yatesch01pg01_62 11/19/03 5:41 PM Page 22

1.2 MEASURING

Got some spare time?

Do people have more or less free time now than they did a generation ago? A book titled *The Overworked Americans* says we are working more than ever. Another book, titled *Time for Life,* says that we have more free time than ever. Is somebody lying with statistics?[13]

To see if free time is increasing, we must **measure** "free time." That is, we must reduce the vague idea to a number that can go up or down. The first step is to say what we mean by free time: something like "time when you're not working, not traveling to or from work, not doing household chores, not . . ." You see that we can get different numbers by changing our list of "nots."

Once we decide what free time is, we must actually produce the numbers. We might ask a sample of people how they spent their time yesterday. Don't trust their memory? Afraid they will exaggerate how long they worked? If so, we might ask people to keep a diary of how they spend their time today. Of course, the really busy people may forget to record all their work time. Not only is it hard to say exactly what "free time" is, but it's hard to attach a number to measure whatever we say it is.

Sample surveys are wonderful. Well-designed experiments are even more wonderful. But in the end, we need to turn ideas like "free time" or "pain relief" or "income" into numbers. Don't trust the numbers until you know how this was done.

© NAS. Reprinted with special permission of North American Syndicate.

" WOW! SIX FEET LONG AND THREE FEET DEEP."

yatesch01pg01_62 11/19/03 5:41 PM Page 23

ACTIVITY 1.2A The tilted glass

Materials: *Unlined paper, protractor*

1. On an unlined sheet of paper, carefully trace the tilted glass.

2. The drawing represents a tilted glass of water. Draw a line to indicate the water level if the glass is as full of water as possible.

3. Use a protractor to measure the acute angle that the bottom of the cup makes with horizontal. Record your answer to the nearest degree.

4. Now measure the angle between the line you have drawn and horizontal to the nearest degree.

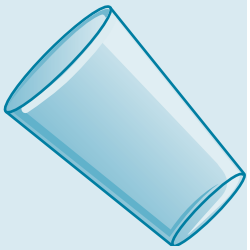
5. Record data from each member of your class in a table like this one.

Tilt of glass	Tilt of my line	Gender

6. Your teacher will provide you with the correct values. How do your measurements compare?

7. Make a dotplot to display the "tilt of glass" measurements for the class.

8. Write a few sentences to describe how the class's measurements compare with the correct values.



9. Calculate the average "tilt of glass" measurement for the class.

10. Now look at the "tilt of my line" measurements. Determine whether students' errors were mostly in one direction from the correct answer.

11. Compare the errors made by male and female students on the "tilt of my line" measurements. (You may want to repeat this activity with more students. Then you can see if your class's results hold true in general.)

yatesch01pg01_62 11/19/03 5:41 PM Page 24

Measurement basics

Statistics deals with numbers. Planning the production of data through a sample survey 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 1.10 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.[14]

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 number 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.

EXAMPLE 1.11 Length, college readiness, highway safety

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

To measure a student's readiness for college, I might ask the student to take the SAT Reasoning exam. The exam form is the *instrument*. The *variable* is the student's score in points, somewhere between 400 and 1600 if I combine the verbal and mathematics sections of the SAT.

yatesch01pg01_62 11/19/03 5:41 PM Page 25

How can I measure the safety of traveling on the highway? I might decide to count the number of people who die in motor vehicle accidents in a year. The government's Fatal Accident Reporting System collects data on all fatal traffic crashes. I can use the government count of traffic deaths as a *variable* to measure highway safety.

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 a valid way to describe the property it claims to measure?
3. How accurate are the measurements?

We don't often design our own measuring devices (for example, we use the results of the SAT or the Fatal Accident Reporting System), so we won't go deeply into these questions. Any consumer of numbers, however, should know a bit about them.

