Exactly how is the variable defined?

2. Is the variable an accurate way to describe the property it claims to measure?

3. How dependable are the measurements?

We don't often design our own measuring devices—we use the results of the SAT or the Fatality Analysis Reporting System, for example—so we won't go deeply into that aspect of measurement. Any consumer of numbers, however, should know a bit about how they are produced.

Know your variables

Measurement is the process of turning concepts like length or employment status into precisely defined variables. Using a tape measure to turn the idea of "length" into a number is straightforward because we know exactly what we mean by length. Measuring college readiness is controversial because it isn't clear exactly what makes a student ready for college work. Using SAT scores at least says exactly how we will get numbers. Measuring leisure time requires that we first say what time counts as leisure. Even counting highway deaths requires us to say exactly what counts as a highway death: Pedestrians hit by cars? People in cars hit by a train at a crossing? People who die from injuries six months after an accident? We can simply accept the government's counts, but someone had to answer those and other questions in order to know what to count. For example, to be included the Fatality Analysis Reporting System, "a crash must involve a motor vehicle traveling on a trafficway customarily open to the public and must result in the death of at least one person (occupant of a vehicle or a non-motorist) within 30 days of the crash." The details of when a death is counted as a highway death are necessary because they can make a difference in the data.

EXAMPLE 3 Measuring unemployment

Each month the Bureau of Labor Statistics (BLS) announces the *unemployment rate* for the previous month. People who are not available for work (retired people, for example, or students who do not want to work while in school) should not be counted as unemployed just because they don't have a job. To be unemployed, a person must first be in the labor force. That is, the person must be available for work and looking for work. The unemployment rate is

$$\text{unemployment rate} = \frac{\text{number of people unemployed}}{\text{number of people in the labor force}}$$

To complete the exact definition of the unemployment rate, the BLS has very detailed descriptions of what it means to be "in the labor force" and what it means to be "employed." For example, if you are on strike but expect to return to the same job, you count as employed. If you are not working and did not look for work in the last two weeks, you are not in the labor force. So, people who say they want to work but are too discouraged to keep looking for a job don't count as unemployed. The details matter. The official unemployment rate would be different if the government were to use a different definition of unemployment.

The BLS estimates the unemployment rate based on interviews with the sample in the monthly Current Population Survey. The interviewer can't simply ask, "Are you in the labor force?" and "Are you employed?" Many questions are needed to classify a person as employed, unemployed, or not in the labor force. Changing the questions can change the unemployment rate. At the beginning of 1994, after several years of planning, the BLS introduced computer-assisted interviewing and improved its questions. Figure 8.1 is a graph of the unemployment rate that appeared on the front page of the BLS monthly news release on the employment situation. There is a gap in the graph before January 1994 because of the change in the interviewing process. The unemployment rate would have been 6.3% under the old system. It was 6.7% under the new system. That's a big enough difference to make politicians unhappy.

Measurements, valid and invalid

No one would object to using a tape measure reading in centimeters to measure the length of a bed. Many people object to using SAT scores to measure readiness for college. Let's shortcut that debate: just measure the height in inches of all applicants and accept the tallest. Bad idea, you say.

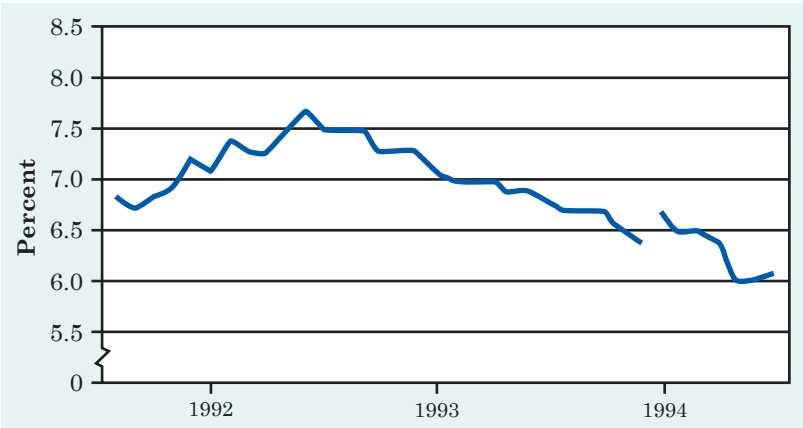


Figure 8.1 The unemployment rate from August 1991 to July 1994. The gap shows the effect of a change in how the government measures unemployment.

Why? Because height has nothing to do with being prepared for college. In more formal language, height is not a *valid* measure of a student's academic background.

Valid measurement

A variable is a **valid** measure of a property if it is relevant or appropriate as a representation of that property.

It is valid to measure length with a tape measure. It isn't valid to measure a student's readiness for college by recording her height. The BLS unemployment rate is a valid measure, even though changes in the official definitions would give a somewhat different measure. Let's think about measures, both valid and invalid, in some other settings.

EXAMPLE 4 Measuring highway safety


Mahaux Photography/Getty Images

Roads got better. Speed limits increased. Big SUVs and crossovers have replaced some cars, while smaller cars and hybrid vehicles have replaced others. Enforcement campaigns reduced drunk driving. How did highway safety change between 2007 and 2012 in this changing environment?

