**Problem Set 1**

This problem set covers material from the first three classes including material from OIS Chapter 1 and Chapter 2 Section 2.1. It makes use similar data to that we discussed in class. Partial credit may be given for answers that are correct in part, but not in full. This problem set is due on Gradescope by Wednesday September 1, 2021 at 11:59 PM.

**Part I: Who Runs for Office and who wins? (22 pts.)**

The 2018 election saw candidates from a wide array of backgrounds run for office. For example, in [this](https://fivethirtyeight.com/features/democrats-primaries-candidates-demographics/) *FiveThirtyEight* piece, Meredith Conroy, Mai Nguyen and Nathaniel Rakich note that “2018 is a new ‘Year of the Woman’ in Democratic primaries. The questions in this section will pertain to this article.

1. First let’s consider the design of the study. (14 pts.)
   1. If you would like to determine if a candidate’s gender relates Democratic primary voters’ vote choice (i.e., if they vote for a female candidate or a non-female candidate), would this be an experiment or an observational study? How do you know? Could you infer causality here? (3 pts.)
   2. In the above study, what is the response variable and what is the explanatory variable? (2 pts.)
   3. What type of variable is each of these variables in terms of how they would be coded in the dataset (i.e., vote choice and candidate’s gender)? Be as specific as possible and tell the reader how you know this. What if we were looking at overall vote percentage for a candidate instead of vote choice? (6 pts.)
   4. Let’s say that you (with permission from your university) invited 1,000 Democratic voters to a lab where they read a randomly assigned vignette about two candidates, one male and one female, and they were asked which theoretical candidate they preferred. What kind of study does this appear to be? How do you know? Could you infer causality here? (3 points)
2. Next, you are going to look at the spread for some of the other variables in the study. (8 pts.)

|  |  |
| --- | --- |
| *Figure 2.1* | *Figure 2.2* |
|  |  |

* 1. Figure 2.1 above is a histogram of the primary vote percentage of candidates with a STEM background. Is this plot skewed or symmetric (and if skewed, in what direction)? Unimodal or bi/multimodal? For both skewness and number of modes, how do you know? (4 points)
  2. Figure 2.2 above is a histogram of the primary vote percentage of candidates who had support from the Democratic party organizations such as the Democratic Congressional Campaign Committee (DCCC) before the primary took place. Is this plot skewed or symmetric (and if skewed, in what direction)? Unimodal or bi/multimodal? For both skewness and number of modes, how do you know? (4 points)

**Part II: Thinking Like a Pollster (15 points)**

A pollster would like your help in designing a study for the upcoming Durham municipal elections this November to determine voter attitudes. (15 points)

1. The pollster presents various research designs.
   1. To save money, the pollster would like to ask the opinions of people on the website of the progressive newspaper *IndyWeek.* What kind of sample is this? Is this approach problematic? Why or why not? (3 points)
   2. Next, the pollster comes back with a completely different approach. They would like to ask every Durham voter their opinion on the election. What would you call this approach? Is this approach problematic? Why or why not? (3 points.)
   3. Now, the pollster would like to go with a midpoint approach by asking 1500 Durham registered voters their opinions with everyone having an equal chance of being selected. What kind of sample is this? Does this approach have any potential biases that could arise? Which is especially concerning? (3 points)
   4. The researcher would like your input on other possibilities. Name two other possibilities for how they could get a sample, how the pollster would carry out the poll, and a possible issue that could arise with that approach. (6 points)

**Part III: Design an Experiment (10 pts.)**

1. You have been asked to design an experiment on a topic of your choice. Discuss how you would go about doing that. In this answer, please present a null and alternative hypothesis and discuss the four principles of experimental design. Please also discuss blinding, the importance of confounding variables, and whether you could infer causality from your experiment. (10 pts.)

