Pieces of Data

☺ If nothing else, please note the following: The word *data* is plural. The singular is *datum*. The only exception to all this is on TV (which I guess is now Hulu or Netflix or Amazon Prime), where *Data* is singular when it refers to the android on *Star Trek*. This is a serious pet-peeve.

Sets of data consist of two or more organized observations. An **observation** is a value (i.e., a number or a category code) that tells you something about something. (Technically, therefore, an observation is a proposition.) More generally, an observation specifies the value of a subject on some dimension. In statistical parlance, these dimensions are called *variables*.

Variables

A **variable** is a property, characteristic, or quality (of a creature, object, or situation) that can take on more than one value. The values of some variables differ mathematically, in which case the variable is called *quantitative* (since it specifies a quantity); other variables take on values that differ in ways that are categorical, in which case the variable is called *qualitative* (since it specifies a quality). Some classic examples from psychology are response time and percent correct for quantitative, and sex and race for qualitative. More on this below.

The quantitative vs. qualitative distinction is inherent to the variable and does not depend on how the variable is being used. Another way that variables can differ, however, does depend on the rôle that the variable is playing. On one side are **independent** or **manipulated variables**. These are properties, characteristics, or qualities that are entirely determined or set by the experimenter. For example, in an experiment concerning the effects of sleep-deprivation, number of hours of sleep -- if it is set by the experimenter -- is an independent variable. On the other side are **dependent** or **measured variables**. These are properties, characteristics, or qualities that are observed as they occur. In the case of sleep deprivation, for example, “grumpiness” on a ten-point scale might be the dependent variable. In all that follows, *independent variable* will be ***IV*** and *dependent variable* will be ***DV***.

☞ Note: IV and DV are not automatic labels for a given property, characteristic, or quality. Some variables appear as IVs in one study and DVs in another (e.g., hours of sleep: this can be the IV in a sleep-deprivation study or the DV in a study of the effects of “mental work”). It is crucial that you correctly identify how a given variable is being used in a given study. The best way to do this is to remember the alternative names for IV and DV and ask: was this variable *manipulated* or *measured* by the researcher?

⊕ Note that SPSS does not always use these same labels and sometimes uses them incorrectly. Most of all, SPSS likes to refer to the predictor variable(s) in regression as IVs, even when they were not manipulated.

There is also what appears to be a third class of variables that lies somewhere between IVs and DVs. These **subject variables** are properties, characteristics, or qualities that vary across research subjects, but are relatively stable within subjects (across time) or extremely difficult to manipulate. The classic examples are race and sex, but “higher-order” examples also exist, such as socio-economic status. In some analyses, SVs are treated as if they were IVs. Technically, this creates a quasi-experiment, as opposed to a “true” experiment (which has real IVs). In other analyses, SVs are used as covariates. More on this when we get to specific analyses. For now, just keep this third type of variables in mind.

Types of Data

Turning now to specifics, there are three main ways in which variables can differ: in terms of their ***scale***, ***precision***, and ***kind***. The first has already been mention; the second will probably be familiar; the third is almost unique to psychology (and the term is an invention of mine).

There are four types of ***scale***:

Nominal - the values differ qualitatively (e.g., race, sex, hair color, etc.)
Ordinal - the values have an order, but no particular spacing (e.g., birth order)
Interval - the values are equally spaced, but zero doesn’t mean *none* (e.g., temp in C or F)
Ratio - the values have equal spacing and zero means *none* (e.g., # of siblings; temp in K)

For the purposes of statistical analysis, these four scales can be divided into three important types: qualitative (i.e., nominals), ordinal, and quantitative (i.e., interval and ratio scales). In other words, no current analysis makes a distinction between interval and ratio scales, so these two can be treated as one.

⊕ Note: SPSS refers to qualitative, ordinal, and quantitative scales as *nominal*, *ordinal,* and *scale*, respectively, which is fine with me.

