Research Methods in Political Science Focus Questions for Unit 2
Last updated: 10/21/2024

Any bolded phrasing notes something that changed after the earliest draft of this guide was posted (Note, some blocks of questions were moved Unit 3 after this guide was first posted)

- Review: Be able to define and give examples of "independent," "dependent," "intervening," and "control" variables. Be able to apply these labels if given the summary of a study.

- Review: What are the two main ways researchers "control" for the effect of certain variables to make sure that any relationship they see between the independent and dependent variables is not spurious? (Hint: in experiments, certain characteristics are kept the same for all observations; however, in large-n statistical studies (like almost all research using survey data), social scientists use statistical programs to calculate and assign the average value of each control variable to each respondent. In other words, if you are studying the effect of more education on the likelihood of voting for Republicans and want to control for gender, the statistical program calculates results as though everyone was half female (i.e., .51 x whatever positive effect being female typically has on voting Democratic).

- Review: What is a theory, what is a hypothesis, and how do hypotheses relate to a study's independent and dependent variables? What characteristics make some hypotheses better than others? Can a study's hypotheses contradict? (Hint: yes they can contradict if you are testing competing theories or casual relationships between the independent variable/s and dependent variable/s as is the case in article you read on the Brazilian election of Bolsonaro article back in Unit 1)

UNIT 2 MATERIAL

Best practices in sampling and "sample error"

- When designing survey "samples" to make inferences about a larger "population," why is it important in most cases that the surveyed sample of individuals be closely "representative" of the population as a whole? What is the goal with a "random sample" (i.e., what is the technique trying to achieve?)? How are "random" samples different from "convenience" samples, where researchers often use census-derived "quotas" to make the sample look like (aka "mirror") the US as a whole? What kind of random sample is the best?

- The gold standard for generating a truly representative survey sample is to use a technique that randomly selects participants. Specifically, the best survey-based studies of political, economic, and social behavior or attitudes use samples in which every person in the target population have the same probability of being included in the survey. The three most common approaches to probability sampling are: "simple," "stratified," and "cluster" samples. Why are stratified samples better than simple random samples when studying a widely distributed population? When are cluster samples used (hint: good for studying students or prisoners)?

- What is the best way to get a random sample of the nation as a whole? Why have researchers mostly stopped using RDD (random-digit-dialing cell and landlines) to build these surveys? How do they do in now? How does the Pew Foundation build its samples? Why is the Pew Foundation using a large "panel" survey to offset the massive expense of creating a high-quality probability sample by mail? Are there any other advantages to this type of panel survey?

- How do researchers use Census results, specifically the "parameters" (versus a sample "statistic") from the nation's decennial survey to make sure that their survey samples are truly representative of the nation if that's the population we want our survey to

represent?

- Oftentimes, researchers use "convenience" samples (i.e., samples *not* compiled using random sampling techniques). Usually, convenience samples are used to cut costs (this, for example, was the case with the sample used in Whitt et al.'s study of polarization). What steps can researchers use to make these non-random samples work better? Hint: Researchers can compile their convenience sample with age, gender, race, urbanicity, region, and partisanship quotas that mirror population characteristics reported in better- quality population studies—including Census parameters. And, they can verify that their survey's results match those for various groups on high-quality random samples. Ideally, many questions asked of a convenience sample will duplicate questions recently administered in higher-quality samples, allowing researchers to determine if their convenience sample differs substantially from what they would see with a random probability sample.

- What is "sample error"? Even if we were able to draw a sample where every American had an equal probability of being included, we would only be observing a portion of the population, and our results would be subject to "sampling error." What is this error, and what is its relationship

- Even if we were able to draw a sample where every American had an equal probability of being included, we would only be observing a portion of the population, and our results would be subject to "sampling error." What is this error, and what is its relationship to the number of individuals in the sample? What is the sampling error for a survey of 500 individuals versus 1,500 respondents? Why not talk to 4,000 or even 10 thousand people every time we sample a population (how much would those larger samples reduce the sample bias related to the probability that our sample deviates from the larger population as a function of being a sample?

- Interpret each element in the following statement, which is similar to what is found in the notes of most surveys whose results are reported by the print media: A (completely made-up by your professor) USA Today/Suffolk University poll reported that 48.1% of Americans think that the US should undertake military action to curtail North Korea's nuclear program. Source: Poll of 1000 registered voters taken Wednesday, October 7 through Friday, October 9m 2017. Margin of error +/-3 points, 95% confidence level.

