**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.

**Part I: Cherie Berry’s Elevator Pictures (30 points)**

In North Carolina, Labor Commissioner Cherie Berry placed her [picture](https://en.wikipedia.org/wiki/Cherie_Berry#/media/File:Cherie_Berry.jpg) on official elevator placards. [Researchers](https://journals.sagepub.com/doi/abs/10.1177/1532673X15602755) found that Berry performed better in counties with more elevators in 2012. Berry won reelection again in 2016. You are tasked with helping a researcher examine whether she continued to do well in elevator-dense counties in 2016.

1. First, the researcher would like to think about the research design of this study by answering the following questions. (13 points)
   1. Is this an observational study or an experiment? How do you know? Will we be able to infer whether a higher concentration of elevators caused better to receive a higher vote share in a county? (3 points)
   2. The researcher thinks that the number of elevators per 1,000 people affects the percentage of vote she receives in a county. Which of these is the response variable? Which is the explanatory variable? (2 points)
   3. What type of variable is each of these variables? Be as specific as possible and tell the reader how you know this. (4 points)
   4. Imagine we were to look instead at the exact number of elevators in a county and the exact number of votes Berry received. What type of variable would each of these be? How do you know? (4 points)
2. Next, the researcher would like you to look at the spread of the some of the variables that are relevant to the study. (8 points)

*Figure 2.1 Figure 2.2*

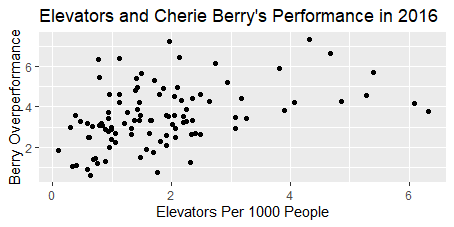
|  |  |
| --- | --- |
|  |  |

* 1. Figure 2.1 above is a histogram of the number of elevators per 1,000 people in North Carolina counties in 2016. 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)

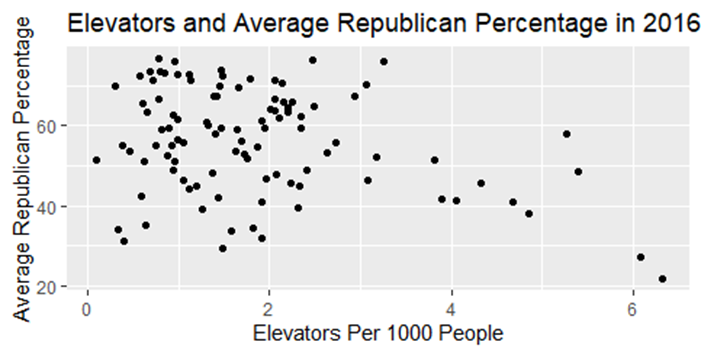
* 1. Figure 2.2 below is a histogram of the average percentage received by Republican candidates for state office in North Carolina counties in 2016. Is this plot skewed or symmetric (and if skewed, in what direction)? Unimodal or bi/multimodal? For both skewness and number of modes, how do you know? (4 points)

1. Now, the researcher has created a series of plots looking at the association between various variables and would like you to characterize the relationships. (9 points)

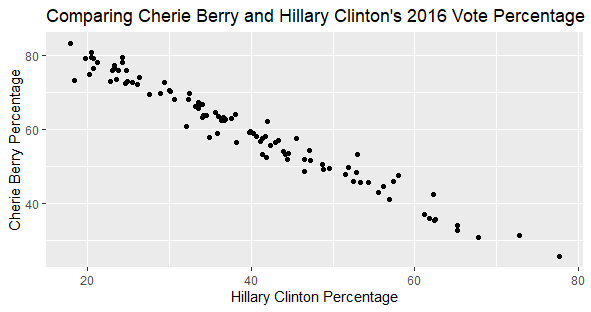
***Figure 3.1***



***Figure 3.2***



***Figure 3.3***



* 1. Figure 3.1 looks at the relationship between the number of elevators per 1,000 people in a county and Berry’s overperformance from the typical Republican. What is the direction of this association? How strong is it? (2 points)
  2. Now, the researcher wants to see how well the typical Republican does in counties with a lot of elevators. Figure 3.2 looks at the relationship between the number of elevators per 1,000 people in a county and the average Republican percentage. What is the direction of this association? How strong is it? (2 points)
  3. Finally, the researcher is interested in how Berry’s relationship compared to how well Democratic presidential candidate Hillary Clinton performed in a county in 2016. Figure 3.3 looks at the relationship between these two variables. What is the direction of this association? How strong is it? (2 points)
  4. You may have noticed two data points in Figure 1 with more than 6 elevators per 1,000 people where Cherie Berry overperformed by less than in several counties that you might expect based on the trend for the rest of the data. Similarly, in Figure 2 these two cases are apart from the rest of the data, with a far smaller average Republican vote percentage than other observations. What might we suspect these two counties to be? Should we immediately throw out these cases? Why or why not? (3 points; aside: these two cases happen to be Durham and Orange counties, where Duke and UNC are located).