The ***precision*** of a set of data is determined by two things: the thing being measured and the method of measuring it. With regard to the thing being measured, there are *discrete* variables, where only some values are possible (e.g., number of children, which must be a whole number), and there are *continuous* variables, where any value between two end points is possible (e.g., response time, which can be any value between zero and infinity).

With regard to the method of measurement, discrete data can be left *as is* or collapsed into larger groups (e.g., 1-2 children, 3-4 children, etc.), which is often said to “lower the precision” of measurement. Lowering precision sounds bad (and it generally is), but collapsing is sometimes required, because certain analyses (e.g., chi-square) have a minimum number of observations per cell in order to be used. Otherwise, discrete data are rarely affected by the method of measurement, because only certain values are possible.

In contrast, the precision of continuous data is (almost by definition) completely determined by the measuring device and/or subsequent rounding. Response time, for example, is usually measured in milliseconds (i.e., rounded off to the nearest thousandth of a second), but is sometimes measured at lower precisions (e.g., tenths of a second) when the recording device has a slower clock speed or the values in milliseconds would be ridiculously large. Likewise, height of the experimenter (which is important in some studies, such as those concerning obedience to authority) is usually rounded off to centimeters, which is a lower precision than millimeters and a higher precision than decimeters. Note, also, that reporting a length or height in meters to two decimal places (e.g., 1.76 m) is the same as using centimeters, because it’s the preciseness that matters, not the units.

Because the above can be confusing, I suggest that we separate the two issues that determine precision and give them separate labels. In particular, I suggest that we refer to the discrete vs. continuous distinction as “class” and that we refer to measurement issues as “resolution.” For example, grouping people via 1-2 kids vs 3-4 kids, etc., would create a “low-resolution measure of the discrete class.” In contrast, measuring response time to the nanosecond would give us a “ridiculously-high-resolution measure of the continuous class.”

With all that said, the precision (class and/or resolution) of a variable has absolutely no effect on how the data are analyzed. It only has a small effect on interpretation and can also determine how the data should be plotted. The one exception to this is when wide-ranging, continuous data are dichotomized (via, e.g., a median split); this has a huge effect on interpretation, because it usually makes the results uninterpretable.

Median splits are bad, m’kay? ☺

Finally, the ***kinds*** of data that psychologists deal with vary considerably. In general, any given set of values can be classified as being from one of three kinds:

Raw - actual observations as originally recorded (e.g., number of self-reported siblings)
Summarized - simple summaries of raw data (e.g., *mean* RT across 20 trials)
Condensed - “scores” on various factors or sub-scales (e.g., a Beck depression score)

As above, the statistical procedures that we will be using make no distinction between these three kinds of data. Put bluntly, SPSS has no idea what kind of data you are using and it wouldn’t really care if it did. It is also possible for different kinds of data to be mixed in a given analysis. For example, “personality” might be operationally defined and quantified as five [condensed] factor scores (as is true under the “Big Five” model) and then used in combination with [raw] gender and [summarized] mean number of miles driven per year. Why would one do this? Maybe in an attempt to predict the frequency of road-rage episodes.

☞ With that said, it is important to note the following: While the kinds of data being used may have no effect on the method of analysis, they can have profound effects on interpretation. For example, because most information-processing studies use the mean of response time across many trials (as opposed to the individual, raw values from trials), the proper conclusion concerns “average” performance and not specific acts. To be clear: when a significant difference is observed between two condition means in the typical response-time experiment, the correct conclusion is that “average” performance in one of the conditions is higher than in the other, not that performance is “always” or even “usually” higher.