We could just count deaths from motor vehicles. The Fatality Analysis Reporting System says there were 41,259 deaths in 2007

and 33,561 deaths five years later in 2012. The number of deaths decreased. These numbers alone show progress. However, we need to keep in mind other things that happened during this same time frame to determine how much progress has been made. For example, the number of licensed drivers rose from 206 million in 2007 to 212 million in 2012. The number of miles that people drove decreased from 3031 billion to 2969 billion during this same time period. If more people drive fewer miles, should we expect more or fewer deaths? The count of deaths alone is not a valid measure of highway safety. So what should we use instead?

Rather than a *count*, we should use a *rate*. The number of deaths per mile driven takes into account the fact that more people drive more miles than in the past. In 2012, vehicles drove 2,969,000,000,000 miles in the United States. Because this number is so large, it is usual to measure safety by deaths per 100 million miles driven rather than deaths per mile. For 2012, this death rate is

$$\frac{\text{motor vehicle deaths}}{\text{100s of millions of miles driven}} = \frac{33{,}561}{29{,}690} = 1.1$$

The death rate fell from 1.4 deaths per 100 million miles in 2007 to 1.1 in 2012. That's a decrease—there were 21% fewer deaths per mile driven in 2012 than in 2007. Driving became safer during this time period even though there were more drivers on the roads.

Rates and counts

Often, a **rate** (a fraction, proportion, or percentage) at which something occurs is a more valid measure than a simple **count** of occurrences.

NOW IT'S YOUR TURN

8.1 Driver fatigue. A researcher studied the number of traffic accidents that were attributed to driver fatigue at different times of the day. He noticed that the number of accidents was higher in the late afternoon (between 5 and 6 P.M.) than in the early afternoon (between 1 and 2 P.M.). He concluded that driver fatigue plays a more prominent role in traffic accidents in the late afternoon than in the early afternoon. Do you think this conclusion is justified?

Using height to measure readiness for college and using counts when rates are needed are examples of clearly invalid measures. The tougher questions concern measures that are neither clearly invalid nor obviously valid.

EXAMPLE 5 Achievement tests

When you take a chemistry exam, you hope that it will ask you about the main points of material listed in the course syllabus. If it does, the exam is a valid measure of how much you know about the course material. The College Board, which administers the SAT, also offers Advanced Placement (AP) exams in a variety of disciplines. These AP exams are not very controversial. Experts can judge validity by comparing the test questions with the syllabus of material the questions are supposed to cover.

EXAMPLE 6 IQ tests

Psychologists would like to measure aspects of the human personality that can't be observed directly, such as "intelligence" or "authoritarian personality." Does an IQ test measure intelligence? Some psychologists say Yes rather loudly. There is such a thing as general intelligence, they argue, and the various standard IQ tests do measure it, though not perfectly. Other experts say No equally loudly. There is no single intelligence, just a variety of mental abilities (for example, logical, linguistic, spatial, musical, kinesthetic, interpersonal, and intrapersonal) that no one instrument can measure.

The disagreement over the validity of IQ tests is rooted in disagreement over the nature of intelligence. If we can't agree on exactly what intelligence is, we can't agree on how to measure it.

Statistics is little help in these examples. The examples start with an idea like "knowledge of chemistry" or "intelligence." If the idea is vague, validity becomes a matter of opinion. However, statistics can help a lot if we refine the idea of validity a bit.

EXAMPLE 7 The SAT again

"SAT bias will illegally cheat thousands of young women out of college admissions and scholarship aid they have earned by superior classroom performance." That's what the organization FairTest said when the 1999 SAT scores were released. The gender gap was larger on the math part of the test, where women averaged 495 and men averaged 531.

Fifteen years later, in 2014, the gap remained. Among high school seniors, women averaged 499 and men 530 on the math part of the test. The federal Office of Civil Rights says that tests on which women and minorities score lower are discriminatory.

The College Board, which administers the SAT, replies that there are many reasons some groups have lower average scores than others. For example, more women than men from families with low incomes and little education sign up for the SAT. Students whose parents have low incomes and little education have, on the average, fewer advantages at home and in school than richer students. They have lower SAT scores because their backgrounds have not prepared them as well for college. The mere fact of lower scores doesn't imply that the test is not valid.

Is the SAT a valid measure of readiness for college? "Readiness for college academic work" is a vague concept that probably combines inborn intelligence (whatever we decide that is), learned knowledge, study and test-taking skills, and motivation to work at academic subjects. Opinions will always differ about whether SAT scores (or any other measure) accurately reflect this vague concept.

Instead, we ask a simpler and more easily answered question: do SAT scores help predict students' success in college? Success in college is a clear concept, measured by whether students graduate and by their college grades. Students with high SAT scores are more likely to graduate and earn (on the average) higher grades than students with low SAT scores. We say that SAT scores have *predictive validity* as measures of readiness for college. This is the only kind of validity that data can assess directly.

What can't be measured matters One member of the young Edmonton Oilers hockey team of 1981 finished last in almost everything one can measure: strength, speed, reflexes, eyesight. That was Wayne Gretzky, soon to be known as "the Great One." He broke the National Hockey League scoring record that year, then scored yet more points in seven different seasons. Somehow the physical measurements didn't catch what made Gretzky the best hockey player ever. Not everything that matters can be measured.

