**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. In 2020, Berry retired and Republican Josh Dobson won the election to replace her. Researchers are interested in exploring Dobson’s performance compared to Berry’s in order to see whether her appeal in counties with lots of elevators truly was unique.

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 Dobson 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 change in vote between 2016 and 2020 (i.e., how Dobson compared relative to Berry). 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 Dobson received relative to Berry. 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.

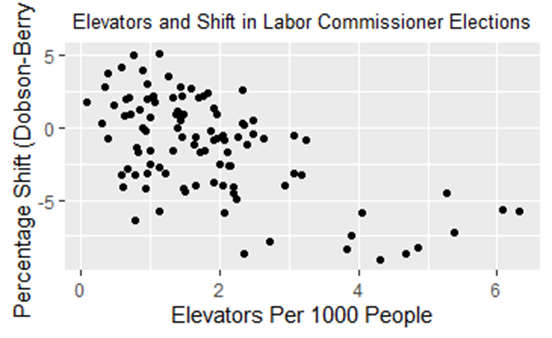
*Figure 2.1 Figure 2.2*

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

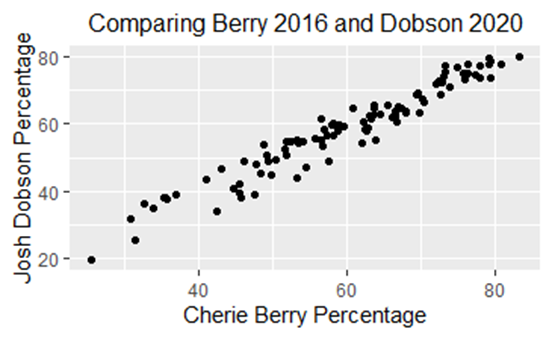
* 1. Figure 2.1 above is a histogram of the number of elevators per 1,000 people in North Carolina counties in 2016. Is this plot skewed or symmetric (and if skewed, in what direction)? Unimodal or bi/multimodal? For both skewness and number of modes, how do you know? (4 points)
  2. Figure 2.2 above is a histogram of the percentage received by Josh Dobson in 2020 minus the percentage received by Cherie Berry in North Carolina counties in 2016. Counties with a positive value were one where Dobson did better in 2020 than Berry did in 2016, while those with a negative value saw Berry do better 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)

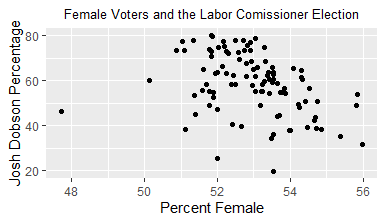
***Figure 3.1***



***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 the vote shift between 2016 and 2020 (Dobson vote percentage in 2020 minus Berry vote percentage in 2016). Is there an association and in what direction? How strong is it? (2 points)
  2. Now, the researcher wants to compare Berry and Dobson’s overall performance (setting elevators aside for the moment). Figure 3.2 looks at the relationship between the percentage of voters who supported Cherie Berry in 2016 and the percentage who supported Josh Dobson in 2020. Is there an association and in what direction? How strong is it? (2 points)
  3. Finally, the researcher is interested in looking whether the percentage of voters in a county who are female relates to the vote percentage for Josh Dobson in 2020. 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 a data point in Figure 3.3 that is distant from all other points. This county has far fewer female voters (as a %) than other counties. What might we suspect this county to be? Should we immediately throw out these cases? Why or why not? (3 points; aside: this happens to be Anson County on the South Carolina border).

**Part II: From Cherie Berry to Mary Berry (24 points)**

In 2016, British host Mary Berry left the [Great British Baking Show](https://www.goodhousekeeping.com/life/entertainment/a34451757/why-did-mary-berry-leave-great-british-baking-show-bake-off/) after it moved from the BBC to Channel 4 and was replaced by new judge Prue Leith. The Great British Baking Show has decided to see whether readers preferred Berry or Leith and would like your input. (15 points)