- The Pew Foundation's surveys sometimes include additional, larger oversamples of subgroups, such as African Americans or young people. In other surveys, Pew looks only at respondents in these groups. Why can't reliable information about these individuals just be calculated from a typical survey's results? Why is it generally a problem when survey results from a nationally representative survey are reported in a way that includes demographic subgroups (e.g., union members, Latinos, evangelical Protestants, African-Americans, etc.)?

- If you want to know about the distribution of views or behaviors in one of these subgroups, why is it better to use a 1000-person sample where each member of the subgroup has a roughly equal probability of being included in your sample rather than a much larger sample that has combined (that is, "pooled") subpopulation respondents from many different surveys all asking the same questions?

Dealing with sample bias and other issues that might cause survey results to not reflect public opinion or behavior correctly

- What is "sample bias," and what types of individuals are more likely to be included or omitted from the typical survey sample? What document do you look at when downloading a survey to see how biases are being addressed? What are some of the strategies researchers can use to get typically underrepresented groups (e.g., young conservatives and various non-white groups) to participate in

surveys?

- How is post-survey weighting used to address issues of atypically higher or low response rates in surveys for some subpopulations? What are the "population parameters," that researchers use to weight data? Where do researchers get this information, and how are it used before and after a survey to make sure that a survey represents the population it is supposed to?

- How can questionnaire design influence the extent to which respondents' answers on a survey correspond to their actual beliefs? What is "social desirability bias," "acquiescence bias," and "priming." What steps can researchers take to at least partially address each of these issues? (In thinking about this question you will find it helpful to know what "list experiment," "forced choice," and "rotated" questions involve. Be able to give some examples (From Pew's explanation of how they write survey questions) on how seemingly modest differences in a question's wording can elicit quite different response patterns. Be able to give examples where list experiments show that American's "public views" on issues are quite different than their "private views," for at least some issues (see in class charts and the short Axios reading). How do scholars measure prejudice against other groups, since few people openly express hostility against many of the groups that they may be biased against?

In what ways is survey research improving over time even as response rates decline?

- Why is vote choice polling so helpful to social scientists (hint: the outcomes of elections allow us to see how well our survey methods are doing and to make improvements)? Why is it harder to estimate an election's outcome in advance versus people's preferred candidate going into that election?

- Recognizing that most political surveys administered prior to 2022 included a lower percentage of Trump voters than his election-day share, Pew released a study that recalculated previously published findings for a wide range of issues to see how boosting (i.e., over-weighting) the share of Trump voters in samples to match his actual share of the electorate in the 2020 election would change our understanding about how Americans feel on many divisive issues. How much did over-weighting Trump voters change Pew's results? Is there any reason to believe that survey research can produce accurate results most of the time even if it sometimes under- or over-weight certain types of people?

- According to Nate Silver at FiveThirtyEight, why did so many polling analysts underestimate Donald Trump's probability of winning in 2016 (Silver made different "assumptions" and said on the eve of the 2016, that collectively and systematically analyzing the polls gave Trump a 30% chance)? How has polling improved with each election since then? (Specifically, what steps have been taken to reduce survey bias—that is, the under-representation of certain types of voters in polls)? Why have vote-choice surveys been less accurate in presidential elections featuring Trump than in off-year elections when he hasn't been on the ballot?

Finding and starting to analyze survey datasets with statistical software? And other tools that help us to understand how our statistical research fits into the larger social science literature

- In class and various SPSS workshop assignments we will be working with instructor-modified, smaller versions of the American Values Survey from the Public Religion Research Institute (PRRI). The methodology for that survey is explained in last part of the dataset's questionnaire, which is in the same folder. How did the PRRI survey assemble a sample that looks like America concerning race, age, gender, and several other variables? Why is this a better sample than the convenience sample used in many social science studies (including the "DataPrac" dataset that came with your textbook)? Why is it a better sample than mostly phone surveys, which are still used by many US news organizations?

- What is Google Scholar, and how is it different than the regular version of Google? How do you search for a phrase rather than just one term in GS? How do you omit certain words from a search (for example, if you were looking for research with a variable measuring pride in being American but wanted to omit Lee-Greenwood-related citations from showing up)? How can you get GS to show

you links to articles in HPU's library if you are away from the university? How do you get GS to show you how to format its citations to meet different citation style requirements? How do you get GS to show you only recent articles, say those from the last few years? Now, the most important question you need to know the answer to when doing a GS search: If you find a particularly promising article in a search, how do get GS to show you more citations that are similar?

- Why is the Pew Research Center such a good source for public opinion data? How do you find datasets on the website? What are some the major research units at the Pew Research Center, and what groups or topics does each focus on?