Predictive validity

A measurement of a property has **predictive validity** if it can be used to predict success on tasks that are related to the property measured.

Predictive validity is the clearest and most useful form of validity from the statistical viewpoint. "Do SAT scores help predict college grades?" is a much clearer question than "Do IQ test scores measure intelligence?" However, predictive validity is not a yes-or-no idea. We must ask *how accurately* SAT scores predict college grades. Moreover, we must ask *for what groups* the SAT has predictive validity. It is possible, for example, that the SAT predicts

college grades well for men but not for women. There are statistical ways to describe "how accurately." The Statistical Controversies feature in this chapter asks you to think about these issues.

STATISTICAL CONTROVERSIES

SAT Exams in College Admissions

Susan Stava/The New York Times/Redux

Colleges use a variety of measures to make admissions decisions. The student's record in high school is the most important, but SAT scores do matter, especially at selective colleges. The SAT has the advantage of being a national test. An A in algebra means different things in different high schools, but an SAT Math score of 625 means the same thing everywhere. The SAT can't measure willingness to work hard or creativity, so it won't predict college performance exactly, but most colleges have long found it helpful.

The accompanying table gives some results about how well SAT scores predict first-year college grades from a sample of 151,316 students in 2006. The numbers in the table say what percentage of the variation among students in college grades can be predicted by SAT scores (Critical Reading, Writing, and Math tests combined), by high school grades, and by SAT and high school grades together. An entry of 0% would mean no predictive validity, and 100% would mean predictions were always exactly correct.

How well do you think SAT scores predict first-year college grades? Should SAT scores be used in deciding college admissions?

	All institutions	Private institutions	Public institutions
SAT	28%	32%	27%
School grades	29%	30%	28%
Both together	38%	42%	37%

Measurements, accurate and inaccurate

Using a bathroom scale to measure your weight is valid. If your scale is like many commonly used ones, however, the measurement may not be very accurate. It measures weight, but it may not give the true weight. Let's say that originally your scale always read 3 pounds too high, so

$$\text{measured weight} = \text{true weight} + 3 \text{ pounds}$$

If that is the whole story, the scale will always give the same reading for the same true weight. Most scales vary a bit—they don't always give the same reading when you step off and step right back on. Your scale now is somewhat old and rusty. It still always reads 3 pounds too high because its aim is off, but now it is also erratic, so readings deviate from 3 pounds. This morning it sticks a bit and reads 1 pound too low for that reason. So the reading is

$$\text{measured weight} = \text{true weight} + 3 \text{ pounds} - 1 \text{ pound}$$

When you step off and step right back on, the scale sticks in a different spot that makes it read 1 pound too high. The reading you get is now

$$\text{measured weight} = \text{true weight} + 3 \text{ pounds} + 1 \text{ pound}$$

You don't like the fact that this second reading is higher than the first, so you again step off and step right back on. The scale again sticks in a different spot and you get the reading

$$\text{measured weight} = \text{true weight} + 3 \text{ pounds} - 1.5 \text{ pounds}$$

If you have nothing better to do than keep stepping on and off the scale, you will keep getting different readings. They center on a reading 3 pounds too high, but they vary about that center.

Your scale has two kinds of errors. If it didn't stick, the scale would always read 3 pounds high. That is true every time anyone steps on the scale. This systematic error that occurs every time we make a measurement is called *bias*. Your scale also sticks—but how much this changes the reading differs every time someone steps on the scale. Sometimes stickiness pushes the scale reading up; sometimes it pulls it down. The result is that the scale weighs 3 pounds too high on the average, but its reading varies when we weigh the same thing repeatedly. We can't predict the error due to stickiness, so we call it *random error*.

Errors in measurement

We can think about errors in measurement this way:

$$\text{measured value} = \text{true value} + \text{bias} + \text{random error}$$

A measurement process has **bias** if it systematically tends to overstate or understate the true value of the property it measures.

A measurement process has **random error** if repeated measurements on the same individual give different results. If the random error is small, we say the measurement is **reliable**.

To determine if the random error is small, we can use a quantity called the **variance**. The variance of n repeated measurements on the same individual is computed as follows:

1. Find the arithmetic average of these n measurements.

2. Compute the difference between each observation and the arithmetic average and square each of these differences.

3. Average the squared differences by dividing their sum by $n - 1$. This average squared difference is the variance.

A reliable measurement process will have a small variance.

For the three measurements on our sticky scale, suppose that our true weight is 130 pounds. Then the three measurements are

$$130 + 3 - 1 = 132 \text{ pounds}$$
$$130 + 3 + 1 = 134 \text{ pounds}$$
$$130 + 3 - 1.5 = 131.5 \text{ pounds}$$

