Public Policy 529
Fall 2023: Problem Set #1
R Version

Due Monday, September 11

At this stage of the course, a key goal is for you to begin working with your stats software. As a result, this problem set is much more software-oriented than those that you will do later. I say this because many students experience a sudden panic, thinking that they need to become proficient very fast. Indeed, one's first exposure does involve learning several new things all at once. With that foundation in place, however, we learn new things more incrementally.

Dominique is providing an overview of Stata and R in section. Additionally, please note the various resources on Canvas. First, in the Media Gallery, you will find a video that I produced to introduce students to R and RStudio. I provide a tour of the interface and a demonstration of how to use the software for basic descriptive statistics.

Second, in the Files section of Canvas, there is a help folder for R. There, you can find the "R-help" document that contains examples with explanations of pretty much every type of task that you will need for this course. This should probably be your first resource when you have questions.

Third, in the Pages section, there are links to several free resources for learning R.

We will use some data from the 2020 version of the American National Election Studies (ANES) dataset for this problem set. The data file, `anes2020subset`, is downloadable from the Canvas website for the course in the Files→Datasets directory. The version for R has the file extension .RData. Create a folder on the computer you are using that will serve as your "working directory." After you have downloaded the dataset, move it into this working directory.

Next, launch RStudio and change the working directory to be the folder you just created. It's easiest to do this through the menus at first. You will find "Set Working Directory" in the Session menu, where there is a submenu option for "Choose Directory..." This will bring up a directory navigation window. Navigate to where your working directory is located, click on it, and click "Open." When you complete this action, the command to set the working directory will appear in the Console window. Examine the syntax so that you can learn the command.

Next, you will need to load your dataset. Assuming it is sitting in your working directory, and you have successfully changed R's working directory to be that folder, you can issue a simple command to load the data: `load("anes2020subset.RData")`

If the dataset is not in your working directory, this will not work. You would either need to supply the full directory path to where the data is located, or you can use the File→Open menu option to navigate to where the dataset is located and open it. Once the dataset loaded properly, RStudio's Global Environment window will list `anes2020` as a data object. In RStudio, you can browse the data by typing the command `View(anes2020)` in the Console window. Note that the "V" must be capitalized.

I recommend that you first experiment with R by typing commands into the RStudio Console to make sure you are using the correct commands. As you learn the syntax that you need to perform the correct steps for this problem set, you can copy and paste the commands into your R script. This file provides a set of instructions for R. The goal for you is to create and run a script that will produce an error-free rendition of all the analysis in this problem set. When you are done, save this file and turn it in with your problem set answers.

To open a new R script, click the new file icon, which is in the upper left (it has a plus sign over a blank document), then select "R Script." Or, choose File→New File→R Script from the menus. This brings up a blank text document in RStudio's upper left pane. You can type your commands into this document, or you can copy and paste them from the Console window. A sample R script is contained in the R Help on Canvas in the Files section.

1. How many observations does the dataset have? In RStudio, this information is available in the Global Environment window or through the command `nrow(anes2020)`, which tells R to report the number of rows in the data frame `anes2020`.

2. Find the answers to the following questions.

   (a) Let's examine the variable `SciImptCovid`, which asks how important science should be for decisions about Covid-19. Specifically, we will use two functions that provide basic information about a variable's distribution: `summary()` and `table()`.

   It is important to note that the output of these functions will depend on whether the variable is categorical or quantitative. In R, categorical variables are of a class called "factor," while quantitative variables are of a class called "numeric." To identify the class of a variable, you can use the function `class()`. For example, `class(anes2020$SciImptCovid)`.

   First, run the command `summary(anes2020$SciImptCovid)`. From the Console window, copy both the command and the resulting output, and paste it into your answers. If you are familiar with RMarkdown and would like to use that instead, please feel free.

   Note that, inside the parentheses, you first identify the data frame that contains the variable, then you put the name of the variable after a dollar sign. Since R allows us to have multiple data frames in memory, we need to specify which data frame we are using.

Second, run this command: `table(anes2020$SciImptCovid)`. Again, copy and paste the command and output into your answers. Continue to do this for all the commands you run to answer questions for this problem set.

Compare the output from these two commands. In your answers, identify the main difference.

(b) In R, the symbol `NA` means missing data. These respondents either refused to answer, could not be reached for the post-election survey, or broke off the interview before this question.

Answer the following questions. Should we be concerned about these cases with missing data when we want to perform analysis with this variable? What could potentially be a problem?