Complications

One complication to the above arises when the kind of data (especially raw vs. summarized or condensed) is crossed with the precision (especially discrete vs. continuous). For example, assume that you are conducting a study concerning the correlates of “average family size.” One way to operationally define “average family size” is in terms of the mean number of children (per household) in the immediate and preceding generations. By this definition, to find the value of "average family size" for a given subject, you would calculate the mean of the number of children across the subject, the subject's brothers and sisters, the subject's parents, and any and all directly related aunts and uncles. Specific example: I have a daughter and a son, while my brother has no children; my parents [obviously] have two, my paternal uncle has none, my paternal aunt has three, and my maternal aunts have four, two, and none. Therefore, my "average family size" is the mean of 2, 0, 2, 0, 3, 4, 2, and 0, which is 13/8 or 1.625. The weirdness here is that "number of children" is clearly discrete, because it must be a whole number, but this particular "average family size" is not a whole number. So what is the precision of "average family size"?

The key to resolving this is to focus on the final value of the variable (i.e., the actual datum that the subject will contribute to the statistical analysis), as opposed to where this value came from. To be clear: yes, "number of children" (for any one household) is discrete, but the analysis doesn't concern the number of children for a single family unit; it concerns average number of children across multiple units. Thus, "average family size" (using the operational definition given above) is continuous, because it can take on any value between zero and infinity. Well, maybe not infinity, but you get the idea.

There will never be a conflict or any question as to how to describe a variable, as long as you keep track of what you are talking about at any given moment. Keeping track is very important, because the precision of a condensed or summarized datum can easily be different from the precision of the raw data from which this value came.

Another example. Assume that you are interested in the correlates of hair color. On one hand, you could use the peak (or the mean) of the spectrum of the light rays reflected from the subject's head in broad-band white light and, therefore, have a continuous variable. On the other hand, you could choose a small set of categories (e.g., red, brown, black, blond, and other) and use a nominal scale. The decision is yours. The questions to keep in mind when deciding are (1) which is more likely to provide psychologically meaningful information? and (2) which can be done using the equipment and time that is allotted to this study? In the case of hair color, I don't see much of a problem making this decision. However, it isn't always so easy. This is one reason that you were given an advisor and access to a library full of journals. If in doubt, ask or see what other people have done. But please stop and see if you agree before following others.

Other Important Data Terms

As you read through the next few chapters, several other terms will appear. Some of these have a colloquial meaning that is different from the technical meaning used in statistics, which can cause some grief.

population: The set of all values of a variable -- together with their relative frequencies -- that exist in the world. Note: the values do not have to correspond to people. For example, "heights of grass" is a population when it refers to every blade of grass in the world.

population distribution: This is the same thing as *population*, but puts extra stress on the relative frequencies of the values. Instead of a list of all values (which will have many repetitions), you have a probability for each possible value, instead.

sample: A set of observed values collected as a package under similar conditions. Note: samples can involve more than one observation per subject. If it has only one, then it is a univariate sample; if it has more than one, then it's a multivariate sample; two = bivariate; three = trivariate; etc.

sampling distribution: (Note: it's *sampl**ing*** distribution, not *sample* distribution.) This is the theoretical set of values -- together with their relative frequencies -- that a variable that is calculated *from a sample* can take on. For example, we are often interested in the *sampling distribution of the mean*, which is the theoretical distribution of the means from an infinite number of samples of fixed size from a given population.

random sample: A sample that was collected on the proviso that the inclusion of any particular value had no effect on whether another particular value was also included - or - A sample that was collected such that the probability of a given value being included is exactly the same as the relative frequency of that value in the population.

Low-frequency Terms

sampling population: The subset of the population that *could* have been included in your study. This term is used to stress that some members of (general) population were not eligible. The statement that "the college undergraduate is the Norwegian white rat of cognitive psychology" is usually meant as a criticism in that it stresses how the sampling population is different from the (entire) population. If no ancillary arguments are made -- e.g., that the sampling population is a random selection from the (entire) population -- then the only justifiable conclusions from the study are those that only apply to the sampling population.

target population: The subset of the population that you are interested in making a statement about. This term is also used to stress that the study does not concern every member of the (entire) population. Example: Person X (sneering): "the college undergraduate is the Norwegian white rat of cognitive psychology." Person Y (sneering back): "that's OK, because they are my target population."