The average of these three measurements is

$$(132 + 134 + 131.5)/3 = 397.5/3 = 132.5 \text{ pounds}$$

The differences between each measurement and the average are

$$132 - 132.5 = -0.5$$
$$134 - 132.5 = 1.5$$
$$131.5 - 132.5 = -1$$

The sum of the squares of these differences is

$$(-0.5)^2 + (1.5)^2 + (-1)^2 = 0.25 + 2.25 + 1 = 3.5$$

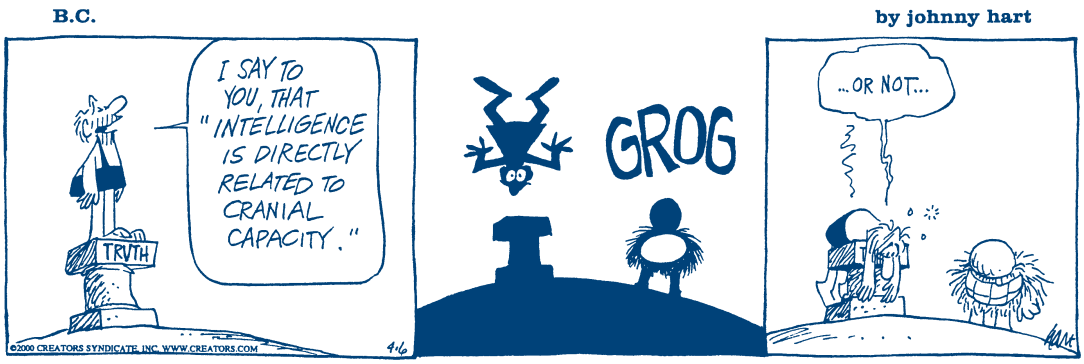
and so the variance of these random errors is

$$3.5/(3 - 1) = 1.75$$

In the text box, we said that a reliable measurement process will have a small variance. In this example, the variance of 1.75 is quite small relative to the actual weights, so it appears that the measurement process using this scale is reliable.

A scale that always reads the same when it weighs the same item is perfectly reliable even if it is biased. For such a scale, the variance of the measurements will be 0.

Reliability says only that the result is dependable. Bias means that in repeated measurements the *tendency* is to systematically either overstate or understate the true value. It does not necessarily mean that every measurement overstates or understates the true value. Bias and lack of



By permission of Johnny Hart and Creators Syndicate, Inc.

reliability are different kinds of error. And don't confuse reliability with validity just because both sound like good qualities. Using a scale to measure weight is valid even if the scale is not reliable.

Here's an example of a measurement that is reliable but not valid.

EXAMPLE 8 Do big skulls house smart brains?

In the mid-nineteenth century, it was thought that measuring the volume of a human skull would measure the intelligence of the skull's owner. It was difficult to measure a skull's volume reliably, even after it was no longer attached to its owner. Paul Broca, a professor of surgery, showed that filling a skull with small lead shot, then pouring out the shot and weighing it, gave quite reliable measurements of the skull's volume. These accurate measurements do not, however, give a valid measure of intelligence. Skull volume turned out to have no relation to intelligence or achievement.

NOW IT'S YOUR TURN

8.2 The most popular burger joint. If you live in the United States, you may have heard your friends debate whether In-N-Out Burger or Five Guys Burgers and Fries has the better hamburger. According to the Consumer Reports National Research Center, In-N-Out Burger ranked first in both 2011 and 2014. Five Guys Burgers and Fries ranked third in 2011 but seventh in 2014. Is this proof that In-N-Out has better burgers? Do you think these ratings are biased, unreliable, or both? Explain your answer.

Improving reliability, reducing bias

What time is it? Much modern technology, such as the Global Positioning System, which uses satellite signals to tell you where you are, requires very exact measurements of time. In 1967, the International

Committee for Weights and Measures defined the second to be the time required for 9,192,631,770 vibrations of a cesium atom. The cesium atom is not affected by changes in temperature, humidity, and air pressure, like physical clocks are. The National Institute of Standards and Technology (NIST) has the world's most accurate atomic clock and broadcasts the results (with some loss in transmission) by radio, telephone, and Internet.

EXAMPLE 9 Really accurate time

NIST's atomic clock is very accurate but not perfectly accurate. The world standard is Coordinated Universal Time, compiled by the International Bureau of Weights and Measures (BIPM) in Sèvres, France. BIPM doesn't have a better clock than NIST. It calculates the time by averaging the results of more than 200 atomic clocks around the world. NIST tells us (after the fact) how much it misses the correct time by. Here are the last 12 errors as we write, in seconds:

0.0000000075	0.0000000012
0.0000000069	−0.0000000020
0.0000000067	−0.0000000045
0.0000000063	−0.0000000046
0.0000000041	−0.0000000042
0.0000000032	−0.0000000036

In the long run, NIST's measurements of time are not biased. The NIST second is sometimes shorter than the BIPM second and sometimes longer, not always off in the same direction. NIST's measurements are very reliable, but the preceding numbers do show some variation. There is no such thing as a perfectly reliable measurement. The average (mean) of several measurements is more reliable than a single measurement. That's one reason BIPM combines the time measurements of many atomic clocks.

Scientists everywhere repeat their measurements and use the average to get more reliable results. Even students in a chemistry lab often do this. Just as larger samples reduce variation in a sample statistic, averaging over more measurements reduces variation in the final result.

Use averages to improve reliability

No measuring process is perfectly reliable. The **average** of several repeated measurements of the same individual is more reliable (less variable) than a single measurement.

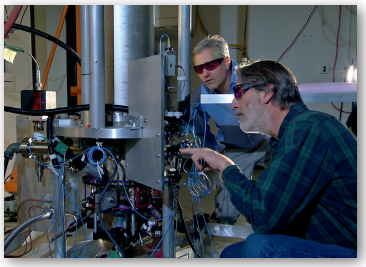


Figure 8.2 This atomic clock at the National Institute of Standards and Technology is accurate to 1 second in 6 million years. (Source: NIST.)

