Methods Camp Assignment 1

Day 1

Methods Camp Teaching Staff

Today you will be working with data from the American National Election Studies (ANES) from 2016. You will specifically be exploring how a respondent's degree of opposition to free trade is related to their views about three presidential candidates at the time: Trump, Sanders, and Clinton.

Please put your answers in these callout boxes, and make sure that when you want us to check your code, you use `echo = T`

> Answer
>
> Answers go in boxes like these so Sofia and Christina can check them easily!
>
> ```
> # code chunks should go in with echo = T so we can check your work!
> ```

Loading and examining the data

- Load the anespilot16.csv data, view its dimensions, check what type (class) of object it is, and view the first few rows and first 10 variables.
- Load the dplyr and ggplot2 packages.
- Come up with one question you have about the data and show which functions you used to answer it.

Create variables of interest to review different data types

Create a vector, *varsinclude*, that contains the following variables of interest for our analysis:

Feelings about free trade:

- freetrade: views about free trade

Rankings of which issues are most important:

- ISSUES_OC14_10: where the respondent ranks unemployment as an important issue (out of 21 issues)

Feelings about candidates:

- ftsanders: feeling thermometer for Sanders
- fttrump: feeling thermometer for Trump
- fthrc: feeling thermometer for Hillary Clinton

Demographic variables:

- gender
- race
- educ
- birthyr
- weight

Check your intuition

What type (i.e. numeric, logical, etc.) of vector should *varsinclude* be? Confirm your answer by using the "class" command.

Creating a new data frame

Use either the appropriate dplyr command to create a new data.frame, *anes2*, that only includes those relevant variables. *Hint: if using dplyr, the command all_of may be useful.*

Then, check the dimensions (should be 1200 x 9).

Creating new factor variables

Now, we're going to work on getting the variables into more usable form.

Using the codebook for the study and the appropriate R command, create a new data.frame (*anes3*) that is the same as *anes2*. Then in *anes3*, create a new factor variable-racenew- that collapses the racial groups into the following categories and labels them appropriately:

- 1 = white
- 2 = black

- 3 = hispanic
- 4 = other (includes all non-hispanic categories)

Save the results into a new data.frame, *anes3*.

Relabeling variables

Now we're going to clean some of the other variables.

Use the codebook and the appropriate R command in dplyr, create labels for the levels of the following variables. When creating these labels, it's up to you whether you want to create new variables or just save over the existing names.

- gender
- educ
- freetrade

Save the results in *anes3*.

Practicing logical statements

To make sure the continuous variables are usable, use the "summary" command to view the range of the following heat thermometer variables we will be examining (make sure to use indexing so it only summarizes those three variables):

- fttrump
- ftsanders
- fthrc

The feeling thermometers should go from 0-100, but what is the range of these three variables?

Use the codebook to:

1. For each of the variables, recode values above 100 to NA
2. Exclude from the data observations that have NA for at least one of the heat thermometer variables - there should be 1188 observations remaining

You want to see if there's a meaningful enough subset of respondents who oppose free trade to look at whether they're more likely to support Trump or Sanders than Clinton.

Recoding variables with conditional statements

Next, create a new variable, `clintonsupport` that is set to 1 if 1) the respondent has a feeling thermometer rating for Clinton that is greater than 50 or 2) the respondent's feeling thermometer rating for Trump and Sanders is lower than for Clinton. Otherwise, set `clintonsupport` to 0. Verify that the sum of `clintonsupport` is 560.

Often times, we will want to convert age, a continuous variable, into a categorical one by using age groups. Estimate the age of the respondent by assuming all were born January 1st of their birth year, and create a new variable called `age` (hint: survey was fielded in 2016). Then, create a new variable, `age_group`, that categorizes the respondent into the following age groups:

- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65+

Use `table()` to print the number of respondents in each age group.

Descriptive exploration

Print the number and percentage of respondents who oppose free trade (includes those who oppose free trade a little, a moderate amount, and a lot).

How many and what % of respondents oppose free trade?

Digging Deeper

Going back to the *anes3* data, do the following:

Create a numeric version of the freetrade variable called freetradenum- this will allow you to take the mean

Use dplyr to find the mean views about trade (higher = more opposition) for each educational category, using *na.rm* in the *mean* command to exclude NA's from this calculation (the *mean* command won't run if you have missing values), and then to rank the education groups from most opposed to least opposed.

Which educational groups are most opposed to free trade?

Summarizing the data

Now, use dplyr to find the mean thermometer rating for Trump, Sanders, and Clinton by education group, creating three summary measures using the *summarise* command: *trumpheat*, *sandersheat*, *hilheat*. Display the data so that rows are education categories and the columns are the candidates.

You're also interested in descriptively exploring what role gender might play in these views– find the mean degree of opposition to free trade for each education and gender category (e.g., males with 2 year, females with 2 year, males with some college, etc.) using dplyr

Then, order from most opposed (highest score) to least opposed (lowest score)

What patterns, if any, do you notice?

You're confused about why postgrad females seem one of the most opposed to free trade, and guess it has something to do with sample size – add a column to print the number of individuals in each group.

Matrix Review

standardize weighting

Often, when we are working with surveys, we might want to use weights to make sure that our sample is representative of the population. In this case, we have a variable called *weight* that we can use to weight our data. A weighted average can be expressed as:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where w_i is the weight for each observation and x_i is the value of the variable we are averaging.

Calculate the weighted average of the feeling thermometers for Trump, Sanders, and Clinton. Hint: use the *weighted.mean* command to do this:

Now, let's repeat this calculation but **only** using matrix operations. While this is not the clearest way to do this, sometimes when you're working with large datasets, it can be useful to use matrix operations to speed up calculations.

Matrix `ft` below is a matrix that contains the feeling thermometers for Trump, Sanders, and Clinton. The vector `w` is the weight variable.

```
# comment out the following lines to run the code
#ft <- as.matrix(anes3[, c("fttrump", "ftsanders", "fthrc")])
#w <- as.matrix(anes3$weight)
```

Calculate the weighted averages using `%*%`. Hint: begin by using `dim()` to check the dimensions of `ft` and `w` to make sure they are compatible for matrix multiplication. If they're not, what do you need to do to make them compatible?