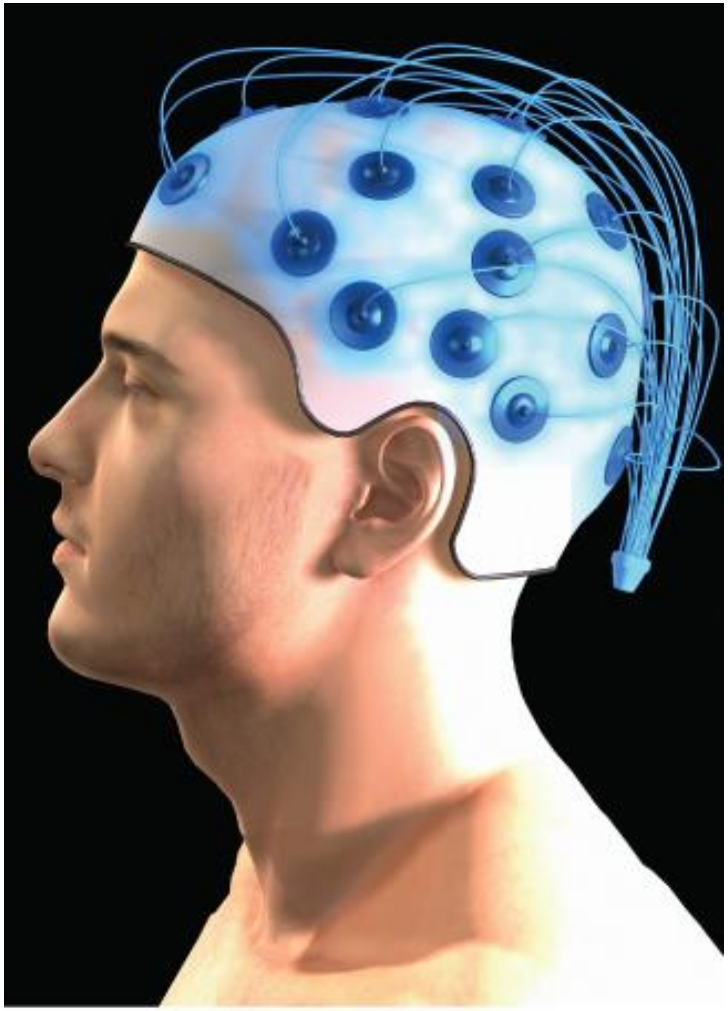
Chapter 8

Measuring

Lecture Slides

Case Study: Measuring 1

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.



MedicalRF.com/Corbis

Case Study: Measuring 2

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 or a formula that includes grades in school?

Case Study: Measuring 3

It is hard to say exactly what intelligence is and difficult to attach a number to measure it. In the end, can we 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 1

Statistics deals with data, and the data may or may not be numbers.

Once we have our sample respondents or our experimental subjects, we must 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?

Once we define the variables we want to measure, then we can decide how to measure them.

Measurement Basics 2

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

Example: Length, college readiness, highway safety 1

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**.

Example: Length, college readiness, highway safety 2

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.** They are determined by a complicated scoring system described at the SAT website (www.collegeboard.com).

Example: Length, college readiness, highway safety 3

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.

Measurement Basics 3

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?

Know Your Variables 1

Measurement is the process of turning concepts such as 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.

Know Your Variables 2

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?

Example: Measuring unemployment 1

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.

Example: Measuring unemployment 2

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

The BLS has very detailed descriptions of what it means to be "in the labor force" and what it means to be "employed."

- If you are on strike but expect to return to the same job, you are 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.

Measurements, valid and invalid 1

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. Why? Because height has nothing to do with being prepared for college.

We say height is not a **valid measure** of a student's academic background.

