1 Measurement Scales

Statistics operate on a *data set*. The data set may be viewed as a two dimensional matrix, very similar to a blank spreadsheet found in many contemporary software packages such as Excel. The rows of the data matrix are *observations*. In neuroscience, observations are usually organisms (humans, rats, mice) but occasionally they may be other phenomena such as cell cultures. The columns of the data matrix consist of attributes—sex, parietal lobe activity in a PET (positron emission tomography) scan, number of bar presses—measured on the observations. This chapter explains *measurement scales*, the different mathematical classes for the attributes.

1.1 Measurement Scales: Traditional Classification

Statisticians call an attribute on which observations differ a *variable*. The type of unit on which a variable is measured is called a *scale*. Traditionally, statisticians talk of four types of measurement scales: (1) *nominal*, (2) *ordinal*, (3) *interval*, and (4) *ratio*.

1.1.1 Nominal Scales

The word *nominal* is derived from *nomen*, the Latin word for name. Nominal scales merely name differences and are used most often for qualitative variables in which observations are classified into discrete groups. The key attribute for a nominal scale is that there is no inherent quantitative difference among the categories. Sex, religion, and race are three classic nominal scales used in the behavioral sciences. Taxonomic categories (rodent, primate, canine) are nominal scales in biology. Variables on a nominal scale are often called *categorical* variables.

1.1.2 Ordinal Scales

Ordinal scales rank-order observations. Class rank and horse race results are examples. There are two salient attributes of an ordinal scale. First, there is an underlying quantitative measure on which the observations differ. For class rank, this underlying quantitative attribute might be composite grade point average, and for horse race results it would be time to the finish line. The second attribute is that individual differences individual on the underlying quantitative measure are either unavailable or ignored. As a result, ranking the horses in a race as 1^{st}, 2^{nd}, 3^{rd}, etc. hides the information about whether the first-place horse won by several lengths or by a nose.

There are a few occasions in which ordinal scales may be preferred to using a quantitative index of the underlying scale. College admission officers, for example, favor class rank to overcome the problem of the different criteria used by school districts in calculating GPA. In general, however, measurement of the underlying quantitative dimension is preferred to rank-ordering observations because the resulting scale has greater statistical power than the ordinal scale.

1.1.3 Interval Scales

In ordinal scales, the interval between adjacent values is not constant. For example, the difference in finishing time between the 1st place horse and the 2nd horse need not the same as that between the 2nd and 3rd place horses. An interval scale has a *constant interval* but lacks a true 0 point. As a result, one can add and subtract values on an interval scale, but one cannot multiply or divide units.

Temperature used in day-to-day weather reports is the classic example of an interval scale. The assignment of the number 0 to a particular height in a column of mercury is an arbitrary convenience apparent to everyone anyone familiar with the difference between the Celsius and Fahrenheit scales. As a result, one cannot say that 30° C is twice as warm as 15° C because that statement involved implied multiplication. To convince yourself, translate these two into Fahrenheit and ask whether 86° F is twice as warm as 50° F.

Nevertheless, temperature has constant intervals between numbers, permitting one to add and subtract. The difference between 28° C and 21° C is 7 Celsius units as is the difference between 53° C and 46° C. Again, convert these to Fahrenheit and ask whether the difference between 82.4° F and 69.8° F is the same in Fahrenheit units as the difference between 127.4° F and 114.8° F?

1.1.4 Ratio Scales

A ratio scale has the property of equal intervals but also has a true 0 point. As a result, one can multiply and divide as well as add and subtract using ratio scales. Units of time (msec, hours), distance and length (cm, kilometers), weight (mg, kilos), and volume (cc) are all ratio scales. Scales involving division of two ratio scales are also themselves ratio scales. Hence, rates (miler per hour) and adjusted volumetric measures (mg/dL) are ratio scales. Note that even though a ratio scale has a true 0 point, it is possible that the nature of the variable is such that a value of 0 will never be observed. Human height is measured on a ratio scale but every human has a height greater than 0. Because of the multiplicative property of ratio scales, it is possible to make statements that 60 mg of fluoexetine is three times as great as 20 mg.

1.2 Measurement Scales: Other Views

We presented the traditional classification of measurement scales because the terminology is widely used in statistics. Modern statisticians, however, recognize problems in this system and urge students to be cognizant of several other properties of measurement scales (Mosteller & Tukey, 1977; Velleman & Wilkinson, 1993). We review some of these properties here.

