Public Policy 529
Fall 2023: Problem Set #2

Due Tuesday, September 19

1. Provide an example of a variable for which the mode applies as a measure of central tendency but the mean and median do not.

2. Each of these terms is a value of a variable: 10 hours, Buddhist, strongly disagree, 3 classes, citizen. Provide the measurement level of the associated variable.

3. A researcher is studying pollutant levels in the area around a mining operation. Her set of 8 soil samples shows the following levels of arsenic (measured in parts per million):

   4, 5, 2, 4, 8, 35, 7, 5

   (a) Using this sample, calculate the mean, median, mode, range, interquartile range, variance and standard deviation. Do not use built-in software functions to find these statistics, as the point is for you to understand what the computer is doing. You may use your software as a kind of calculator if you wish.

   (b) Which of these measures above are sensitive to extreme values?

   (c) What measure of central tendency do you think is the best summary of the distribution? Explain your answer.

   (d) Suppose that 3 of these soil samples were taken from parks (coded as 1), while the other 5 were not (coded as 0). Find and interpret the mean, median and mode of this dichotomous variable.

4. In the 2014 General Social Survey, respondents were asked how many adults (age 18+) live in their household. The frequency distribution of the variable looks like this:

	Frequency	Percentage	Cum. %
1	828	32.74%	32.74%
2	1,318	52.12%	84.86%
3	264	10.44%	95.29%
4	89	3.52%	98.81%
5	23	0.91%	99.72%
6	6	0.24%	99.96%
7	1	0.04%	100.00%
Total	2,529	100.0%	

(a) What is the measurement level of this variable?

(b) Use all appropriate measures of central tendency to describe this sample distribution.

(c) Which of these measures of central tendency do you think is best for communicating the story of this variable's distribution? Why?

(d) What is the interquartile range of this distribution?

(e) Would you describe this distribution as having skewness? If so, what kind?

5. For this question, use the `anes2016subset` dataset.

(a) The variable `NewsAttn` indicates the respondent's self-reported level of attention to the news. In R or Stata, produce a frequency table and use it to calculate the proportion of the sample that either pays "a lot" or "a great deal" of attention to the news.

Note: connecting this to probability theory, we might think of these two outcomes as comprising an event: respondent is a heavy news consumer. The proportion you just calculated represents the probability that a randomly selected member of the sample is a heavy news consumer. We might write that as P("a lot" or "a great deal").

Note to R users: you can use commands from Problem Set 1 or you may want to consider reading section 3.4 of the R-help document on Canvas to learn how to use the `questionr` library to create tables with relative frequencies and cumulative percentages like those from Stata's `tabulate` command.

(b) What is the probability that a person selected at random from this sample pays "a moderate amount" of attention to the news or less? How does this probability relate to what you calculated in part (a)?

(c) If we randomly select a respondent, what is P("a little")? What is P(~ "a little")?

(d) Now make a table that shows the joint frequency distribution of two variables: `NewsAttn` and `WrongTrack`. The variable `WrongTrack` indicates whether the respondent believes that the United States is "going in the right direction" or is "on the wrong track."

In both Stata and R, we just add another variable to the `tab` and `table()` commands respectively. For Stata, the command is `tab NewsAttn WrongTrack`.

For R, we can insert the `table()` function inside a function called `addmargins()` to add row and column totals. The R command is: `addmargins(table(anes2016$NewsAttn, anes2016$WrongTrack))`.

(e) In this sample, what proportion of the respondents believe the country is on the wrong track?

(f) Among those who pay "a lot" or "a great deal" of attention to the news, what is the probability that a randomly selected respondent says the country is on the wrong track?

(g) For a randomly selected respondent, what is P("wrong track" or "a great deal")? What is P("wrong track" and "a great deal")?

6. Using the anes2016subset dataset, make a box plot that shows the distribution of the variable BLMtherm within each of the categories of the variable PartyID, which indicates the political party affiliation of the respondents. Describe any differences you see in the central tendency and dispersion of the BLM thermometer between these categories.

   The Stata command is: `graph box BLMtherm, over(PartyID)`. The R command is: `boxplot(BLMtherm ~ PartyID, data = anes2016)`. Note: copying and pasting the R command will produce an error due to hidden code to make the tilde. Just type it. Also, there is no hyphen in `boxplot`.