1. First, the Great British Baking Show would like your input about a variety of possible research designs.
   1. First, the Great British Baking Show comes to you and offers two possible options: either putting a poll on their webpage or asking every viewer to respond. What are each of these approaches called? What are the issues with each? (6 pts.)
   2. Now, the Great British Baking Show offer another approach: asking 1500 viewers 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? Is this approach better than the previous two options? (3 points)
   3. The Great British Baking Show would like your input on other possibilities. Name two other possibilities for how they could get a sample, how the Great British Baking Show would carry out that approach, and a possible issue that could arise with that approach. (6 points)
2. Now, the Great British Baking show would like to carry out an experiment to see whether viewers prefer Berry to Leith. (9 points; let’s imagine this is before COVID-19 when it would be okay to gather a group of people in a room.) Let’s say that Berry is the treatment group and Leith is the control group. (You might actually do this slightly differently if you were to do this experiment in real life, but setting this here for simplicity’s sake.)
   1. How should you assign subjects to the Berry and Leith groups? (2 points)
   2. Voters who have a favorable opinion of the BBC may respond differently than those who don’t like the BBC. How might you go about addressing this when assigning people to groups? What is this approach called? (2 points)
   3. Viewers watching the Leith episode can smell fresh cookies being baked down the hall while those watching the Berry episode cannot. Is this potentially problematic? Why or why not? (2 points)
   4. If the Great British Baking Show finds that readers like the Mary Berry episodes better, 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)

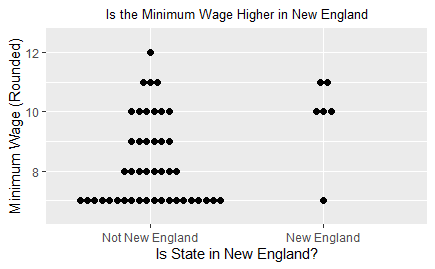
**Part III: Minimum Wage and Poverty (46 points)**

An economic policy scholar would like your help on a project about the minimum wage level and poverty. They are especially interested in looking at relationships between unionization and inequality and whether minimum wages tend to be higher in New England states.

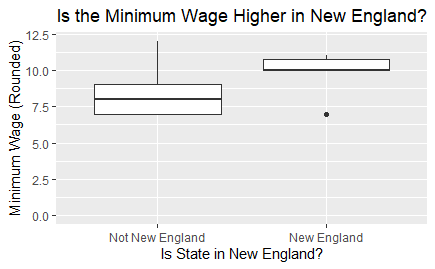
1. The researcher made a scatterplot comparing the minimum wage in 2018 to the poverty rate. (11 points)
   1. Which of these is the explanatory variable and which is the response variable? Is any association you might find necessarily causal in this observational data? (3 points)
   2. There are four 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. New England tends to be more progressive than the rest of the country and the researcher is interested in whether minimum wages are higher there. (36 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 in the box on the left represent? (Note: I’m looking for a conceptual answer here, rather than an exact number.) Why are there only two lines for New England? Comparing the two boxplots, what does this suggest about the minimum wage in New England compared to the rest of the country? (5 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 for the non-New England boxplot, but one dot for the New England box plot? (3 points.)
  2. The researcher would like you to do some calculations of some measures of central tendency of minimum wage data from New England in 2018 that are available in the below table. (I’ve rounded the values here to make the calculations a bit easier.) Please calculate the mean, median, and mode. Please show your work or justify your answer when relevant. (6 points).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| State | CT | MA | ME | NH | RI | VT |
| Minimum Wage (Rounded) | $10 | $11 | $10 | $7 | $10 | $11 |

* 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. There was an outlier below the 1st quartile for New England in the dotplot for part b. Why is there is an outlier below the 1st quartile, but not one above the 3rd quartile? Using the definition for outliers, please calculate what values would be outliers to answer this question. (For the purposes of this problem, please keep the two points used in the calculation of the median in the data in your calculation of 1Q and 3Q.) (5 points.)
  3. New Hampshire tends to be the most conservative New England state. (It was the last New England to vote Republican for president and Republicans have a [state government trifecta](https://ballotpedia.org/State_government_trifectas).) What happens to the mean, median, and mode when you exclude New Hampshire from the cases considered to be in New England? Please show your work. Finally, consider what this shows about the robustness of the mean and median as measures of central tendency. (7 points)