Measurements, valid and invalid 2

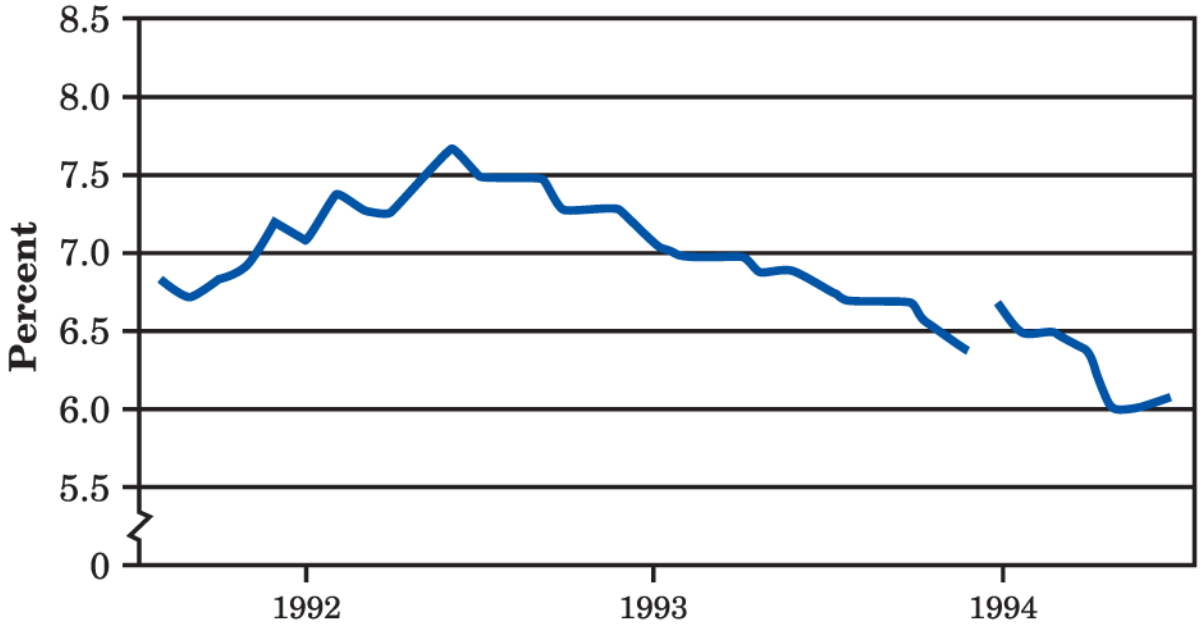
A variable is a **valid** measure of a property if it is relevant or appropriate as a representation of that property.

It is valid to measure length with a tape measure.

It is not 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.

Measurements, valid and invalid 3



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

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

Example: Measuring highway safety 1

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.

Mahaux Photography/Getty Images

How did highway safety change between 2007 and 2012 in this changing environment?

We could just count deaths from motor vehicles.

Example: Measuring highway safety 2

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. However, we need to keep in mind other things that happened during this same time frame to determine how much progress has been made.

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.

Example: Measuring highway safety 3

The count of deaths alone 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 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.

Example: Measuring highway safety 4

For 2012, this death rate is

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

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.

Measurements, Valid and Invalid 4

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

Measurements, Valid and Invalid 5

Is the SAT a valid measure of readiness for college?
"Readiness for college academic work" is a vague concept that probably combines many factors.

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?

Measurements, Valid and Invalid 6

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.

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.

Measurements, Valid and Invalid 7

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?"

But, we must ask how *accurately* SAT scores predict college grades.

Measurements, accurate and inaccurate 1

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 three pounds too high. Then,

Measured weight = true weight + three pounds

Measurements, accurate and inaccurate 2

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 three pounds too high because its aim is off, but now it sticks a bit and reads one pound too low for that reason. Then,

Measured weight = true weight + three pounds – one pound

Measurements, accurate and inaccurate 3

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

Measured weight = true weight + three pounds + one 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

measured weight = true weight + three pounds − 1.5 pounds

Measurements, accurate and inaccurate 4

Your scale has two kinds of errors.

If it didn't stick, the scale would always read three 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. The result is that the scale weighs three 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**.

Measurements, accurate and inaccurate 5

We can think about errors in measurement this way:

measured value = true value + bias + 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.**

Measurements, accurate and inaccurate 6

To determine if the random error is small, compute 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.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{where} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

A reliable measurement process will have a small variance.

Measurements, accurate and inaccurate 7

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.

Bias and lack of reliability are different kinds of error.

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. For example, if a scale weighs a 200-pound man as 198, 202, 205, and 195 pounds, the scale is valid but not reliable.

Example: Do big skulls house smart brains? 1

In the mid-19th 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.

Example: Do big skulls house smart brains? 2

These accurate measurements do not, however, give a valid measure of intelligence.

Skull volume turned out to have no relation to intelligence or achievement.

Paul Broca's measuring process was reliable, but not valid.

Improving reliability, reducing bias 1

Scientists everywhere repeat their measurements and use the average to get more reliable results.

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.

Improving reliability, reducing bias 2

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.

Example: Measuring unemployment again 1

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 re-interview 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.

Example: Measuring unemployment again 2

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 1

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.

Pity the poor psychologist 2

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.

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: Authoritarian personality? 1

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.”

Example: Authoritarian personality? 2

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.

Pity the poor psychologist 3

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.

Pity the poor psychologist 4

2. The F in “F-scale” stands for Fascist. 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?

Pity the poor psychologist 5

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. So 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?

Pity the poor psychologist 6

Psychologists lack answers to these questions. The first two are controversial because not all psychologists think about human personality in the same way. The last two questions do not have simple answers.

“Measurement” seems straightforward when we measure weight but is complicated when we try to measure human personality.

Pity the poor psychologist 7

Be wary of statistical “facts” about squishy topics like authoritarian personality, intelligence, and readiness for college.

The numbers look solid, 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 1

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.

Statistics in Summary 2

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.

Statistics in Summary 3

Also ask if there are **errors in measurement** 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.

Statistics in Summary 4

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.