Know your variables

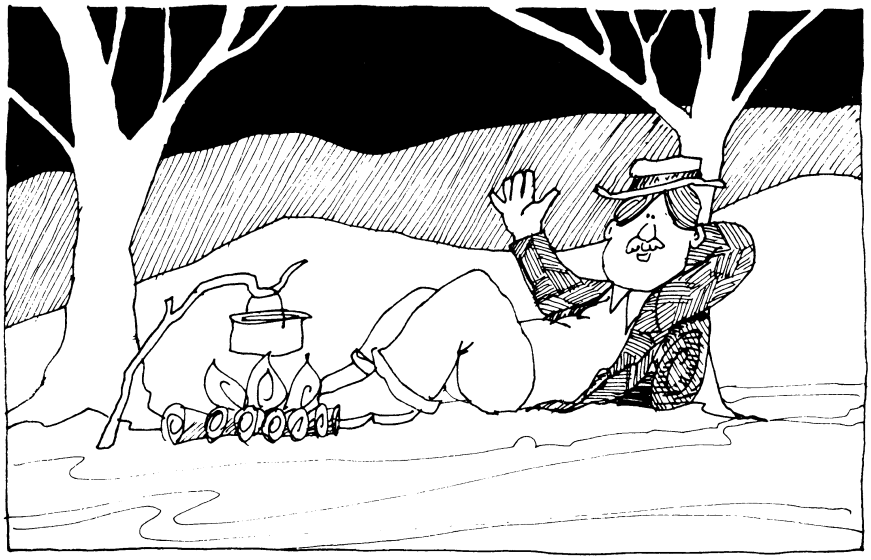
Measurement is the process of turning concepts like length or employment status into precisely defined variables. Using a tape measure to turn the idea of "length" into a number is straightforward because we know exactly what we mean by length. Measuring college readiness is controversial because it 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 free time requires that we first say what time counts as "free." 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 6 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, a person must die within 30 days of an accident to count as a traffic death. These details are a nuisance, but they can make a difference.

EXAMPLE 1.12 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, she must be available for work and looking for work. The unemployment rate is

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

yatesch01pg01_62 11/19/03 5:41 PM Page 26



"Unemployed? Not me, I'm out of 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 used 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 1.1 on the facing page 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 in 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 change to make politicians unhappy.

yatesch01pg01_62 11/19/03 5:41 PM Page 27

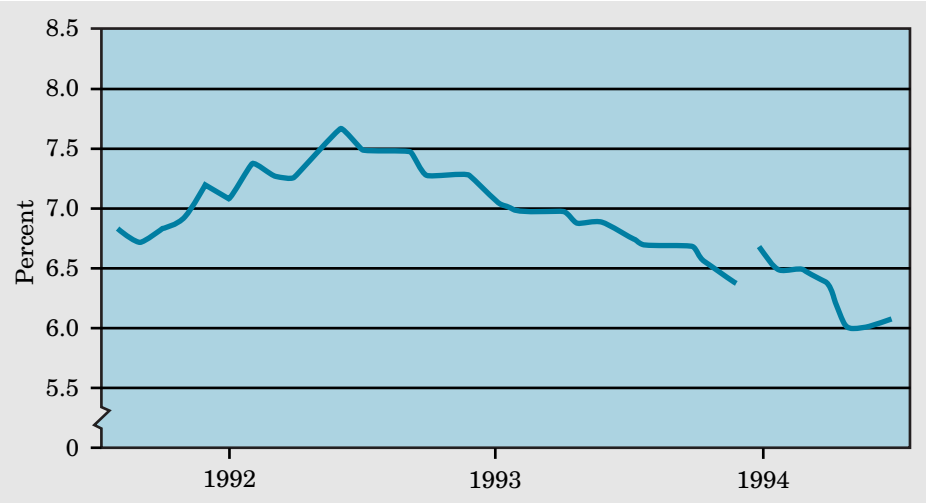


FIGURE 1.1 The unemployment rate from August 1991 to July 1994. The gap shows the effect of a change in how the government measures unemployment.

