Vanderbilt University Political Science - Stats I
Fall 2024 - Prof. Jim Bisbee

Lecture 1

- Housekeeping.
  - Syllabus:
    - Deliverables.
      - Attendance.
      - Class and lab.
      - Homeworks.
      - Exams.
    - Grades.
    - Help.
      - Introduce Martin.
      - Propose Martin's office hours; check to see if works.
    - Books.
    - R.
    - Brightspace. [check.]
    - Hand out questionnaires.
- Why are you here? (i.e., why are you getting a Ph.D. in political science?)
  - You are here because you enjoy asking and answering questions about politics.
    - (If not, you're in the wrong line of work.)
  - So what kinds of questions do you find interesting? List on board.
- This is a course about quantitative analysis in political science.

  - Quantiative analysis takes its place alongside other methodologies in empirical political science. It is particularly helpful and appropriate for particular kinds of social scientific tasks.
    * Tends to be based on: *numerical* measurements of specific aspects of social and political phenomena; [contrast to: non-numerical]
    * Is more interested in developing and testing generalizable theories about *multiple* phenomena; [contrast to: particular cases]
    * Seeks measurements and analyses that are easily *replicable* by other researchers. [contrast to: difficult to replicate in its entirety.]
  - Contrast this to *qualitative* research [on all three aspects].
  - In addition, appropriateness of *qualitative* research depends on...
    * Hypothesis testing vs. hypothesis generation
    * Agreed-upon measures of concepts vs. those still up for debate
    * Analyst's willingness to apply less or more *structure* to the study of the phenomenon

- When political scientists work with data, we generally do so for three reasons:
  - What can we say about the data we have?
  - What can we say about the data we don't have (but we know is out there in the world) based on the data we do have?
  - What can we say about the data we'd expect to see under a hypothetical scenario, based on the data we do have?

- These activities generally require three kinds of statistics:
  - Descriptive
  - Inferential (from samples to populations)
  - Prediction (from models to hypothetical scenarios)
    * And lingering behind all of this is a kind of statistic we'll use all the time: the *test* statistic.

- What's a *statistic*?
  - It's a number that summarizes data.

0.1 Variables 101

-
  - We study **units**. They are the level at which we wish to make statements about a social process. Units are often also known as **cases**.
    - People, Counties, Nations, Dyads.
  - Units have **attributes**. An attribute is any characteristic of a unit that in theory might distinguish it from other units.
  - **Variables** are *logical* groupings of *mutually exclusive* attributes. The analyst assigns each attribute a **value**. We then say that a variable takes on a value for any particular unit.
    - The variable "hair color" takes on the value "red" for the unit Me.
    - The variable "party affiliation" takes on the value "Republican" for the unit Mitt Romney.
  - For ease of data manipulation in quantitative analysis, we typically assign **scores** to each potential value a variable can take on. The scores are simply numbers associated with each value. Sometimes the scores are meaningful in their own right; often they are not. But it's generally a lot easier to enter and manipulate data via numbers than via the words we use to describe values.
  - Note that all of these–which units, which attributes, how to group into variables, how to score the variables–are *choices* that must be considered and justified by the analyst. Typically there are conventions within subfields of political science that either must be adhered to (which is what we usually do), or if we depart from them we need to justify this departure. Sometimes the departure itself is a noteworthy innovation. Going to leave aside here a whole field of theory on how we move from concepts to **measures**. In this class, in most cases, we'll take the measures as given.

0.2 Levels of measurement

- Variables can be measured at various levels. We'll consider four such levels in our class:

  - Nominal
  - Ordinal
  - Interval
  - Ratio

- Variables measured at the **nominal** level take on values that cannot be ordered in any logical way.
  - Egs.

- Variables measured at the **ordinal** level taken on values that can be rank ordered, but that's it. We know nothing about how much more or less one value is than another.
  - Egs.

- Variables measured at the **interval** level take on values whose differences can be meaningfully compared. I.e., the interval between two values is meaningful.
  - Eg: Fahrenheit. The Year.

- Variables measured at the **ratio** level take on values such that the value of zero is meaningful in a specific sense: it means *nothing* of the quantity being measured..
  - Eg. income. height. age...number of wars...miles.

- Mathematical operations:
  - **Nominal:** equality.
  - **Ordinal:** equality. greater than or less than.
  - **Interval:** addition and subtraction; averages
  - **Ratio:** multiplication and division (i.e. ratios) - "twice as tall," "six times as wealthy"
  - Note that we generally have more information as we move up the ladder. Your first instinct should be to use measures that retain as much information about your units as possible.