Before this review, however, we must point out a well-established principle about measurement scales and statistics. The order of the traditional measurement scales presented above—nominal, then ordinal, then interval, then ratio—is from weakest to strongest in terms of statistical inference. If there is a choice among measurement scales, then always select the *highest* (i.e., *strongest*) scale. Hence, an interval scale should be preferred to a nominal scale, an interval scale to an ordinal scale, and so on.

1.2.1 Continuous and Discrete Variables

A continuous variable has an infinite number of possible values between any two points on the measurement scale. For example, mouse weight will have an infinite number of possible values between 25 grams and 26 grams because one could always add extra decimal places to the measurement.

A discrete variable on the other hand can only take on a limited number of values. By their nature, all categorical variables are discrete, but so are many variables measured on ratio scales. One very important type of discrete variable measured on a ratio scale is a *count* such as the number of pups in a rat litter or number of correct responses on memory task. Counts are always positive integers.

1.2.2 Bounded Variables

All variables probably have mathematical bounds imposed by nature—the weight of an adult human brain, for example, has a lower and upper bound even though the exact numbers for these bounds may be unclear. The term *bounded variables*, on the other hand, refers to measurement scales with a mathematical boundary. Counts, for example, have a lower bound of 0 while percents have a lower bound of 0 as well as an upper bound of 100.

Bounded variables may—not do not *have to*—present problems for statistics. For example, many statistics apply the normal curve to data. The equation for the normal curve assumes that scores are symmetric around the mean and can range from negative infinity to positive infinity (although as scores deviate more and more from the mean, the probability of observing them becomes less and less). Many variables measured as percents, however, may have values that cluster close to 0 (or 100). In such cases, mathematical transformations of the original variables are used and the statistical analysis is performed on the transformed values.

1.2.3 Categorical versus Categorized Variables

True categorical variables place observations into groups in which there is no implication about differences in magnitude between the groups. A *dichotomous* variable has two mutually exclusive categories with no implication of a difference in magnitude between the categories. Sex (male versus female) is dichotomous as is virginity (either you are or you were). The term *binary variable* is synonymous with dichotomous variable. The phrase *polychotomous* refers to a categorical variable with more than two mutually exclusive classes.

Categorized variables, on the other hand, develop arbitrary—but often meaningful—classes from a variable that is inherently quantitative. A *dichotomized* variable is one that has an underlying quantitative scale in which an arbitrary cut point is used to divide observations into two classes. Classic examples are the cutoffs for cholesterol levels or blood pressure that are used in medicine to make treatment decisions about hypercholesterolemia and hypertension. The term *polychotomized* applies to two or more cut points that result in more than two ordered classes—e.g., blood pressure could give hypotensive, normotensive, and hypertensive groups.

As in most attempts at classification, there are large gray areas that are not well captured by distinguishing categorical from categorized variables. Cancer is clearly categorical but within cancer victims it is important to quantify the stage of the cancer. Psychiatric disorders provide another gray area. Many regard a diagnosis of antisocial personality disorder as a categorized variable reflecting an underlying continuum of prosocial to antisocial behavior, but the same consensus will not be found for schizophrenia.

A further gray area is the *ordered group variable*. An ordered group variable has one (or possibly more) underlying quantitative dimension(s) but lacks clear cutpoints to separates the groups. Division of professorial ranks into assistant, associate, and full professors is an example of an ordered group variable. Some classification systems in psychopathology also use ordered groups. For example, the "schizophrenia spectrum" may be defined as not affected, schizotypal personality disorder, and schizophrenia.

1.2.4 A Group is not a Group is not a Group

One of the most persistent statistical problems in medicine and neuroscience is a failure to distinguish categorical variables from categorized and ordered groups. There is a tendency to view "a group" much as Gertrude Stein viewed a rose—a group is a group is a group. Such a statement is tantamount to statistical heresy because the optimal procedures for analyzing categorical variables are *not* the same as those for categorized or ordered groups.

In truly categorical variables, *the ordering of the groups is completely immaterial*. For example, in a chart, the bar for males could come before or after the one for females. Now consider a dose-response study that administers either 0 mg, 10mg, 15mg, or 20mg of an active drug to a randomly chosen rat. Imagine the response from the editor and reviewers to a submitted journal article that contained a figure starting with a bar from the 20 mg group, followed by one from the 10 mg group, then the placebo, and finally the 15 mg group. Yet the statistical procedure used to test for differences among means often treats this ordering as just as perfectly fine and logical.