EXERCISES

1.24 Measuring pain The Department of Veterans Affairs (DVA) offers medical care to 3.4 million patients. 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.
(a) For the four existing vital signs, identify the measuring instrument used, the units of measurement, and the variable.
(b) Help out the DVA: how would you measure a patient's pain?

1.25 Measuring life's quality Is life in Britain getting better or worse? The usual government data don't say. So the British government announced that it wants 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? Be sure to identify the instrument and units for each variable you suggest.

1.26 Lightning strikes You can measure your distance from a lightning flash by counting the number of seconds between the flash and the thunderclap. Divide your time by 5, and you have the approximate distance in miles.
(a) Identify the instrument, units, and variable in this setting.
(b) Do you think this is a valid way to measure the distance to the lightning strike? Why or why not?
(c) How accurate do you think the measurements would be? Explain.

yatesch01pg01_62 11/20/03 9:05 AM Page 28

1.27 The heat is on According to *The Boy Scout Handbook,* you can measure the temperature of a fire using a "hand thermometer." It says, "Hold your palm where the food will be cooking. . . . Count 'one-and-one, two-and-two, . . . ,' and so on, for as many seconds as you can hold your hand still." Based on the number of seconds counted, you can determine the temperature. For example, 2 or 3 seconds would indicate a temperature of 400 to 450 °F.

(a) Identify the instrument, units, and variable in this setting.

(b) Do you think this is a valid way to measure temperature? Why or why not?

(c) How accurate do you think the measurements would be? Explain.

What are your units?

Not paying attention to units of measurement can get you into trouble. In 1999, the *Mars Climate Orbiter* spacecraft 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 operated the spacecraft thought the numbers were in metric system units (kilograms, kilometers). There went $125 million.

1.28 The BLS Go to the Bureau of Labor Statistics Web site: www.bls.gov.

(a) What is the current unemployment rate?

(b) The BLS is also responsible for measuring the consumer price index (CPI) each month. What is the current CPI? What does this number mean?

(c) Investigate at least one other variable that the BLS measures. Describe what you find in a few sentences.

Valid and invalid measurements

No one would object to using a tape measure reading in centimeters to measure the length of my 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. 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

yatesch01pg01_62 11/19/03 5:41 PM Page 29

"This is our new college admissions sizer. I got the idea from the carry-on luggage sizers at the airport."

somewhat different measure. Let's think about measures, valid and invalid, in some other settings.

EXAMPLE 1.13 Measuring highway safety

Roads get better. Speed limits increase. Big SUVs replace cars. Enforcement campaigns reduce drunk driving. How has highway safety changed over time in this changing environment?

We could just count deaths from motor vehicles. The Fatal Accident Reporting System says there were 41,508 deaths in 1991 and 41,945 deaths 10 years later in 2000. But the number of licensed drivers rose from 169 million in 1991 to 191 million in 2000. The number of miles that people drove rose from 2172 billion to 2750 billion. If more people drive more miles, there may be more deaths even if the roads are safer. The count of deaths is not a valid measure of highway safety.

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 2000, vehicles drove 2,750,000,000,000 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 2000, this death rate was

$$\frac{\text{motor vehicle deaths}}{\text{100s of million of miles driven}} = \frac{41{,}945}{27{,}500} = 1.53$$

yatesch01pg01_62 11/19/03 5:41 PM Page 30

The death rate fell from 1.91 deaths per 100 million miles in 1991 to 1.53 in 2000. That's a big change—there were 20% fewer deaths per mile driven in 2000 than a decade earlier. Driving has been getting safer.

Rates and counts

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

Using height to measure readiness for college, and even 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 1.14 Achievement tests

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

EXAMPLE 1.15 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 that no one instrument can measure.

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

yatesch01pg01_62 11/19/03 5:41 PM Page 31

Statistics is little help in these examples. They start with an idea like "knowledge of statistics" 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 1.16 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."[15] 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. 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.