- Three trickier cases:
  - Richter (ratio, although slightly tricky: $y_{Richter} = \log_{10}[\text{shaking amplitude}]$, so each unit represents a *doubling* of amplitude.)
  - Celsius has a meaningful zero (the freezing point of water), but it is *not* ratio level. Why? Because "zero degrees Celsius" $\neq$ zero of the quantity "temperature." [E.g., is a 30-degree Celsius day (86F) thirty times as hot as a one-degree Celsius day (32F)? Not meaningful.] Contrast to "zero miles" or "zero years old." Celsius is interval level. Contrast to Kelvin, measured at the absolute zero. Zero degrees K = -273.15 C; 1 degree K = 0 C, and so on.
  - Latitiude, longitude: for all intents and purposes, ratio. Of course, zero on the latitiude scale doesn't mean "zero distance." It means we are at the Equator. But as a unit of measure, to travel two degrees latitude is to travel twice as far as one degree latitude. If you travel from the Equator to Vermont, at 45 degrees latitude North, you've covered *half the distance* to the North Pole (at 90 degrees latitude). Therefore ratio.
- A special case: the dichotomous variable. Two ways to consider it: as a nominal variable with two categories, or as a ratio variable with two values, zero and one. E.g. "gender" versus "female."
- Choice of level of measurement is often up to the analyst. Many underlying concepts yield several choices for levels of measurement. Again, your first instinct should be to use measures that retain as much information about your units as possible. E.g.:
  - Location of residence: region of country [nominal], county [nominal], zip code [nominal!], latitude and longitude [ratio].
  - Hair color: common usage [nominal], amount of pheomelanin and eumelanin (hair pigments) [ratio]
  - Income: poor, middle class, rich [ordinal]; dollars per year [ratio]

0.3 Data structures

- The typical way that we store data (and really, the way that we think about data structures) is via a **data table**. Excel, any statistical software program. Each unit is given a row. Each variable is given a column. [You could of course do it the other way, but as in many things you'll find that you save time and headaches by being consistent with convention.] The scores (and/or in some cases, the values) the variables take on for each case are entered in the cells. [Draw on board.]

0.4 Summarizing data: displays

- So we've got a group of units, and we know the values taken on by a set of variables in this set of units. Our next move is typically to say something meaningful about the group with the data at hand. [Gesture to table.] Why not just present this?

- You're not going to get any more detailed than this. But descriptions of data usually involve tradeoffs between detail and parsimony.

- So let's move to a higher level of abstraction. A frequency table.

  - Displays the frequency with which a variable takes on values in a group of units.
  - Columns with value, number of units, proportion of units
  - One row for each value.
  - And for ordinal, interval, and ratio-level data: cumulative percentile.

- When displaying frequency distributions, sometimes it makes more sense to do so graphically. Typically graphs move a bit toward parisomony at the expense of some detail. [scale on board] What information is missing from the graph that isn't in the table?

  - Another choice: **recode** the data into categories and then produce either a frequency table or a bar chart of the categories. [e.g. on board]

- When we move to interval- or ratio-level data with lots of categories, a histogram is almost always preferred to a frequency table.

0.5 Summarizing data: measures of central tendency and dispersion

- An even higher level of abtraction: measures of central tendency and measures of dispersion.
  - Measures of **central tendency** tell us about the *typical value* of the variable in a group of units.
    - mode: the value of the variable most frequently observed in the group of units (all LOMs; wait to say this): ;
    - median: the value of the smallest observation for which the cumulative percentage is 50 or greater (ordinal and up);
    - mean - the average: $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ (interval and up).
    - Note that all of these measures summarize all the observations of a variable with just one number.
  - Measures of **dispersion** typically accompany measures of central tendency. They provide a sense of the "spread" of a variable's distribution – a.k.a. the amount of *variation* in its distribution.
    - *range* (the difference between largest and smallest values);
    - *IQR* (the difference between the values at the 75%ile and 25%ile,
    - *variance* $s^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$. Note different from book, because more intuitive. The standard deviation (s.d.) is simply $\sqrt{s^2}$. The s.d. is a more informative measure of dispersion: it is the average distance of an observation from the mean. It is measured in the same units as the variable itself. What sign must the variance and thus s.d. always be? (LOM: interval or higher):
  - Furthermore, social scientists often speak *qualitatively* about the frequency distribution of a variable:
    - It may be symmetric [draw] or skewed [draw], also number of children on handout.
      - Typically the median is a better measure of central tendency for variables with skewed distributions than the mean, because it is resistant to outliers. Use age at married example.
    - It may be unimodal or bimodal [number of children]

* There are *quantitative* ways to speak about the distribution of a variable. For example the **skewness** of a distribution is typically calculated as $g_1 = \frac{1}{N \cdot s^3} \sum_{i=1}^{N} (y_i - \overline{y})^3$ This statistic will be zero in a perfectly symmetric distribution. It will be *negative* if observations below the mean tend to be farther way from it than observations above the mean/data are skewed left/long left "tail"). If will be *positive* if data are skewed right (vice versa/long right "tail"). You rarely see this statistic used in political science, but it will pop out [handout].