Unfortunately, there is no similarly straightforward way to reduce the bias of measurements. Bias depends on how good the measuring instrument is. To reduce the bias, you need a better instrument. The atomic clock at NIST (Figure 8.2) is accurate to 1 second in 6 million years but is a bit large to put beside your bed.

EXAMPLE 10 Measuring unemployment again

Measuring unemployment is also "measurement." The concepts of bias and reliability apply here just as they do to measuring length or time.

The Bureau of Labor Statistics checks the *reliability* of its measurements of unemployment by having supervisors reinterview about 5% of the sample. This is repeated measurement on the same individual, just as a student in a chemistry lab measures a weight several times.

The BLS attacks *bias* by improving its instrument. That's what happened in 1994, when the Current Population Survey was given its biggest overhaul in more than 50 years. The old system for measuring unemployment, for example, underestimated unemployment among women because the detailed procedures had not kept up with changing patterns of women's work. The new measurement system corrected that bias—and raised the reported rate of unemployment.

Pity the poor psychologist

Statisticians think about measurement much the same way as they think about sampling. In both settings, the big idea is to ask, "What would happen if we did this many times?" In sampling, we want to estimate a population parameter, and we worry that our estimate may be biased or vary too much from sample to sample. Now we want to measure the true value of some property, and we worry that our measurement may be biased or vary too much when we repeat the measurement on the same individual. Bias is systematic error that happens every time; high variability (low reliability) means that our result can't be trusted because it isn't repeatable.

Thinking of measurement this way is pretty straightforward when you are measuring your weight. To start with, you have a clear idea of what your "true weight" is. You know that there are really good scales around: start at the doctor's office, go to the physics lab, end up at NIST. You can measure your weight as accurately as you wish. This makes it easy to see that your bathroom scale

always reads 3 pounds too high. Reliability is also easy to describe—step on and off the scale many times and see how much its readings vary.

Asking "What would happen if we did this many times?" is a lot harder to put into practice when we want to measure "intelligence" or "readiness for college." Consider as an example the poor psychologist who wants to measure "authoritarian personality."

EXAMPLE 11 Authoritarian personality?

Do some people have a personality type that disposes them to rigid thinking and to following strong leaders? Psychologists looking back on the Nazis after World War II thought so. In 1950, a group of psychologists developed the "F-scale" as an instrument to measure "authoritarian personality." The F-scale asks how strongly you agree or disagree with statements such as the following:

- Obedience and respect for authority are the most important virtues children should learn.
- Science has its place, but there are many important things that can never be understood by the human mind.

Strong agreement with such statements marks you as authoritarian. The F-scale and the idea of the authoritarian personality continue to be prominent in psychology, especially in studies of prejudice and right-wing extremist movements.

Here are some questions we might ask about using the F-scale to measure "authoritarian personality." The same questions come to mind when we think about IQ tests or the SAT exam.

1. Just what is an "authoritarian personality"? We understand this much less well than we understand your weight. The answer in practice seems to be "whatever the F-scale measures." Any claim for validity must rest on what kinds of behavior high F-scale scores go along with. That is, we fall back on predictive validity.

2. The F in "F-scale" stands for Fascist. As the second question in Example 11 suggests, people who hold traditional religious beliefs are likely to get higher F-scale scores than similar people who don't hold those beliefs. Does the instrument reflect the beliefs of those who developed it? That is, would people with different beliefs come up with a quite different instrument?

3. You think you know what your true weight is. What is the true value of your F-scale score? The measuring devices at NIST can help us find

a true weight but not a true authoritarianism score. If we suspect that the instrument is biased as a measure of "authoritarian personality" because it penalizes religious beliefs, how can we check that?

4. You can weigh yourself many times to learn the reliability of your bathroom scale. If you take the F-scale test many times, you remember what answers you gave the first time. That is, repeats of the same psychological measurement are not really repeats. Thus, reliability is hard to check in practice. Psychologists sometimes develop several forms of the same instrument in order to repeat their measurements. But how do we know these forms are really equivalent?

The point is not that psychologists lack answers to these questions. The first two are controversial because not all psychologists think about human personality in the same way. The second two questions have at least partial answers but not simple answers. The point is that "measurement," which seems so straightforward when we measure weight, is complicated indeed when we try to measure human personality.

There is a larger lesson here. Be wary of statistical "facts" about squishy topics like authoritarian personality, intelligence, and even readiness for college. The numbers look solid, as numbers always do. But data are a human product and reflect human desires, prejudices, and weaknesses. If we don't understand and agree on what we are measuring, the numbers may produce more disagreement than enlightenment.

STATISTICS IN SUMMARY

Chapter Specifics

- To **measure** something means to assign a number to some property of an individual.
- When we measure many individuals, we have values of a **variable** that describes them.
- Variables are recorded in **units**.
- When you work with data or read about a statistical study, ask if the variables are **valid** as numerical measures of the concepts the study discusses.
- Often, a **rate** is a more valid measure than a **count**.
- Validity is simple for measurements of physical properties such as length, weight, and time. When we want to measure human personality

and other vague properties, **predictive validity** is the most useful way to say whether our measures are valid.