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.

yatesch01pg01_62 11/19/03 5:41 PM Page 32

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." Application 1.2A uses one of these descriptions to help give you the big picture. It appears that SAT scores do have moderate predictive validity, and that they are about equally valid for different groups of students. Differences among groups in SAT scores generally reflect unequal environments that bring about unequal preparation for college.

APPLICATION 1.2A SAT Exams in College Admissions

Colleges use a variety of measures to make admissions decisions. The student's record in 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 a math SAT score of 620 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.

How well do SAT scores predict first-year college grades? The table below gives some results from a sample of 48,039 students:[16]

	All students	Men only	Women only	Black students
SAT	27%	26%	31%	25%
School grades	29%	28%	28%	24%
Both together	37%	36%	38%	34%

The numbers in the table say what percent of the variation among students in college grades can be predicted by SAT scores (math and verbal combined), by high school grades, and by SAT scores and high school grades together. An entry of 0% would mean no predictive validity, and 100% would mean predictions were always exactly correct. Use the table above to answer the following questions.

1. Do SAT scores predict college grades as well as high school grades do? Explain.
2. Does combining SAT scores and high school grades help predict college grades better than using either variable alone? Justify your answer.
3. Do SAT scores and high school grades predict college grades about equally well for males, females, and black students? Explain.

yatesch01pg01_62 11/19/03 5:41 PM Page 33

APPLICATION 1.2A SAT Exams in College Admissions *(continued)*

4. How effective are SAT scores and high school grades in predicting college grades?
5. In addition to SAT scores and high school grades, what other characteristics (variables) do you think should be taken into consideration in predicting success in college?

Selective colleges are justified in paying some attention to SAT scores, but they are also justified in looking beyond SAT scores for the motivation that can bring success to students with weaker academic preparation. The SAT debate is not really about the numbers. It is about how colleges should use all the information they have in deciding whom to admit, and also about the goals colleges should have in forming their entering classes.

EXERCISES

1.29 Counting the unemployed? We could measure the extent of unemployment by a count (the number of people who are unemployed) or by giving a rate (the percent of the labor force that is unemployed). The number of people in the labor force grew from 107 million in 1980, to 126 million in 1990, and to 140 million at the beginning of 2000. Use these facts to explain why the count of unemployed people is not a valid measure of the extent of unemployment.

1.30 Measuring intelligence "Intelligence" means something like "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'?" or "Who won the last soccer World Cup?"

1.31 School bus safety The National Highway Traffic Safety Administration says that an average of 11 children die each year in school bus accidents, and an average of 600 school-age children die each year in auto accidents during school hours. These numbers suggest that riding the bus is safer than driving to school with a parent. The counts aren't fully convincing, however. What rates would you like to know to compare the safety of buses and private autos?

1.32 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

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.

yatesch01pg01_62 11/19/03 5:41 PM Page 34

a broad range of jobs. As in the case of the SATs, 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.

1.33 Fighting cancer Congress wants the medical establishment to show that progress is being made in fighting cancer. Some variables that might be used are the following:
(a) Total deaths from cancer. These have risen sharply over time, from 331,000 in 1970, to 505,000 in 1990, and to 553,000 in 2000.
(b) The percent of all Americans who die from cancer. The percent of deaths due to cancer rose steadily from 17.2% in 1970 to 23.5% in 1990 but then leveled off to 23.0% in 2002.
(c) The percent of cancer patients who survive for 5 years from the time the disease was discovered. These rates are rising slowly. For whites, the 5-year survival rate was 50.9% in the 1974 to 1979 period and 62.0% from 1989 to 1997.

None of these variables are fully valid as a measure of the effectiveness of cancer treatment. Explain why both (a) and (b) could increase even if treatment is getting more effective, and why (c) could increase even if treatment is getting less effective.

1.34 Online IQ tests Search for an online IQ test. When you find one that will take no more than about 15 minutes, take the test. Write a few sentences about whether you think the test provides a valid measurement of your intelligence.