Even though there are four groups of rats, these groups are not categorical. They differ on a quantitative dimension and hence are categor*ized*. Both the ordering of the groups in the hypothetical figure as well as the statistics for testing means should reflect this fact. To follow the principle outlined at above, a variable using the ratio scale of 0, 10, 15, and 25 should be analyzed in place of a nominal variable that effectively treats any single ordering of the groups as just as logical and rational as any other possible ordering of the groups.

From a purely statistical viewpoint there is one overriding principle—*whenever possible, analyze the quantitative variable in place of the categorized group variable*. In the case of ordered groups, then use an ordinal scale. One will usually gain greater statistical power from the quantitative variable and quantitative variables can undercover important nonlinear relationships that are sometimes impossible to detect with group

variables[1]. We will have more to say about this matter later, but the principle is so important this it should be memorized here.

1.3 Psychometrics

Many areas of human neuroscience gather data using *psychometric instruments*—interviews, questionnaires, and tests. Two essential properties of psychometric instruments are their reliability and the validity.

1.3.1 Reliability

Reliability is defined as the *repeatability* of measurement. If one gave the same interview, questionnaire, or test on two occasions, then how well would the results on the first administration predict those of the second administration? If the level of prediction is high, then the psychometric instruments is said to be reliable. If predictability is poor, then the instrument is unreliable.

The time interval between the administrations of the measuring instruments depends on the nature of the trait under study and on the type of reliability. One type of reliability, termed *internal consistency reliability*, actually measures reliability for a single administration. Internal consistency reliability is appropriate only for scales where each item measures the same trait[2]. Here, the reliability consists of how well scores on, say, a subset of items predict scores on other subsets of items. When all the items on a single scale—e.g., the word in a vocabulary test—are purported to measure the same trait—vocabulary ability—then the scale should have high internal consistency reliability.

Some traits—mood, for instance—can change in a matter of hours. Hence, a neuroscientist interested in affect might examine the reliability of a mood measure using a time interval of an hour or two. Other traits—e.g., intelligence—might be expected to change little over the course of a week or two. Hence, the time interval required to assess the reliability of a measure of intelligence could be much greater than that for mood. Often, psychometricians will administer a different form of the test on each occasion to control for memory effects[3].

It is important to distinguish reliability from *trait stability*, although in practice the two concepts blend into each other rather than being discrete entities. The time interval for reliability testing depends on a theoretical implication and/or common-sense judgment about how quickly individual differences in trait may change for natural reasons. Stability, on the other hand, is an empirical question about how well one can predict trait scores over a long period of time. Reliability of measurement is a prerequisite for assessing trait stability. If the measure has poor reliability, then any assessment of stability must be questioned. High reliability, on the other hand, does not

---

[1] This principle applies to the analysis of data and should not be used to justify abandonment of categorized variables in clinical practice.

[2] Psychometricians term such scales *homogeneous scales*. Scales composed of items measuring different traits are called *heterogeneous scales*.

[3] The different forms of a test are referred to as *parallel forms*.

imply high stability. A measure of mood may be reliable, but the correlation between scores assessed now and a year from now may be low.

1.3.2 Validity

The validity of a psychometric instrument is defined as the extent to which the instrument measures what it purports to measure. The concept of test validity has a bewildering array of flavors, only some of which can be detailed here.

1.3.2.1 Content Validity

Content validity is defined as the extent to which there is common-sense agreement between an item on a scale and the trait that the scale claims to measure. For example, consider the following item: "the meaning of the word *garrulous* is _____." The item would have good content validity if the scale were a measure of vocabulary but poor content validity if the scale measured working memory.

1.3.2.2 Criterion Validity

Criterion validity is the extent to which scores on the instrument predict some preordained criterion measure usually—and sometimes, euphemistically—referred to as the "gold standard." Virtually all structured interviews designed to diagnose psychiatric disorder assess validity using this yardstick. Here, the results using a newly developed interview are compared to those using the best available interview (or some combination of interview and clinical judgment). If the results agree, then the new interview is said to be valid. Note that this procedure can result in the evolution of a "gold standard." The new interview, for example, could replace the old one as the preferred criterion for assessing validity.

1.3.2.3 Construct Validity

Construct validity (Cronbach & Meehl, 19xx) consists of the extent to which the instrument correlates with those variables that, according to theory and previous research, it should correlated with and does not correlate with other variables that it should not predict. For example, a measure of intelligence might be expected to predict grades in school and results from standardized academic achievement tests but not personality traits.

1.3.3 Relationship between Reliability and Validity

(Not written yet)