VARIABLES

1. 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 social science research (and most survey research in political science). A particular research project focuses on a particular set or *population* of individuals. The National Election Studies focus on members of the U.S. *voting age population* (VAP), so the NES surveys collect data pertaining to individual of voting age. Relevant characteristics of individuals concerning which data may be collected include: *gender*, *race*, *education*, *party identification*, *ideology*, *opinion on abortion*, *level of political trust*, etc., so these are all examples of *variables* that pertain to *individuals*. Other kinds of empirical political science research focus on individuals but on more specialized populations such as *members of Congress*. Variables of interest pertaining to members of Congress (but not to rank-and-file Americans) include *number of terms served*, *last re-election margin*, *committee assignments*, *roll-call vote on a bill*, *ADA* (or other) *ratings* (of voting record), etc. Still other kinds of empirical political science research focus on quite different units of analysis, such as (i) *Presidential elections*, (ii) *states*, *counties,* and other legal *jurisdictions*, (iii) *districts*, *wards*, or *precincts* (common units of analysis in electoral research based on election records rather than survey data), (iv) *legislatures*, (v) etc.

For students of comparative and international politics, *nations* often constitute the unit of analysis. Variables pertaining to nations include *population*, *GNP*, *per capita income*, *literacy rate*, *military spending as a percent of GNP*, *type of party system* (one-party, two-party, multi-party), *regime* (democratic, authoritarian, etc.).

Households often constitute the unit of analysis in sociological or economic research. (The Current Population Survey and the Panel Study of Income Dynamics are surveys of households.) 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., states in Presidential elections from 1824 to 2000.

Variables are the building blocks of empirical social/political science 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 typical level of interest among voters, or the average rate of turnout in recent elections?

(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 voters have about the same level of interest or are some very interested while others not interested at all? Do all elections have about the same level of turnout, or do some have high turnout while others have low turnout?

(3) How are two variables *related* or *associated* in a set of data? For example, is the level of interest among voters related to their level of education? Does the level of turnout in elections depend on how close the election is (or is expected to be)?

(4) Does one variable have *causal impact* on another variable? For example, does high education cause people to become more interested in politics? Does a close contest cause more voters to turn out and vote?

2. Variables and Their Values

To repeat, variables *vary* — they take on different values from case to case from time to time.

Thus, associated with every variable is a range (often a list) of possible *values*. For example, PARTY IDENTIFICATION is a variable (pertaining to individuals and which can vary both over cases [different people have different party identifications at a given point in time] and over time [a given person's party identification may change over time]). In the U.S. context, possible values of PARTY ID include REPUBLICAN, DEMOCRAT, INDEPENDENT (or perhaps refinements like STRONG REPUBLICAN, WEAK DEMOCRAT, etc., and/or other values like MINOR PARTY). VOTED IN 2000 ELECTION is another variable pertaining to individuals, with just two possible values, YES and NO. 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%.

As a reminder that any variable must have a range of two or more possible values, it is generally useful to give variables names like *LEVEL OF* EDUCATION, *WHETHER* VOTED IN 2000 ELECTION, *SIZE OF* POPULATION, *TYPE OF* REGIME, *LEVEL OF* TURNOUT, *DIRECTION OF* IDEOLOGY, etc.

The actual value of a variable in a particular case is called an *observation* (or *observed value*). Thus we "observe" that Joe Smith (the case) has the PARTY IDENTIFICATION (the variable) WEAK DEMOCRAT (the observed value). Likewise the 2000 Presidential election (the case) has a LEVEL OF TURNOUT (the variable) of 51% (the observed value).

In survey research, each respondent is a case and each question generates a variable, because different respondents can give different answers. If a question is closed form, the (multiple-choice style) options constitute its possible values (which will likely be numerically coded in the data spreadsheet). If the question is open ended, the possible values depend on coding practices that the researchers must develop and apply.

3. 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., WHETHER VOTED IN 2000 ELECTION. 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., PARTY IDENTIFICATION, TYPE OF REGIME). However, in a data array (such as the Student Survey data spreadsheet that you have previously received) these verbal values are typically recorded in terms of *numerical codes* (to save space and to facilitate machine processing).

A variable is *quantitative* if its (true, not coded) values are given by *numbers* (e.g., LITERACY RATE, SIZE OF HOUSEHOLD, LEVEL OF INCOME, LEVEL OF TURNOUT). In a data array, such values are typically recorded in terms of their actual numerical values.

Notice that substantively related variables may be of different types depending on the unit of analysis to which they pertain. For example, TURNOUT *pertaining to individuals* in a single NES study is a dichotomous variable, i.e., a given respondent reports that he or she either voted in the election or did not. However, TURNOUT *pertaining to elections* (or jurisdictions) is a quantitative (and interval and essentially continuous — see below) variable with possible values ranging from 0% to 100%.

4. 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. These distinctions are really relevant only as they pertain to non-dichotomous variables (with three or more possible values).