Accurate and inaccurate measurements

Using a bathroom scale to measure your weight is valid. If your scale is like mine, however, the measurement may not be very accurate. Think about my bathroom scale. It measures my weight, but it may not give my true weight. My scale reads 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: it reads 3 pounds too high because its aim is off. But it is also erratic. Most scales vary a bit—they don't always give the same reading when you step off and step right back on. My scale is somewhat old and rusty. This morning it sticks a bit and reads one-half pound too low for that reason. So the reading is

$$\text{measured weight} = \text{true weight} + 3 \text{ pounds} - 0.5 \text{ pound}$$

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

$$\text{measured weight} = \text{true weight} + 3 \text{ pounds} + 0.25 \text{ pound}$$

yatesch01pg01_62 11/19/03 5:41 PM Page 35

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

My scale has two kinds of errors. If it didn't stick, the scale would always read 3 pounds high. That would be true every time anyone stepped on the scale. A systematic error that occurs every time we make a measurement is called *bias.* My scale also sticks—but how much this changes the reading is different 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 measurements this way:

measured value = true value + bias + random error

A measurement process has **bias** if it systematically overstates or understates 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.**

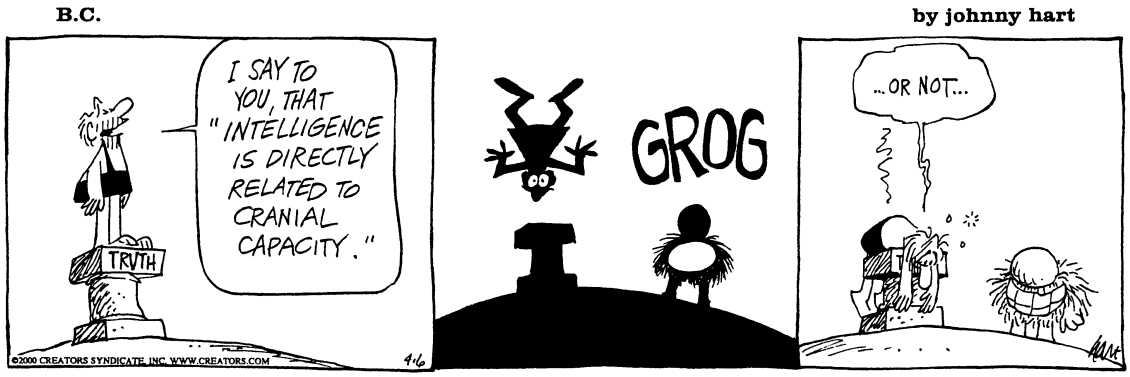
A scale that always reads the same when it weighs the same item is perfectly reliable even if it is biased. Reliability only says that the result is repeatable. Bias and lack of reliability are different kinds of error. And don't confuse reliability with validity just because both sound like good qualities. 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 1.17 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 and then pouring out the shot and weighing it gave quite reliable measurements of the skull's volume. These accurate

yatesch01pg01_62 11/19/03 5:41 PM Page 36

measurements do not, however, give a valid measure of intelligence. Skull volume turned out to have no relation to intelligence or achievement.



By permission of Johnny Hart and Creators Syndicate, Inc.

ACTIVITY 1.2B How long is a minute?

Materials: *Stopwatch for each pair of students*

Some people say that a minute seems like a long time. Others say, "Time flies." How well can you determine a minute?

1. You and your partner should sit as far as possible from other students in the class. Make sure you cannot see a clock from your position. Remove your watches.

2. You will take turns timing and measuring. The timer tells the measurer when to begin. When the measurer believes a minute has passed, she should quietly say, "Stop." At that point, the timer should stop the stopwatch and record the time that has passed to the nearest *tenth* of a second. Do not tell your partner how much time actually passed! Reset the stopwatch and switch roles.

