**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 Berry 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 Berry 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)

* 1. 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)

1. Next, the researcher would like you to look at the spread of the some of the variables that are relevant to the study.

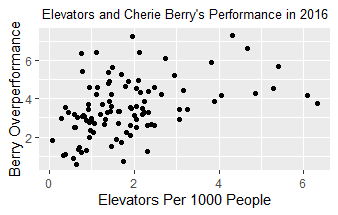
*Figure 2.1 Figure 2.2*

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

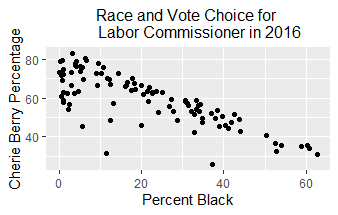
* 1. Figure 2.1 above is a histogram of the number of elevators per 1,000 people in North Carolina counties in 2016. Is this plot skewed or symmetric (and if skewed, in what direction)? Unimodal or bi/multimodal? For both skewness and number of modes, how do you know? (4 points)
  2. Figure 2.2 above is a histogram of the percentage received by Cherie Berry 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)

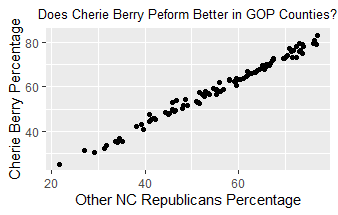
***Figure 3.1***



***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. Is there an association and in what direction? How strong is it? (2 points)
  2. Now, the researcher wants to see how well Cherie Berry does in counties with a high percentage of Black voters. Figure 3.2 looks at the relationship between the percentage of voters who are Black and Cherie Berry’s percentage. Is there an association and in what direction? How strong is it? (2 points)
  3. Finally, the researcher is interested in how Berry’s relationship compared to how well other North Carolina Republican candidates for state office performed in a county in 2016. Figure 3.3 looks at the relationship between these two variables. Is there an association and in what direction? How strong is it? (2 points)
  4. You may have noticed two data points in Figure 3.2 where less than 50 percent of the population is Black, but Berry received 25-32% of the vote. This is a far lower percentage that other counties where a similar percentage of voters are Black. 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: 538’s New Look (27 points)**

This year, [*FiveThirtyEight*](https://fivethirtyeight.com/features/how-we-designed-the-look-of-our-2020-forecast/) redesigned the look of their forecast page. Imagine if they would like to see if readers like this year’s look or if they prefer the old look. (15 points)

1. First, *FiveThirtyEight* would like your input about a variety of possible research designs.
   1. To save money, *FiveThirtyEight* would like to ask the opinions of people by using a Twitter poll*.* What kind of sample is this? Is this approach problematic? Why or why not? (3 points)
   2. Next, FiveThirtyEight comes back with a completely different approach. They would like to ask every single reader their opinion on the new look. What would you call this approach? Is this approach problematic? Why or why not? (3 points.)
   3. Now, *FiveThirtyEight* would like to go with a midpoint approach by asking 1500 readers 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, *FiveThirtyEight* would like to get readers’ opinions on the inclusion of forecast mascot [Fivey Fox](https://www.tweet247.net/united%20states/fivey+fox) on the new forecast and has gotten permission 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 forecast with Fivey Fox, while the other will be shown the forecast without Fivey Fox. Which group will be shown which? What might you give the group that doesn’t see Fivey Fox? 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 have a favorable opinion of foxes may respond 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! Five foxes ran through the room where viewers were to be shown the forecast with Fivey Fox. Is this potentially problematic? Why or why not? (2 points)
   5. If *FiveThirtyEight* finds that readers like Fivey Fox, should they potentially do the same study again if they really want to be sure about their result? Why? What is this called? (3 points)

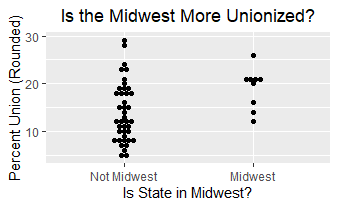
**Part III: Unions and Economic Inequality (43 points)**

A labor historian would like your help on a project about unionization and economic inequality in the early 1990s. 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.

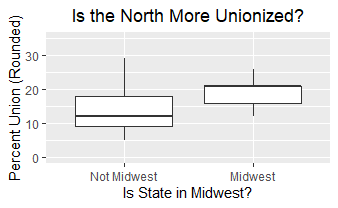
1. The researcher made a scatterplot comparing the percent of people in a union in 1990 to income inequality (measured using the Gini Index, which in theory goes from 0 to 1; question is worth 11 points).
   1. Which of these is the explanatory variable and which is the response variable? Is any association 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; these four things come from the Coursera videos.)

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

1. Michigan had the highest rate of unionization nationwide in 1976, but by 1990, both Hawaii and New York had a higher rate of unionization. The labor historian wonders if the Midwest is still the cradle of unions it was a few decades earlier. (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) | 16 | 21 | 20 | 12 | 26 | 21 | 14 | 21 | 21 |

* 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. 21 percent of Pennsylvanians were unionized in 1990. 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 does not include Pennsylvania, how large or how small do observations have to be to count as outliers? Please show your work. (4 points)