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. A nominal variable typically has a name like NAME OF ____, TYPE OF ____, NATURE OF ____, KIND OF ____, etc. In

a data array such as the Student Survey or SETUPS data, numerical codes for nominal variables must be assigned /to values in an essentially arbitrary manner. (TYPE OF) RELIGIOUS AFFILIATION is a standard example of a nominal variable (with values of PROTESTANT, CATHOLIC, etc.). Given an election with three or more candidates, CANDIDATE PREFERENCE is a nominal variable.

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. 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 ordinal variable typically has a name like DIRECTION OF ___, EXTENT OF ____ , LEVEL OF ____ , DEGREE of ____, etc. In a data array, numerical codes can (and should) be assigned to values in a manner consistent with their natural ordering. If the natural ordering is from LOW to HIGH, the codes should likewise run from lower to higher numbers. If the natural ordering is not from LOW to HIGH, e.g., DIRECTION OF IDEOLOGY, the two extreme values (or "poles"), e.g., MOST LIBERAL and MOST CONSERVATIVE, should be assigned the minimum and maximum code values, but which gets which is arbitrary (and intermediate values, e.g., MODERATE, should be assigned intermediate codes.) Note, however, that DIRECTION OF IDEOLOGY could be renamed DEGREE OF LIBERALISM, which does range from LOW [i.e., "least liberal"] to HIGH ["most liberal"]. Or we could reverse the "polarity" of the variable and call it DEGREE OF CONSERVATISM, ranging from LOW [i.e., "least conservative" (or "most liberal")] to HIGH ["most conservative" (or "least liberal")].

   Opinion variables with closed-form values running from STRONGLY AGREE (or APPROVE) to STRONGLY DISAGREE (or DISAPPROVE) are also ordinal in nature. Since the value INDEPENDENT is usually deemed to fall "between" DEMOCRAT and REPUBLICAN, (three-category) PARTY IDENTIFICATION is usually deemed to be ordinal, as is the standard seven-category PARTY IDENTIFICATION (formed by the composition of Student Survey Q1-2 and corresponding to SETUPS V09). However, this works only if we treat cases with "minor party" values as missing data.

   Note: In practice, a data spreadsheet displays numerical codes for missing data ("unobserved" values), which cannot be part of any natural ordering. Typically missing data is assigned numerical codes such as 0, 9, 99, etc. (One has to tell SPSS the "missing data" code(s) of each variable, so that it can set cases so coded aside when it makes calculations.)

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. An interval variable typically has a name like LEVEL OF ____ , DEGREE OF ____ , AMOUNT OF ____ , etc. Given interval variables, actual numerical values (rather than numerical codes) are normally entered into a data array, though sometimes (numerically coded) *class intervals* as used instead (e.g., SETUPS V60 [AGE]), as will be discussed later. Variables like 7-CATEGORY PARTY IDENTIFICATION are often treated as interval variables (I did this in computing and displaying "Mean Party Identification" in charts based on cumulative Student Survey and NES data), though the justification for doing this may be a bit weak.

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 represents 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 LEVEL OF SAT (or IQ) SCORE (pertaining to individuals), DEGREE OF TEMPERATURE (Fahrenheit or Celsius, pertaining to locales [cross-sectional] or days [longitudinal]). IDEOLOGY variables (e.g., Student Survey Q27-37 and SETUPS V34-39), as well as PARTY IDENTIFICATION and OPINION variables, are often deemed to be interval (since we may average their values) rather than merely ordinal, but they certainly are not ratio.

   Examples of ratio variables include NUMBER OF CHILDREN or AGE (pertaining to individuals), NUMBER OF MEMBERS (pertaining to households [or legislatures]), SIZE OF POPULATION (pertaining to nations [or other jurisdictions]), LEVEL OF INCOME (pertaining to individuals or households), LEVEL OF PROFITS (pertaining to business firms) or SIZE OF BUDGET SURPLUS (pertaining to governments). (Note that, even though LEVEL OF PROFITS or SIZE OF BUDGET SURPLUS can have negative values, their zero points are not arbitrary; however, ratio comparisons can only be made between observed values with the same [positive or negative] sign).

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

5. 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), NUMBER OF MEMBERS (of councils or committees), and similar *counts* provide examples of discrete variables.

6. 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. LEVEL OF TEMPERATURE (of days), HEIGHT, WEIGHT, and AGE (of individuals), provide examples of continuous variables (though we typically round off to the nearest degree, inch, pound, year, etc.). IDEOLOGY might be thought of as a "truly" continuous variable, which is *measured imprecisely* (to be covered in the next Handout) in whole number values only.

   Some variables are in principle discrete but are "virtually" continuous because they have so many possible (numerical) values. Examples include SIZE OF POPULATION and LEVEL OF TURNOUT. Indeed, "truly discrete" interval variables with more than a half dozen or so possible discrete values are often most conveniently analyzed as if they were continuous.