3. Continue timing and measuring until each partner has measured a minute three times. Do not share your results until everyone else in the class has finished.

4. Examine your three measurements. How close did you get to a minute? Were all of your errors in the same direction? Find your average measurement.

5. Now look at your partner's data. Answer the same questions you did in Step 4.

6. Share data with your classmates. Which student was most accurate? Who was most reliable?

Extension: Analyze the entire class's measurements by making a graph. How accurate was the class as a whole? How reliable were the class's measurements?

yatesch01pg01_62 11/19/03 5:41 PM Page 37

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. Time starts with the earth's path around the sun, which lasts one year. But the earth is much too erratic. Since 1967, time starts with the standard second, and the second is defined to be the time required for 9,192,631,770 vibrations of a cesium atom. Physical clocks are bothered by changes in temperature, humidity, and air pressure. The cesium atom doesn't care. People who need really accurate time can buy atomic clocks. The National Institute of Standards and Technology (NIST) keeps an even more accurate atomic clock and broadcasts the results (with some loss in transmission) by radio, telephone, and the Internet.

EXAMPLE 1.18 Really accurate time

NIST's atomic clock is very accurate, but not perfectly accurate. The world standard is Universal Coordinated 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) by how much it misses the correct time. Here are the last 10 errors as we write, in seconds:[17]

0.000000007	0.000000000
0.000000005	−0.000000003
0.000000006	−0.000000005
0.000000000	−0.000000001
0.000000002	−0.000000001

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 numbers above 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. Averaging over more measurements reduces variation in the final result.

yatesch01pg01_62 11/19/03 5:41 PM Page 38

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.

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 1.2) is accurate to 1 second in 6 million years and is a bit large to put beside your bed.

FIGURE 1.2 This atomic clock at the National Institute of Standards and Technology is accurate to 1 second in 6 million years. (Photo courtesy of John Wessels, NIST Time and Frequency Division.)

EXAMPLE 1.19 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.

yatesch01pg01_62 11/19/03 5:41 PM Page 39

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.

EXERCISES

1.35 Measuring crime, I 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 National Crime Victimization Survey publishes data based on a national sample of more than 43,000 households. The victim survey shows about two and a half 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.)

1.36 Measuring crime, II Each year, the National Crime Victimization Survey asks a sample of more than 43,000 households whether they have been victims of crime and, if so, the details. In all, more than 80,000 people answer these questions. 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?

1.37 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.

1.38 Where to live? Each year, *Money* magazine ranks the 300 largest metropolitan areas in the United States in an article on the best places to live. First place in 1997 went to Nashua, New Hampshire. Nashua was ranked 42nd in 1996 and 19th in 1995. Monmouth and Ocean Counties in New Jersey were ranked 167th in 1995, 38th in 1996, and 3rd in 1997. Are these facts evidence that *Money*'s ratings are invalid, biased, or unreliable? Explain your choice.

yatesch01pg01_62 11/19/03 5:41 PM Page 40

1.39 In the trenches Two investigators measure the depth of an oceanic trench. The first one makes 20 independent measurements and reports the average of the values obtained. The second person reports the results of a single measurement using the same process.
(a) Which investigator will have more reliability in the number reported, or will they have about the same reliability?
(b) Which investigator will have more bias in the number reported, or will they have about the same bias?

1.40 Healthy babies Apgar scores are a measurement of an infant's overall health taken a few minutes after birth. The score ranges from 0 (dead) to 10 ("perfect health") and is based on tests of the baby's heart and breathing rate, muscle tone, and other criteria. A critic gives three reasons why the Apgar score isn't a perfect measurement:
Reason I: There are many important facets of health that aren't measured by the score.
Reason II: A doctor's rating may be affected by being present at the birth; a doctor may give an unwarranted low value to a baby whose birth was difficult, for example.
Reason III: Two different doctors may give different Apgar scores, even when measuring the same baby at the same time.
(a) Which of these criticisms concerns the validity of the Apgar score? Which one concerns its reliability? Which one concerns the bias in the measurement?
(b) Suppose two doctors both judge an infant's health using the Apgar system, and the average of their two values is taken as the "official" Apgar score. Will this improve the validity of the measurement? The reliability of the measurement? Reduce the bias of the measurement?