- Also ask if there are **errors in measurement** that reduce the value of the data. You can think about errors in measurement like this:

$$\text{measured value} = \text{true value} + \text{bias} + \text{random error}$$

- Some ways of measuring are **biased**, or systematically wrong in the same direction.
- To reduce bias, you must use a better **instrument** to make the measurements.
- Other measuring processes lack **reliability**, so that measuring the same individuals again would give quite different results due to **random error**.
- A reliable measuring process will have a small **variance** of the measurements. You can improve the reliability of a measurement by repeating it several times and using the **average** result.

In reasoning from data to a conclusion, we start with the data. In statistics, data are ultimately represented by numbers. The planning of the production of data through a sample or experiment does not by itself produce these numbers. The extent to which these numbers represent the characteristics we wish to study affects the quality and relevance of our conclusions. When you work with data or read about a statistical study, ask exactly how the variables are defined and whether they leave out some things you want to know. This chapter presents several ideas to think about in assessing the variables measured and hence the conclusions based on these measurements.

CASE STUDY EVALUATED The Case Study that opened the chapter is motivated by research conducted in 1991 by Willerman, Schultz, Rutledge, and Bigler. Read about this study in the EESEE story "Brain Size and Intelligence," and use what you have learned in this chapter to answer the following questions.

1. How did the researchers measure brain size? Is this a valid measure of brain size? Is it reliable? Is it biased?
2. How did the researchers measure intelligence? Is this a valid measure of intelligence?
3. The researchers found some evidence that brain size and intelligence are related. However, the study described in Example 8 did not. Discuss the differences in the two studies.

LaunchPad Online Resources
macmillan learning

- *LearningCurve* has good questions to check your understanding of the concepts.

CHECK THE BASICS

For Exercise 8.1, see page 168; for Exercise 8.2, see page 174.

8.3 Comparing marijuana use. According to the 2011–2012 National Survey on Drug Use and Health, among young adults aged 18–25 years, approximately 29.2% of young adults in Arizona, approximately 34.4% of young adults in Michigan, and approximately 30.4% of young adults in Virginia used marijuana in the past year. Which of the following is true?

(a) We cannot use these percentages to compare marijuana use by young adults in these states because we do not know how many young adults used marijuana in the past year in each state.

(b) We cannot use these percentages to compare marijuana use by young adults in these states because we do not know how many young adults live in each state.

(c) We can use these percentages to compare marijuana use by young adults in these states because we are given rates.

(d) None of the above is true.

8.4 What is the instrument of measurement? A college president is interested in student satisfaction with recreational facilities on campus. A questionnaire is sent to all students and asks them to rate their satisfaction on a scale of 1 to 5 (with 5 being the best). The instrument of measurement is

(a) a student.

(b) the questionnaire.

(c) the rating on the scale.

(d) satisfaction.

8.5 Weight at the doctor's office. When you visit the doctor's office, several measurements may be taken, one of which is your weight. A doctor's office encourages patients to keep their shoes on to be weighed and promises to subtract 2 pounds for the weight of the patient's shoes. Which of the following is true once the patient's weight is adjusted by 2 pounds?

(a) This will be a reliable measure of the patient's weight.

(b) This will be a valid measure of the patient's weight.

(c) This will be an unbiased measure of the patient's weight.

(d) None of the above is true.

8.6 Measuring athletic ability. Which of the following is *not* a valid measurement of athletic ability?

(a) Time (in seconds) to run a 100-meter dash

(b) Number of times a person goes to the gym per week

(c) Maximum weight (in pounds) a person can bench-press

(d) Number of sit-ups a person can do in one minute

8.7 Comparing teaching assistants (TAs). Professor Holmes has two teaching assistants who grade homework for a Statistics 101 course. Professor Holmes gives each of the two teaching assistants a rubric (a clear scoring guide) for the TAs to use when they grade the assignments. Holmes gives each TA the same student's paper to grade and has each TA grade the paper according to the rubric. Professor Holmes is doing this to try to guarantee the scores given by the TAs are

(a) not biased.

(b) predictive.

(c) reliable.

(d) valid.

CHAPTER 8 EXERCISES

8.8 Counting the unemployed? We could measure the extent of unemployment by a count (the number of people who are unemployed) or by a rate (the percentage of the labor force that is unemployed). The number of people in the labor force grew from 115 million in June 1985, to 132 million in June 1995, to 149 million in June 2005, to 157 million in June 2015. Use these facts to explain why the count of unemployed people is not a valid measure of the extent of unemployment.

8.9 Measuring a healthy lifestyle. You want to measure the "healthiness" of college students' lifestyles. Give an example of a clearly invalid way to measure healthiness. Then briefly describe a measurement process that you think is valid.

8.10 Rates versus counts. Customers returned 40 cell phones to Verizon this spring, and only 15 to Best Buy next door. Verizon sold 800 cell phones this spring, while Best Buy sold 200.

(a) Verizon had a greater number of cell phones returned. Why does this not show that Verizon's cell phone customers were less satisfied than those of Best Buy?

(b) What is the rate of returns (percentage of cell phones returned) at each of the stores?

