Statistical Thinking for the 21st Century

Chapter 2

Working with data

2.3 What makes a good measurement?

In many fields such as psychology, the thing that we are measuring is not a physical feature, but instead is an unobservable theoretical concept, which we usually refer to as a *construct*. For example, let's say that I want to test how well you understand the distinction between the four different scales of measurement described above. I could give you a pop quiz that would ask you several questions about these concepts and count how many you got right. This test might or might not be a good measurement of the construct of your actual knowledge — for example, if I were to write the test in a confusing way or use language that you don't understand, then the test might suggest you don't understand the concepts when really you do. On the other hand, if I give a multiple choice test with very obvious wrong answers, then you might be able to perform well on the test even if you don't actually understand the material.

It is usually impossible to measure a construct without some amount of error. In the example above, you might know the answer but you might mis-read the question and get it wrong. In other cases there is error intrinsic to the thing being measured, such as when we measure how long it takes a person to respond on a simple reaction time test, which will vary from trial to trial for many reasons. We generally want our measurement error to be as low as possible.

Sometimes there is a standard against which other measurements can be tested, which we might refer to as a "gold standard" — for example, measurement of sleep can be done using many different devices (such as devices that measure movement in bed), but they are generally considered inferior to the gold standard of polysomnography (which uses measurement of brain waves to quantify the amount of time a person spends in each stage of sleep). Often

the gold standard is more difficult or expensive to perform, and the cheaper method is used even though it might have greater error.

When we think about what makes a good measurement, we usually distinguish two different aspects of a good measurement.

2.3.1 Reliability

Reliability refers to the consistency of our measurements. One common form of reliability, known as "test-retest reliability", measures how well the measurements agree if the same measurement is performed twice. For example, I might give you a questionnaire about your attitude towards statistics today, repeat this same questionnaire tomorrow, and compare your answers on the two days; we would hope that they would be very similar to one another, unless something happened in between the two tests that should have changed your view of statistics (like reading this book!).

Another way to assess reliability comes in cases where the data includes subjective judgments. For example, let's say that a researcher wants to determine whether a treatment changes how well a child interacts with other children, which is measured by having experts watch the child and rate their interactions with the other children. In this case we would like to make sure that the answers don't depend on an unreliable individual rater — that is, we would like for there to be high *inter-rater reliability*. Inter-rater reliability can be assessed by having more than one rater perform the rating, and then comparing their ratings to make sure that they agree well with one another.

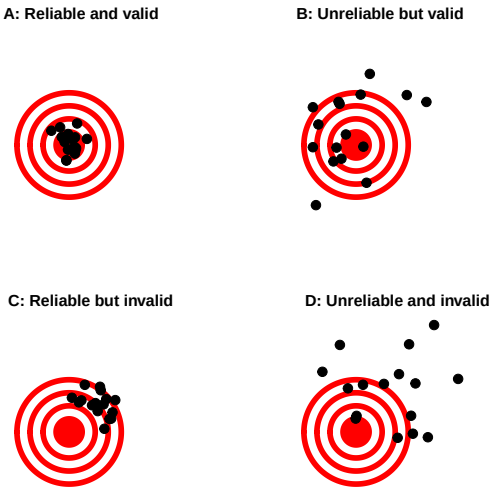

Figure 2.1: A figure demonstrating the distinction between reliability and validity, using shots at a bullseye. Reliability refers to the consistency of location of shots, and validity refers to the accuracy of the shots with respect to the center of the bullseye.

2.3.2 Validity

Reliability is important, but on its own it's not enough: After all, I could create a perfectly reliable measurement on a personality test by re-coding every answer using the same number, regardless of how the person actually answers. We want our measurements to also be *valid* — that is, we want to make sure that we are actually measuring the construct that we think we are measuring (Figure 2.1). There are many different types of validity that are commonly discussed; we will focus on three of them.

Face validity. Does the measurement make sense on its face? If I were to tell you that I was going to measure a person's blood pressure by looking at the color of their tongue, you would probably think that this was not a valid measure on its face. On the other hand, using a blood pressure cuff would have face validity. This is usually a first reality check before we dive into more complicated aspects of validity.

A construct is an idea that cannot be directly observed.

Construct validity. Is the measurement related to other measurements in an appropriate way? This is often subdivided into two aspects. *Convergent validity* means that the measurement should be closely related to other

measures that are thought to reflect the same construct. Let's say that I am interested in measuring how extroverted a person is using a questionnaire or an interview. Convergent validity would be demonstrated if both of these different measurements are closely related to one another. On the other hand, measurements thought to reflect different constructs should be unrelated, known as *divergent validity*. If my theory of personality says that extraversion and conscientiousness are two distinct constructs, then I should also see that my measurements of extraversion are *unrelated* to measurements of conscientiousness.

Predictive validity. If our measurements are truly valid, then they should also be predictive of other outcomes. For example, let's say that we think that the psychological trait of sensation seeking (the desire for new experiences) is related to risk taking in the real world. To test for predictive validity of a measurement of sensation seeking, we would test how well scores on the test predict scores on a different survey that measures real-world risk taking.