1.41 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.

APPLICATION 1.2B MERMAID

Mermaid is a system for monitoring pollution in coastal waters, estuaries, rivers, and lakes. It was developed by the GKSS Research Center in Geesthacht, Germany. (MERMAID stands for Marine Environmental Remote-Controlled Measuring and Integrated Detection.) Of course, the concept of "pollution" is very complex. So the MERMAID system measures dozens of different variables such as water pH, phosphate and oxygen levels, and temperature. A MERMAID system is actually a collection of modules that transmit signals to a land-based station from an ocean platform, a buoy, or even an

yatesch01pg01_62 11/19/03 5:41 PM Page 41

APPLICATION 1.2B MERMAID *(continued)*

unmanned ship. (The picture shows a MERMAID platform in the Wadden Zee, the Netherlands.)

The GKSS Research Center is very particular about the accuracy of the devices included in the MERMAID system. For example, when the phosphate-measuring module was used many times on the same water sample, the results were so consistent that they varied from each other by only a few parts per billion. Also, the center compared their automatic measuring devices with the most modern "gold-standard" method of conducting each of the chemical analyses and found that the results were essentially identical.

1. From the above paragraph, explain in your own words how the GKSS Research Center is addressing each of the following issues in its MERMAID systems:

(a) Measurement reliability

(b) Bias

(c) Validity

2. In what units are each of the variables mentioned above measured? You may need to visit the MERMAID Web site to be sure (w3g.gkss.de/mermaid).

EXPLORING THE WEB

You can get the time directly from the atomic clock at NIST at www.time.gov. There is some error due to Internet delays, but the display even tells you roughly how accurate the time on the screen is. For a glimpse of the elaborate system that lies behind everyday measurements of length, weight, and time, visit the Web site of the International Bureau of Weights and Measures, www.bipm.fr.

yatesch01pg01_62 11/19/03 5:41 PM Page 42

STATISTICS IN SUMMARY

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. 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. Ask if the variables are **valid** as numerical measures of the concepts the study discusses. 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 measurements** that reduce the value of the data. You can think about errors in measurement like this:

measured value = true value + bias + 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.** You can improve the reliability of a measurement by repeating it several times and using the average result.

SECTION 1.2 EXERCISES

1.42 Old trees The age of a pine tree was measured five times using a new electronic probe inserted in the tree's trunk. The measured values were 43, 40, 45, 44, and 41 years old. Later this tree was cut down and, by counting the growth rings, it was determined that the tree was really 34 years old. Does this new electronic device for measuring the age of trees have a greater problem with bias or with reliability? Explain.

1.43 Capital punishment Between 1977 and 2001, 749 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 2000 population of these states:

State	Population (thousands)	Executions
Alabama	4,447	23
Arkansas	2,673	24
Florida	15,982	51
Missouri	5,595	53
Nevada	1,998	9
Texas	20,852	256
Virginia	7,079	83

yatesch01pg01_62 11/19/03 5:41 PM Page 43

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

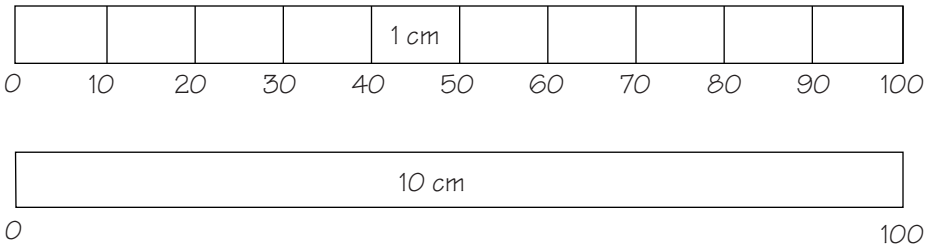
$$\text{rate per million} = \frac{\text{executions}}{\text{population in thousand}} \times 1000$$

Arrange the states in order of the number of executions relative to population. Are Florida and Texas still high by this measure?

