LS 123 Data, Prediction, and Law

Spring 2022

Meeting 2

**Sample Survey Data Plus Introduction to Jupyter Notebooks and Python Basics**

1. **Announcements**
   1. Did people get the Google survey? My goal is to get project groups together by early next week (hard to do while people can still drop)
   2. Course Assistants to help with labs, Python, etc.—office hrs on Piazza
      1. Anna Gueorguieva: office hrs on Piazza
      2. Muskaan Soti: office hrs
      3. Zaina Syed:
   3. Ilya and I also will have office hrs: mine are Tues afternoon (link on bCourses homepage)
   4. We will jump right into using Pandas, which has its own set of methods but which schematically is pretty similar in that it is rows of cases and columns of variables that you can slice, sort, and manipulate; it does require some practice so that is why we are starting with it and using a nice, clear, survey research dataset, the 2016 National Election Study
   5. Go ahead and open Lab 3 on Datahub, as well as the ANES 2016 Codebook since it may help with what I will say
2. **Sample surveys: the “gold standard” for social science data** 
   1. Again, why ANES 2016
      1. First, I know a lot about the domain bec of academic training, and domain knowledge is key to being able to work with data
      2. Second, we know the context of 2016
      3. Third, it is sample survey data, of the sort you see frequently in social science – think quality polling, or the General Social Survey, or the American Community Survey
      4. These are good at establishing what the population looks like but they are very expensive
   2. American National Election Study: every election year since 1948! (which means that what once was political science is going to be history)
   3. Labor intensive, historically—face-to-face interviews, both pre and post election
      1. very expensive: like 10 million per election cycle in presidential election years
      2. Need to have many interviewers in many places, armed with laptops to help administer the survey questions
      3. Need to provide an incentive for participants ($25-100 for the face to face and up to $80 for the online survey)
      4. Note that the online component is relatively new (I think it first was conducted this way in 2012), and that it used to be all face-to-face
   4. Sampling technique
      1. Not a simple random sample of eligible voters (where everyone would have an equal probability of being chosen)
      2. But, the idea of keeping close to an equal probability of being chosen is retained while at same time getting coverage
      3. Stratified Random Sample
         1. choose categories of interest and then draw at random within them, trying to keep the probability of being drawn, as a household, equal across units
         2. Primary sampling units: The PSUs were counties, or combinations of counties to form a minimum population of 50,000, or, in the case of Los Angeles County, half-counties (i.e. L.A. County was divided into two PSUs due to its large size).
         3. Counties were stratified by Census region, prevalence of poverty and members of minority groups, and population size, and then selected at random with probability proportional to the number of adult citizens
         4. except that Cook County, IL, Harris County, TX, Maricopa County, AZ, and both halves of Los Angeles, CA, were selected with certainty
         5. secondary sampling units: census block groups
            1. repeat the process with these smaller units: draw with probability proportional to population (which is I think how census block groups are constructed) from each of 4 census block groups in primary sampling unit
            2. within those groups of census blocks you choose household addresses at random
            3. then finally find adult respondent in each household
      4. Cannot use the plain old standard error calculation but instead must account for the fact that the respondents are clustered in census block groups, which in turn are clustered in the primary sampling units (counties, groups of counties, or county parts)—we cannot assume error in uncorrelated within the sampling unit
      5. Note that some places are not chosen probabilistically: Los Angeles, that this only surveys US citizen adults eligible to vote, and that only people with addresses are surveyed; AK and HI were only part of the online survey, not the in-person
      6. Sometimes the sampling strategy is to oversample segments of the population that you want to learn more about and then downweight their responses proportionally
         1. that’s why there is a sample weight variable, to account for the fact that respondents are not drawn with equal probability
         2. sometimes that is due to the sampling strategy, and sometimes it is part of the design in order to be able to answer certain questions
         3. For example, what if we want to know what African-American Republicans think? If we take an equal probability sample, we will not get enough to make generalizations (so 13 in a sample of 1000, maybe)
         4. So we can oversample African Americans (now even oversample may not capture enough)
         5. ANES 2016 did not oversample in this way
   5. What is this good for? A couple of things
      1. Making causal inferences
         1. Respondents are asked lots and lots of questions about all sorts of things (which we see in the codebook and in the lab exercises)
         2. This can allow us to test relationships between, e.g., individual demographic attributes (race, education, home ownership, age) and, e.g., policy preferences for things like spending, criminal justice policy, and so on
         3. This is a very common approach in political science and sociology—we want to theorize what makes people think in a certain way and then test our hypotheses based on those theories
         4. A very famous example is Phil Converse’s hypothesis about holding a consistent political ideology—not very many people do, and those who do have a consistent ideology tend to have more education
      2. Generalizing claims to the population of US voters
         1. Since it is a sample, and since cases are weighted, you can adjust the raw responses by the individual case weights (which account for how likely that respondent was to be drawn) in order to find out how the population of voters is distributed on various attitudes
         2. For example, V161343 (What do you think about roughing up protesters? Is it justified?)
         3. Go look it up in the codebook PDF file [**demo**] by doing a search for V161343—we can say that 35% of eligible voters say it is not at all justified to rough up protesters
         4. Note for this item that responses are on a scale of 1 to 5; we will talk about that once we have gotten into the lab, but that allows political scientists to say that there are systematic differences among groups of voters (e.g. Democrats and Republicans) and maybe even to calculate a mean score for each group (if you are willing to make assumptions about the measure you’ve created—more later)
         5. This way researchers get a detailed view of what eligible voters think about a host of topics, and it will behave in statistically understandable ways (that is, if there were repeat sampling, we have a confidence interval in which we would expect the measurements to fall, etc)
3. notes on Python in Jupyter notebooks environment (not part of lecture)
   1. talking points on programming in Python
      1. Always prefer simplicity over complexity
      2. Comment your code
      3. Code should be interpretable by other users without your instructions
      4. Avoid duplication of information, confusing or vague names for objects, etc.
      5. Code should be fully executable without human instruction
   2. I do want to emphasize one thing: **comment your work**, including inside each code cell, since it helps you to remember what you were doing and why, and it helps other people understand that too