(c) Use the rates of returns that you calculated to explain to a friend which store you would suggest your friend go to to purchase a cell phone.

8.11 Seat belt safety. The National Highway Traffic Safety Administration reports that in 2013, 9777 occupants of motor vehicles who were wearing a restraint died in motor vehicle accidents and 9580 who were not wearing a restraint died. These numbers suggest that not using a restraining device is safer than using one. The *counts* aren't fully convincing, however. What *rates* would you like to know to compare the safety of using a restraint with not using one?

8.12 Tough course? A friend tells you, "In the 7:30 A.M. lecture for Statistics 101, 9 students failed, but 20 students failed in the 1:30 P.M. lecture. The 1:30 P.M. prof is a tougher grader than the 7:30 A.M. prof." Explain why

the conclusion may not be true. What additional information would you need to compare the classes?

8.13 Obesity. An article in the June 30, 2010, *Columbus Dispatch* reported on the prevalence of obesity among adults in the 50 states. Based on information in the article, California has approximately 6.7 million obese adults, and Texas has approximately 5.2 million. On the other hand, Mississippi has a little over 730,000 obese adults. Do these numbers make a convincing case that California and Texas have a more substantial problem with obesity than Mississippi?

8.14 Capital punishment. Between 1977 and 2014, 1394 convicted criminals were put to death in the United States. Here are data on the number of executions in several states during those years, as well as the estimated June 1, 2014, population of these states:

State	Population (thousands)	Executions
Alabama	4,849	56
Arkansas	2,966	27
Delaware	936	16
Florida	19,893	89
Indiana	6,597	20
Nevada	2,839	12
Oklahoma	3,878	111
Texas	26,957	518

Texas and Florida are among the leaders in executions. Because these are large states, we might expect them to have many executions. Find the *rate* of executions for each of the states just listed, in executions per million population. Because population is given in thousands, you can find the rate per million as

$$\text{rate per million} = \frac{\text{executions}}{\text{population in thousands}} \times 1000$$

Arrange the states in order of the number of executions relative to population. Are Texas and Florida still high by this measure? Does any other state stand out when you examine the rates?

8.15 Measuring intelligence. One way "intelligence" can be interpreted is as "general problem-solving ability." Explain why it is *not* valid to measure intelligence by a test that asks questions such as

Who wrote "The Star-Spangled Banner"?
Who won the last soccer World Cup?

8.16 Measuring life's quality. Is life in Britain getting better or worse? The usual government data do not measure "better" or "worse" directly. So the British government announced that it wanted to add measures of such things as housing, traffic, and air pollution. "The quality of life is not simply economic," said a deputy prime minister. Help them out: how would you measure "traffic" and its impact on the quality of life?

8.17 Measuring pain. There are 9 million enrollees in the Department of Veterans Affairs health care system. It wants doctors and nurses to treat pain as a "fifth vital sign" to be recorded along with blood pressure, pulse, temperature, and breathing rate. Help out the VA: how would

you measure a patient's pain? [*Note*: There is not one correct answer for this question.]

8.18 Fighting cancer. Congress wants the medical establishment to show that progress is being made in fighting cancer. Here are some variables that might be used:

1. Total deaths from cancer. These have risen sharply over time, from 331,000 in 1970, to 505,000 in 1990, to 572,000 in 2011.
2. The percentage of all Americans who die from cancer. The percentage of deaths due to cancer rose steadily, from 17.2% in 1970 to 23.5% in 1990, then leveled off around 23.2% in 2007.
3. The percentage of cancer patients who survive for five years from the time the disease is discovered. These rates are rising slowly. The five-year survival rate was 50% in the 1975 to 1977 period and 66.5% from 2005 to 2011.

None of these variables is fully valid as a measure of the effectiveness of cancer treatment. Explain why Variables 1 and 2 could increase even if treatment is getting more effective and why Variable 3 could increase even if treatment is getting less effective.

8.19 Testing job applicants. The law requires that tests given to job applicants must be shown to be directly job related. The Department of Labor believes that an employment test called the General Aptitude Test Battery (GATB) is valid for a broad range of jobs. As in the case of the SAT, blacks and Hispanics get lower average scores on the GATB than do whites. Describe briefly what must be done to establish that the GATB has predictive validity as a measure of future performance on the job.

8.20 Validity, bias, reliability. This winter I went to a local pharmacy to have my weight and blood pressure measured using a sophisticated electronic machine at the front of the store next to the checkout counter. Will the measurement of my weight be biased? Reliable? Valid? Explain your answer.

8.21 An activity on bias. Let's study bias in an intuitive measurement. Figure 8.3 is a drawing of a tilted glass. Reproduce this drawing on 10 sheets of paper. Choose 10 people: five men and five women. Explain that the drawing represents a tilted glass of water. Ask each subject to draw the water level when this tilted glass is holding as much water as it can.

The correct level is horizontal (straight back from the lower lip of the glass). Many people make large errors in estimating the level. Use a protractor to measure the angle of each subject's error. Were your subjects systematically wrong in the same direction? How large was the average error? Was there a clear difference between the average errors made by men and by women?

Figure 8.3 A tilted glass, for Exercise 8.21. Can you draw the level of water in the glass when it is as full as possible?