- Where do you find information on what topics Pew's research units are investigating at the moment (i.e., their research notes/reports)? What's the difference between a "top-line" report on a dataset (often available earlier) and a survey's dataset?

- If you find a particularly interesting research note on a Pew site, how do you find and download the relevant survey questionnaire and dataset if there is no link on the note? Why does it usually take Pew a year or more in many cases to post their data for public use?

- You read about this in the block of material you read at the end of unit one: How are most specialized Pew surveys administered (answer: with phone calling and using methods to collect a random stratified sample)? How is the American Trends Panel (ATP)—which is now Pew's main survey on American political opinion—different? (Answer: Individuals are asked to participate using the traditional methods for compiling random samples so that every American has an equal opportunity of being asked to be in the sample. However, after being contacted, members of the ATP sample are asked if they would be willing to keep taking additional surveys by computer or smartphone for modest compensation; new respondents selected by best-practice methods are added to replace individuals who drop out. In short, the ATP is not a convenience sample like the DataPrac survey sample is).

Using SPSS syntax to make dataset analysis easier

- Shifting to a completely different topic…. SPSS has three different types of windows: (1) the data editor, which has a tab to view the "data matrix" of entered survey results and another tab called "variable view"; (2) the "output window," which appears anytime SPSS is asked to do something; and (3) the syntax file. What is the main purpose of each window, and what kind of information do they contain?

- With complex datasets, it can be helpful to create a truncated (i.e., "small") version of a dataset that includes only the observations and variables that will be examined in the study. In very general terms, how do you create a truncated dataset in SPSS? For instance, if you have a 13-country study that includes hundreds of variables, including 13 separate, country-specific measures for educational attainment and income, how would you use SPSS syntax to first eliminate respondents from all of the countries except for the one that you care about? How would you then drop all of the irrelevant variables?

- Why is it important to work with a copied dataset rather than your original data when making these types of changes? You don't need to be able to describe the SPSS syntax in any detail, but be able to explain the process of making a small dataset.

- Why do researchers need to use syntax anytime they are modifying a dataset or variables in it? Why do you want to save all modifications to your data (whether it be dropping observations/variables or whenever you create new variables) in a savable syntax file when you could just point and click your way through SPSS menus to do the same thing?

- Why do pretty much all of the variables in Pew datasets (or similar datasets from other prestigious research organizations) need to be recoded before they can be analyzed? What happens if you calculate the mean of a variable that hasn't been recoded (i.e., why does Pew typically assign 9, 99, or 999 values to respondents who answer "don't know" or refuse to answer a question?)?

- Why are dummy (i.e., two possible values) categorical variables—aka "dummy" or "binary" variables—always recoded one if the respondent belongs to the category and 0 if they do not? (As a reminder, the answer to this question is that the mean of a dummy variable will give you its proportion of the population belonging to the group). Incidentally, coding dummy variables 0/1, rather than 1 and 2 as PRRI and Pew do in their original coding also helps interpret regression results, but you won't need to think about this until we get to that part of the course.

- When you are recoding a variable, why is it highly recommended that you do each of the following: (1) Do all of your recoding work in a savable syntax file (even if you are using SPSS's point-click-and-paste to create the code); (2) recode each of your variables of interest into "new" rather than changing the original variable (i.e., coding the new variable into "the same" one, (3) cut-and-paste the original questionnaire wording (or the results of a CODEBOOK command into your syntax, and (4) make a note in the syntax file about any reasoning you have for recoding a variable one way versus another? How do you grey-out notes you put into your syntax that contains information from the

questionnaire and any methodological notes (remember: if you don't tell SPSS that these notes aren't part of your syntax command langauge, your syntax won't run)?

- How do "categorical" (aka nominal), "ordinal," and "interval (aka continuous)" variables differ from one another, and why do you need to know what kind of variables you are working with?

- If we want to know the "central tendency" of each type of variable, what would be the best measure? Hint: the mode is the most frequent response for a variable, but usually it makes sense to look at a frequency distribution because you could have two responses that are just about equally distributed in the data (for example, the mode for gender in most surveys would female, but saying that "female" was the most common response in the survey doesn't represent well what we see in most datasets, which is that close to half the sample identifies as male and about half as female).

- What is the difference between a "mean" and a "median," and why might one measure be preferable over another, depending on how "skewed" the distribution of a variable is? For instance, why is household income in the US typically reported as a median rather than a mean? What does it say that income in the US has a positive skew? Given that household income has a positive skew, will the median or the mean be a larger value??