**Lab 3** (30 minutes)

1. **Lab discussion:** a disquisition on different levels of measurement (powerpoint 02\_SampleSurveyData)
   1. Initially for this lab we used the National Crime Victimization Study, which had a question about theft and a variable about the monetary value of the theft
   2. That is a good ratio measure: each dollar is exactly one measurement unit apart, and there is a meaningful zero on the scale
   3. This leads me to something we will talk about a few times while we talk about ANES—what sort of measure are we talking about [write categories on screen]
      1. Nominal—categorical things like sex, party affiliation, favorite color
      2. Ordinal—things that can be ordered, like feelings or preferences (but where the gap between the values is not entirely clear)
      3. Interval—equal intervals between units of measurement, eg. Degrees Celsius
      4. Ratio—not only are intervals the same but there is a meaningful zero (e.g. kilograms of mass, dollars), such that you can express ratios (3 g is half of 6 g, but 20 degrees is not half of 40 degrees)
   4. Be sure to look at the codebook to see what the response categories are and whether it makes sense to treat the variable as discrete or continuous
      1. But let’s go to V161343 about roughing up protesters—there was a five-point scale, right?
         1. Is this an interval measure?
         2. Sociologists often say no, they are not, just ordinal, but political scientists more willing to say “well, let’s assume that it is an interval measure, since that way we can find a mean, and calculate differences in means, between interesting groups like Democrats and Republicans
         3. We will do more with this assumption next time when we talk about visualization, but it is useful to think about the possible differences in approach depending on what you are doing – but you should know what assumptions you are making
   5. Missing data: it is best when it is handled explicitly as in the codebook, so that you know what happened to the missing cases
      1. Note that the lab deals with it by using shorthand—two assignment statements using the entire dataframe (anes[anes < 0] = np. Nan) that just takes the out of range values as assigns NaN
      2. This also converts it to floating point, which can change the state of the notebook (esp if you run cells out of order)
      3. This is not the kind of shortcut we should use if analyzing data for real, but it is easier to manage in SPSS, for example