8.22 An activity on bias and reliability. Cut five pieces of string having these lengths in inches:

2.9 9.5 5.7 4.2 7.6

(a) Show the pieces to another student one at a time, asking the subject to estimate the length to the nearest 10th of an inch by eye. The error your subject makes is measured value minus true value and can be either positive or negative. What is the average of the five errors? Explain why this average would be close to 0 if there were no bias and we used many pieces of string rather than just five.

(b) The following day, ask the subject to again estimate the length of each piece of string. (Present them in a different order on the second day.) Explain why the five differences between the first and second guesses would all be 0 if your subject were a perfectly reliable measurer of length. The bigger the differences, the less reliable your subject is. What is the average difference (ignoring whether they are positive or negative) for your subject?

8.23 More on bias and reliability. The previous exercise gives five true values for lengths. A subject measures each length twice by eye. Make up a set of results from this activity that matches each of the following descriptions. For simplicity, assume that bias means the same fixed error every time rather than an "on the average" error in many measurements.

(a) The subject has a bias of 0.5 inch too long and is perfectly reliable.

(b) The subject has no bias but is not perfectly reliable, so the average difference in repeated measurements is 0.5 inch.

8.24 Even more on bias and reliability. Exercise 8.22 gives five true values for lengths. A subject measures the first length (true length = 2.9 inches) four times by eye. His measurements are

3.0 2.9 3.1 3.0

Suppose his measurements have a bias of +0.1 inch.

(a) What are the four random errors for his measurements?

(b) What is the variance of his four measurements?

8.25 Does job training work? To measure the effectiveness of government training programs, it is usual to compare workers' pay before and after training. But many workers sign up for training when their pay drops or they are laid off. So the "before" pay is unusually low, and the pay gain looks large.

(a) Is this bias or random error in measuring the effect of training on pay? Why?

(b) How would you measure the success of training programs?

8.26 A recipe for poor reliability. Every month, the government releases data on "personal savings." This number tells us how many dollars individuals saved the previous month. Savings are calculated by subtracting personal spending (an enormously large number) from personal income (another enormous number). The result is one of the government's least reliable statistics.

Give a numerical example to show that small percentage changes in two very large numbers can produce a big percentage change in the difference

between those numbers. A variable that is the difference between two big numbers is usually not very reliable.

8.27 Measuring crime. Crime data make headlines. We measure the amount of crime by the number of crimes committed or (better) by crime rates (crimes per 100,000 population). The FBI publishes data on crime in the United States by compiling crimes reported to police departments. The FBI data are recorded in the Uniform Crime Reporting Program and are based on reports from more than 18,000 law enforcement agencies across the United States. The National Crime Victimization Survey publishes data about crimes based on a national probability sample of about 90,000 households per year. The victim survey shows almost two times as many crimes as the FBI report. Explain why the FBI report has a large downward bias for many types of crime. (Here is a case in which bias in producing data leads to bias in measurement.)

8.28 Measuring crime. Twice each year, the National Crime Victimization Survey asks a random sample of households whether they have been victims of crime and, if so, the details. In all, nearly 160,000 people in about 90,000 households answer these questions per year. If other people in a household are in the room while one person is answering questions, the measurement of, for example, rape and other sexual assaults could be seriously biased. Why? Would the presence of other people lead to overreporting or underreporting of sexual assaults?

8.29 Measuring pulse rate. You want to measure your resting pulse rate. You might count the number of beats in five seconds and multiply by 12 to get beats per minute.

(a) Consider counting the number of beats in 15 seconds and multiplying by 4 to get beats per minute. In what way will this improve the reliability of your measurement?

(b) Why are the first two measurement methods less reliable than actually measuring the number of beats in a minute?

8.30 Testing job applicants. A company used to give IQ tests to all job applicants. This is now illegal because IQ is not related to the performance of workers in all the company's jobs. Does the reason for the policy change involve the *reliability*, the *bias*, or the *validity* of IQ tests as a measure of future job performance? Explain your answer.

8.31 The best earphones. You are writing an article for a consumer magazine based on a survey of the magazine's readers that asked about satisfaction with mid-priced earphones for the iPad and iPhone. Of 1648 readers who reported owning the Apple in-ear headphone with remote and mic, 347 gave it an outstanding rating. Only 69 outstanding ratings were given by the 134 readers who owned Klipsch Image S4i earphones with microphone. Describe an appropriate variable, which can be computed from these counts, to measure high satisfaction with a make of earphone. Compute the values of this variable for the Apple and Klipsch earphones. Which brand has the better high-satisfaction rating?

8.32 Where to work? Each year, *Forbes* magazine ranks the 2000 largest metropolitan areas in the United States in an article on the best places for businesses and careers. First place in 2014 went to Raleigh, North Carolina. Raleigh was ranked third in 2013. Second place in 2014 went to Des Moines, Iowa. Des Moines was ranked first in 2013. Anchorage, Alaska, was ranked 53rd in 2014 but was ranked 18th in 2013! Are these facts evidence that *Forbes*'s ratings are invalid, biased, or unreliable? Explain your choice.

8.33 Validity, bias, reliability. Give your own example of a measurement process that is valid but has large bias. Then give your own example of a measurement process that is invalid but highly reliable.

EXPLORING THE WEB

Follow the QR code to access exercises.