Some background: the sampling methodology for this survey is quite complicated. Some respondents completed the survey online, while others completed it partially online and partially via phone or video interviews. Additionally, the survey took place in two waves: a pre-election survey and a post-election survey. The goal was to interview all respondents both times, but there were 827 respondents to the pre-election survey that did not participate in the post-election survey. Note that the `ScienceExperts` question was asked in the post-election survey. If would like to read more about the survey design, check out the ANES 2020 Codebook file that is in the Files → Datasets folder on Canvas.

(c) Which value of `SciImptCovid` was the most common response? This is the mode of the variable.

(d) We have discussed three basic ways to describe a variable's measurement level: nominal, ordinal, and interval-level. How can we describe the measurement level of this variable? Ignore the missing values.

3. How can we describe the measurement scale (i.e. metric) for each of the variables below? Categorical or quantitative? Nominal, ordinal, or interval-level? Discrete or continuous?

   To see the variable's values, use the `table()` function with each variable name.

   (a) EconWorse

   (b) DiscussPol

   (c) HomeOwnership

   (d) VotedBiden

   (e) Empathy

4. For each of the five variables listed above, identify which measures of central tendency are available. You do not have to calculate or report them.

5. Now use the `summary()` and `table()` functions with the variable `SCOTUStherm`.

   Note: this variable is a "feeling thermometer." The respondent is asked to say how warmly (or coldly) they feel about a person, group, or organization using a 0-100 degree temperature scale. They are told that 50 degrees means neutral feelings. Let's assume that we can treat this variable as having an interval-level measurement scale. In this case, the feeling thermometer measures how the respondent feels about the Supreme Court of the United States.

   Take note of how the output for the `summary()` function differs for `SCOTUStherm` compared to `ScienceExperts`. This is because `SCOTUStherm` is of the numeric class.

   (a) Using the information provided by the commands, find the value of each appropriate measure of central tendency.

   (b) What is the standard deviation of this variable?

   (c) Make a histogram showing the distribution of `SCOTUStherm` and paste this figure into your answers. The command is: `hist(anes2020$SCOTUStherm)`.

6. For this question, you will work with the variable `SpendHighways`, which asks respondents about their opinion regarding the level of government spending on highways.

   (a) Using the `table()` function, produce the frequency distribution of the variable `SpendHighways`. What is the measurement level of this variable?

   (b) Make a bar graph that shows the frequency distribution. Copy and paste it into your answers. The command is `barplot(table(anes2020$SpendHighways))`. If you wish to learn how to make different labels for the bars and a title for the graph, consult the R help document.

   (c) Since `SpendHighways` is of the factor class (i.e. categorical), R will not use this variable in functions that involve mathematical operations. We could, however, use the function `as.numeric()` to have R treat it as numeric. For example, we can embed the `as.numeric()` function inside the `table()` function. The command becomes `table(as.numeric(anes2020$SpendHighways))`. Make that table and note how it compares to what you produced for part (a).

   For `SpendHighways`, does the ordering of the numbers have meaning in terms of the variable? Beyond order, can we think of these numbers as measuring an amount of support for spending, such that each step on the scale is the same amount of change?

   (d) With `SpendHighways` stored in the factor class, R forces you to think before using functions that would do mathematical operations with this variable. The `summary()` function will report only a frequency distribution, for example. If we insert the `as.numeric()` function inside `summary()`, however, things change. Report the results from `summary(as.numeric(anes2020$SpendHighways))`.

   As you know (or will soon learn), the formulas for the mean and standard devia-

tion involve calculations using the variable's values. Why might that be problematic in this case?

Reminder: I would like you to write an R script file that performs all of the needed commands for the analysis above. Include this script file with your answers.

7. Consider two methods to measure democracy in countries. In the first method, a country is coded as a democracy if there are elections and these elections periodically produce a transfer of power to a new person or party. The country is coded as a non-democracy otherwise. In the second method, a set of experts is asked to rate each country on a seven-point scale according to the degree to which the civil liberties and political rights of citizens are protected. On the scale, a 1 means that there there are no protections for civil liberties or political rights, and a 7 means that there are extensive protections for civil liberties and political rights.

   (a) What are the measurement levels of the variables produced by these two measurement strategies?

   (b) Which measurement strategy is more reliable? Explain your reasoning.

   (c) Which measurement strategy is more valid? Explain your reasoning.

8. Consider two methods to measure whether a political candidate received favorable coverage from the news media. In the first method, a random sample of news stories would be split among a team of graduate students. After reading a story, the student would would assign a code indicating whether it was "favorable," "neutral, or "unfavorable" toward the candidate. In the second method, computer software would be used to code each news story by scanning the text to count the appearance of key words that that the researcher has determined to be "negative" words.

   (a) What are the measurement levels of the variables produced by these two measurement strategies?

   (b) Which measurement strategy is more reliable? Explain how you know.

   (c) Which measurement strategy is more valid? Explain how you know.