**Part II: Thinking Like a Pollster (27 points)**

A pollster would like your help in designing a study for the upcoming presidential election to determine voter attitudes. (15 points)

1. First, the pollster would like your input about a variety of possible research designs.
   1. To save money, the pollster would like to ask the opinions of people on the progressive blog *Daily Kos.* 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 American 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 American 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)
2. Now, the pollster is interested in the effect of negative television advertisements on voters and has gotten permission from her university to carry out an experiment. (12 points; let’s imagine this is before COVID-19 when it would be okay to gather a group of people in a room.)
   1. The researcher will divide into a treatment group and a control group. One group will be shown a negative television advertisement, while the other will be shown a non-political advertisement. Which group will be shown which? What would you call the non-political advertisement? Why would you show this group anything? (4 points)
   2. How should you assign subjects to the treatment and control groups? (1 point)
   3. Voters who already support the candidate may respond to a negative advertisement differently than those who oppose them. How might you go about addressing this when assigning people to groups? What is this approach called? (2 points)
   4. Something has gone wrong! The air conditioning has broken in the room where the negative political advertisement is to be shown. Is this potentially problematic? Why or why not? (2 points)
   5. If the pollster finds support for their theory about advertisements, should they potentially do the same study again? Why? What is this called? (3 points)

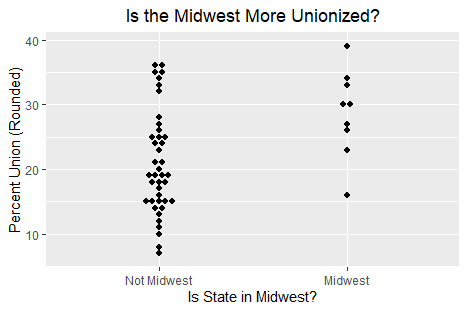
**Part III: Unions and Economic Inequality (43 points)**

A labor historian would like your help on a project about unionization and economic inequality in the mid-1970s. They are especially interested in looking at relationships between unionization and inequality and whether patterns of unionization are different in the Midwest than in the rest of the country.

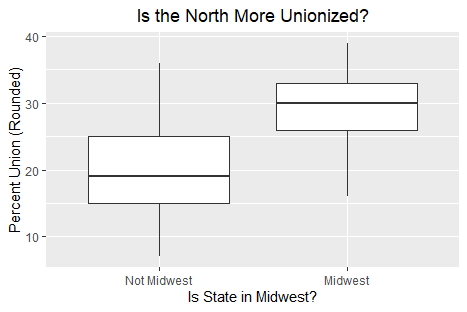
1. The researcher made a scatterplot comparing the percent of people in a union in 1976 to income inequality (measured using the Gini Index, which in theory goes from 0 to 1; question is worth 11 points).
   1. Which of these is the explanatory variable and which is the response variable? Is any association you might find necessarily causal in this observational data? (3 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) (Hint: I’ve provided the graph both with and without a line going through the data to help with this.)

|  |  |
| --- | --- |
|  |  |

1. Michigan had the highest rate of unionization in 1976. The labor historian wonders if this pattern applies to other states in the Midwest. (Here, I define Midwest as states were more than 50 percent of respondents classified the state as such in a 2014 poll for [*FiveThirtyEight*](https://fivethirtyeight.com/features/what-states-are-in-the-midwest/).) (32 points)
   1. First, the researcher makes a dot plot looking at unionization (rounded to the nearest whole number) by region. What does this dot plot suggest about unionization by region? Provide evidence for your conclusion. (2 points)



* 1. Next, the researcher makes a box plot. What do each of the three lines on the box represent? (Note: I’m looking for a conceptual answer here, rather than an exact number.) Looking at the level of each of these lines by region, what does this suggest about unionization in the Midwest compared to the rest of the country? (4 points.)



* 1. The boxplot also has whiskers that extend outward. How far does a whisker extend? What does it mean that there are no dots beyond the whisker for either region? (2 points.)
  2. The researcher would like you to do some calculations of some measures of central tendency of unionization data from the Midwest in 1976 that are available in the below table. Please calculate the mean, median, and mode. Please show your work or justify your answer when relevant. (6 points).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| State | IA | IL | IN | KS | MI | MN | MO | OH | WI |
| Union % (Rounded) | 26 | 30 | 34 | 16 | 39 | 27 | 23 | 33 | 30 |

* 1. The researcher would now like you to calculate several measures of spread: the sample variance and standard deviation. Please show your work. Finally, please discuss two purposes of squaring the deviations of data points from the mean and why a researcher might still prefer to use the standard deviation. (7 points)
  2. According to [*FiveThirtyEight*](https://fivethirtyeight.com/features/what-states-are-in-the-midwest/), some respondents considered Pennsylvania to be a Midwestern state. 35 percent of Pennsylvanians were unionized in 1976. What happens to the mean, median, and mode when you add in Pennsylvania to the set of cases considered to be in the Midwest? Please show your work. Finally, consider what this shows about the robustness of the mean and median as measures of central tendency. (7 points)
  3. There appear to be no outliers in this data. Using the definition of Midwest that includes Pennsylvania, how large or how small do observations have to be to count as outliers? Please show your work. (4 points)