**Part IV: Data on Voter Turnout**

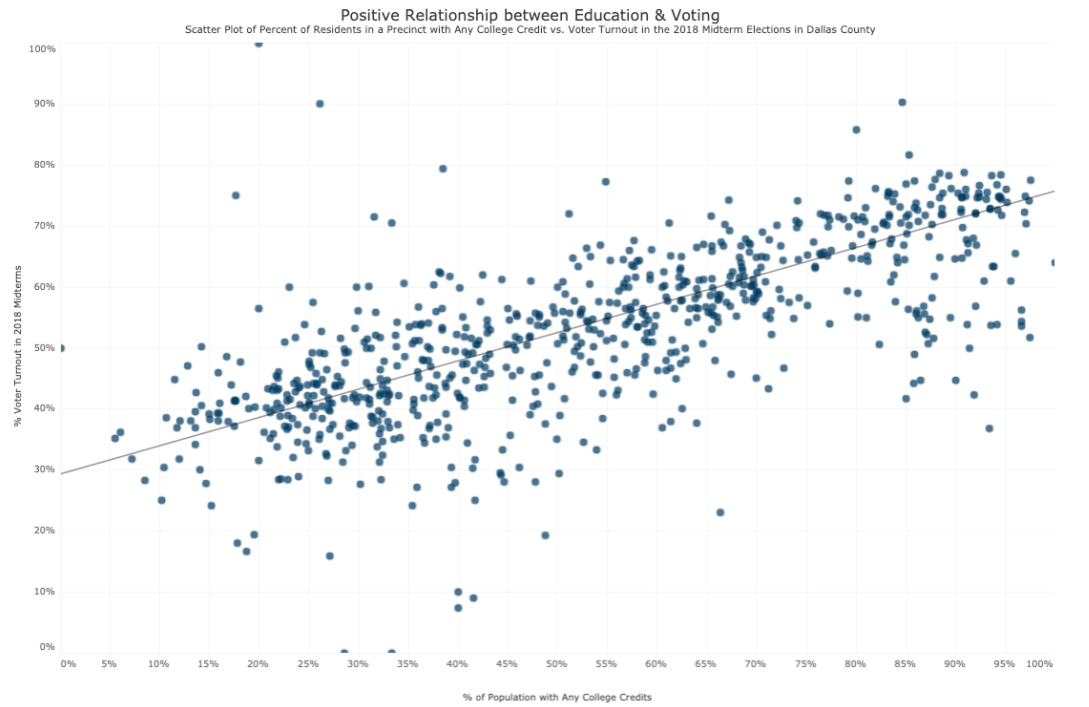
Turnout in the 2018 midterm and 2020 presidential elections rose substantially compared to other recent elections. Using data from the [US Elections Project](http://www.electproject.org/home/voter-turnout/voter-turnout-data), I have constructed the below dot plot and box plot of voter turnout in this millennium by type of election (midterm or presidential) election starting with the 2000 presidential election. I would like you to consider how voter turnout (of the voting eligible population) differs by type of election.

|  |  |
| --- | --- |
| *Dot plot* | *Boxplot* |
|  |  |

1. Please consider the above dot plot and box plot. (17 pts.)
   1. What does this dot plot suggest about turnout by type of election? Provide evidence for your conclusion. (3 pts.)
   2. Now, please turn to the box plot. What do the three lines on the box of a boxplot represent (theoretically, you don’t need to discuss actual values) and why are there only two lines for the midterm data? What does the relative location of the lines for the midterm data relative to the presidential data allow us to conclude about how turnout differs across elections? (6 pts.)
   3. What does the presence of a small whisker above the presidential box plot mean? What does the lack of whiskers elsewhere imply? (4 pts.)
   4. There are dots both above and below each box plot. What do these dots mean and what does the level of the dot above the midterm box plot relative to the dot below the presidential box plot allow us to conclude? (4 pts.)
2. Below are the values of turnout for *midterm* elections since the start of this millennium (rounded to nearest integer). You will be working with this data for this question. (26 pts.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | 2002 | 2006 | 2010 | 2014 | 2018 |
| Turnout | 40 | 40 | 41 | 37 | 50 |

1. Please calculate the mean, median, and mode for midterm turnout. Please show your work or show how you found your answer. (9 pts.)
2. Please calculate the variance and standard deviation for midterm turnout. Please show your work as you find these answers. (6 pts.)
3. 2018 had much higher turnout than any of the other years here. What type of data point does the box plot in the previous question suggest 2018 is here? Would we want to automatically exclude it? Why or why not? (3 pts.)
4. Let’s say we did exclude it. What happens to the mean, median, and mode? Please show your work or explain how you got your answer. What does this suggest about the robustness of the mean and median as measures of central tendency? (8 pts.)
5. The [Commit Partnership](https://commitpartnership.org/about-us) is a non-profit in Dallas, Texas that, according to their website “work(s) to ensure that all North Texas students receive an excellent and equitable education that prepares them to flourish in college and career.” A [post](https://commitpartnership.org/blog/how-to-address-low-voter-turnout-invest-in-education) from 2019 on their website tries to draw a link between education and voter turnout. They provide the below scatter plot of percentage of voters in a Dallas County precinct with college credits on the x-axis and percentage who voted in the 2018 midterms on the y-axis.



* 1. Which of these is the explanatory variable and which is the response variable? Is any association you might find necessarily causal in this observational data? (2 points)
  2. There are four things to consider when evaluating the relationship between numerical variables. What are each of these and how to they apply to this data? (8 points)