- Means are much more useful than medians (i.e., they can be used in more statistical techniques). If a researcher wanted to analyze the mean income in the US (something that happens sometimes when income is used as a control variable), it might make sense to use a "trimmed (aka truncated) mean" to address the fact that the mean household income in the US has a "positive skew." What do the quoted terms mean? For the US, why would it make more sense to trim the mean of the income measure at $200,000 or higher rather than $70,000 and higher? What would be a good way to determine where to trim the data (hint: look at the distribution of income in a frequency chart to see where the skew begins)?

- Why do social scientists typically use "frequency" tables or figures to summarize the distribution of categorical (aka nominal) variables rather than giving just their mode, median, or mean? Why do you have to make sure to present frequency statistics in percentages rather than counts? What happens if you report the mean or median for a partisanship variable that has been coded 1 for Democrats, 2 for Republicans, 3 for independents, 4 for other parties, etc.? Which column do you look at in a frequency chart to get the correct statistic (hint: You want the "valid" percent column because it omits any observations where this variable has been coded as missing data.)

- Why is it so important to complement measures of a variable's "central tendency" with a second statistic that measures its variability (i.e., "dispersion")? Can you describe an example where two cities might have the same mean and median income but radically different standards of living because of how the income is distributed?

- What, in general terms, does the "standard deviation" statistic measure (when used in conjunction with a variable's mean)?

- Can you interpret a standard deviation? As an example, assumed that the mean score (quantitative, plus verbal sections) on a recent SAT was around 1050, with a standard deviation of around 200 points. SAT scores follow a normal distribution. What scores correspond to being one standard deviation below the mean? One standard deviation above it? Roughly what percentiles correspond to being one standard deviation above and below the mean? (Note: Saying a little more than one-third of takes would be 1SD above and a little more than one-third would be one SD below is close enough)

  What score (think 2 standard deviations among the mean) would put someone above the 95th percentile? (Actually, it is a bit higher than the 95th percentile because 95% of test takers would fall into the range of +/- 2 standard deviations from the mean, so someone with a score that is 2 standard deviations above the mean would be in the top 3%). Note that: Someone who is three standard deviations from the mean would be in the top or bottom 1%.

- What would it tell us about the test if the mean score for a future SAT was the same as noted above, but the standard deviation dropped to 10? If that happened, what score would put a student in the 97th percentile or higher?

- What does an interquartile or interquartile range measure (often one or the other is used in conjunction with income or SAT medians, for example)?

- How can a bar chart be used to explore the relationship between one or more independent variables and a dependent variable? Why does it make sense to create these kinds of charts in Excel rather than using SPSS to do so?

- This is a set of questions to make sure that you can apply the main concepts for talking about a study's research design: In your textbook, Chapter 7 (on descriptive statistics) summarizes a study by Besir Ceka and Pedro Magalães that was published in *Comparative Politics*. What is that study's dependent variable? What is the main independent variable? What is the theory about how the independent variable influences the dependent variable? In order to answer this last question, you need to be able to identify the study's intervening variable (hint: which is whether a person lives in a democracy or autocracy). What survey questions do the authors use to measure their independent variable and their dependent variable. What country-level variables are used to measure "the status quo," of a country (which is the intervening variable)? The summary in Chapter 7 doesn't mention the phrase "control variables," but four of them are listed in the descriptive statistics table on pp. 153-154. What are they?

- Now, let's see if you can use the main concepts for talking about a research design to think about a research project you are designing. If you are given an essay question like this one on your Unit 2 exam, be prepared to answer it well (Alternatively, you will be able to write about your research design proposal topic if you are prepared to do so):

  Assume you have been asked to conduct a study on the question of what kind of people think politicians need to be ethical and moral in their personal lives in order to be effective political leaders who make good decisions. For your project, you have located a nationally representative from the Pew Foundation that includes 1,500 respondents. Fortunately for you, the study has all of the variables you will need to operationalize the theory or theories you have in mind to answer your research question. Running a frequency on your dependent variable, you see that Americans as a whole are split on this issue. In a paragraph or so, describe a hypothetical research design for your project that includes:

  (1) at least one dependent variable,

  (2) at least two independent variables,

  (3) one or two theories explaining why you think that independent variables and dependent variable/s are connected.

  (4) at least two plausible, directional hypotheses, to test your theory or theories.

  (5) at least two control variables and an explanation for why that control should be included, and

  (6) at least one statistical method that you will use to test at least one of your hypotheses.

  After you have written the paragraph, label each component with the appropriate numbers listed above (i.e., put #1 next to the dependent variable/s and so on).