1.44 Measuring the moon The diameter of the moon is measured four times by a process that is free of bias. The individual measurements are 2157, 2166, 2162, and 2155 miles, which average out to 2160 miles. One more measurement is about to be taken using the same process. When compared with the estimate of 2160 miles, would you expect this next measurement to be more accurate, just as accurate, or less accurate as a measure of the true diameter of the moon? Explain.

1.45 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?

1.46 Measuring pain Following surgery, patients were given medication to relieve pain. They were then examined at regular intervals. The patients were asked to assess the level of pain they were experiencing at each of the exam times. A pain score was based on the patients' marks on a 10-centimeter line with a range from 0 (no discomfort) to 100 (agonizing pain).[18]
(a) Is the pain score a valid measurement of the amount of pain felt by the patients? Explain your answer.
(b) Consider the two 10-centimeter scales given below. Discuss the reliability of the data collected using each of these scales.

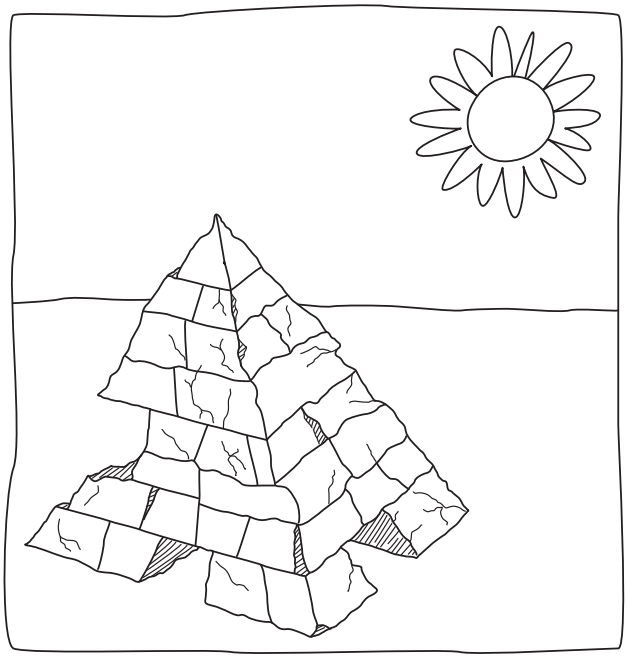


yatesch01pg01_62 11/19/03 5:41 PM Page 44

1.47 Teaching English A teacher wishes to know if a new English curriculum will increase the creativity of her students in writing poetry. At the end of a unit, a sample of each student's poetry is given to a panel of five experts who rate the creativity of the poem on a 10-point scale. For each of the statements below, tell whether it addresses the issue of measurement reliability or measurement validity.
Statement I: In reporting the results, the teacher claimed that the experts' ratings showed just how creative her students were. A critic of the new curriculum argued that the ratings were a poor way to measure creativity.
Statement II: The five experts seemed to rate each student nearly the same. For example, one student's poem got ratings of 8, 7, 8, 8, and 6 by the five experts.

1.48 Measuring good teaching You want to measure the "effectiveness" of teachers at your high school. Give an example of a clearly invalid way to measure good teaching. Then briefly describe a measurement process that you think is valid.

1.49 Activity 1.1 follow-up Refer to Activity 1.1 (page 4).
(a) How reliable are the individual measurements made by the members of the class? Explain briefly.
(b) If every measurement in the class were made by lining up one edge of the textbook with the 1 mark on the ruler rather than the 0 mark, would this affect the reliability of the measurements? Would it affect the bias of the measurements?



If the Pharaoh's architect had used the ruler from statistics class. . .