VARIABLES

Variables and the Unit of Analysis

Variables are characteristics of the "things" that we are studying, commonly called *cases*.

The kind of "thing" that is being studied is called the *unit of analysis*.

Individuals constitute the unit of analysis for much empirical research in social work. A particular research project focuses on a particular set or *population* of individuals. For example,The National Survey of Child and Adolescent Well-Being (NSCAW I and II) focus on children and families that enter the child welfare system, so the NSCAW surveys collect data pertaining to *child welfare involvement*. (see Datasets Available from the National Data Archive on Child Abuse and Neglect (NDACAN) (hhs.gov)). Relevant characteristics of individuals concerning which data may be collected include: *gender*, *race*, *education*, home removal, length of time in foster care, *depressive symptoms*, etc., so these are all examples of *variables* that pertain to *individuals*. Other kinds of empirical research may focus on individuals but on more specialized populations such as *mandatory reporters or persons who use substances*. Variables of interest pertaining to substance users include *polysubstance use, age, residence, and type of social networks*, etc. The type of empirical research that I conduct focuses on a different unit of analysis because I study neighborhood effects of health and well-being. Therefore, the unit of analysis is the neighborhood, typically defined as the census tract in which persons live, travel, etc.

Geographic units of analysis include nations, counties, census tracts, census blocks, and zip codes. The variables apply to the whole unit under investigation and inferences must be carefully made to that unit otherwise you will commit an ecological fallacy. An ecological fallacy is when inferences about lower-level units are made from analyses of higher level units. The most frequently used example is the crime rate. Knowing that the crime rate is high in a county says *nothing about the types of individuals who are committing the crime, including their socio-economic status and demographic characteristics.*

Households often constitute the unit of analysis in social work research. Census data including the Current Population Survey and the Panel Study of Income Dynamics are surveys of households whereby the *household* is the unit of analysis. Variables pertaining to households include *size* (number of people in the household), *type* (single-parent, no children, etc.), *income*, *type of dwelling* (detached house, town house, apartment, etc.), etc.

By definition, variables *vary* — that is, they take on different *values*, either from *case to case* at a particular time (this is called *cross-sectional analysis*), e.g., respondents in a survey, and/or from *time to time* for a particular case or set of cases (this is called *longitudinal analysis*), e.g., intimate partner violence (IPV) prevalence in Los Angeles County during the COVID-19 pandemic.

Variables are the building blocks of empirical research. Researchers ask such questions as the following:

1. What is the *average* or *typical* value of a variable in a set of cases? For example, what is the level of depressive symptoms among youth who are in the foster care system, or what is the prevalence of trauma among firefighters, etc.?
2. How are the values of a variable *distributed* in a set of data, i.e., do most of the same cases have about the same value (*low dispersion*) or do different cases have very different values (*high dispersion*). For example, do all foster youth have about the same level of depressive symptoms or do some have few while others have a lot?
3. How are two variables *related* or *associated* in a set of data? For example, how are depressive symptoms related to length of time in foster care?
4. Does one variable have *causal impact* on another variable? For example, does the trauma associated with home removal cause depressive symptoms in adulthood?

Variables and Their Values

Variables vary, thus, associated with every variable is a range (often a list) of possible *values*.

Education level is a variable (pertaining to individuals and which can vary both over cases (different people have different levels of education at a given point in time) and over time (a given person's education level may change over time). In the U.S. context, possible values of 'educational attainment' include high school grad, GED attainment, grad school, some college, etc. Some variables, such as 'traumatic head injury' has just two possible values, YES and NO, and special care must be taken to analyze it. HEIGHT is a physical variable pertaining to individuals with values that are real numbers (expressed in units such as inches, centimeters, or feet). SIZE a variable pertaining to households with values that are natural (whole) numbers. LEVEL OF TURNOUT is a variable pertaining to elections (or to different jurisdictions in a given election), with values ranging potentially from 0% to 100%.

Types of Variables

Every variable has at least two possible values. (Otherwise a variable could not vary, i.e., take on different values from case to case or from time to time.) A variable is *dichotomous* if it has *exactly* two possible values (often "yes" and "no"), e.g., traumatic brain injury. However, most variables have three or more — or quite likely an infinite number of — possible values.

A variable is *qualitative* if its values are given by *words* (e.g., gender, race). However, in a data array the values are typically recorded in terms of *numerical codes* (e.g. 1 = female; 2 = male; 3 = non-binary, etc.).

A variable is *quantitative* if its values are given by *numbers* (e.g., income, number of years of schooling). In a dataset, such values are typically recorded in terms of their actual numerical values.

Levels of Measurement

It is useful to refine the qualitative/quantitative distinction by further distinguishing among different *types* of variables — or (equivalently) among different *levels of measurement* of pertaining to variables.

1. A *nominal* variable (or a variable *measured at the nominal level*) has values that are *unordered* categories. Given two cases and a nominal variable, we can observe that they have the same value or different values, but (if they have different values) we cannot say that one has the “higher” value and the other “lower,” etc. The numerical codes for nominal variables must be assigned /to values in some manner which is arbitrary. For example, female could be coded 1, or 2, or 1000 the order *does not matter*.
2. An *ordinal* variable (or a variable *measured at the ordinal level*) has values that fall into some kind of natural ordering, often (but not always) running from low to high, i.e. a frequency. Given two cases and an ordinal variable, we can observe that they have the same value or they have different values, and *also* (if they have different values) that one has the “higher” value and the other “lower,” etc., but we *cannot* say *how much* higher or lower. An example is, “during the last week, how many times did you engage in bullying, “never, at least once, two – four times, five or more.'
    - Likert variables with closed-form values running from STRONGLY AGREE (or APPROVE) to STRONGLY DISAGREE (or DISAPPROVE) are also ordinal in nature.
3. An *interval* variable (or a variable *measured at the interval level*) has values that are *real numbers* that can appropriately be added together, subtracted one from another, and averaged. Given two cases and an interval variable, we can say they have the same value or they have different values, and also (if they have different values) that one has the higher value and the other lower, etc., and *also* (since we can subtract one value from another) *how much* higher or lower one value is than the other, i.e., we can determine the magnitude of the *interval* separating them and thus say how “far apart” the cases are with respect to the variable.
4. A *ratio* variable (or a variable *measured at the ratio level*) is an interval variable that has values that are real numbers such that one can appropriately *divide* the value of one by the value of another (i.e., compute their *ratio*) and say, for example, than one case has *twice* the value of another. This requires that the variable have a *non-arbitrary zero value*, which re- presents in some sense the complete absence of the characteristic or property to which the variable refers.
    - Examples of interval variables that are *not* ratio include SAT or IQ SCORE and temperature (Fahrenheit or Celsius).
    - Examples of ratio variables include number of children (0 is meaningful) or age (pertaining to individuals).

Quantitative [interval and ratio] variables may be either “discrete” or “continuous.” (Qualitative [nominal and ordinal] variables are almost always treated as discrete variables.)

1. A *discrete* variable has a finite (and typically small) number of possible values that usually correspond to *whole numbers* (integers) only. Number of children (in households) and similar *counts* provide examples of discrete variables.

2. A *continuous* variable can have *any real number* (at least within some range) as a value (i.e., including fractional values between the integers). A continuous variable has (in principle) an *infinite number* of possible values, so that additional possible values always lie between any two distinct values of the variable. Height, weight and age are examples of continuous variables.

Some variables are in principle discrete but are “virtually” continuous because they have so many possible (numerical) values. Examples may include age, etc.