Lab 9: People's Park Survey

In answering the following questions, it will be helpful to consult

1. the email from Chancellor Christ,
2. slide describing the background and methodology, and
3. the original questionnaire.

Part I: Understanding the Context of the Data

Based on your interpretation of these documents, address the following questions.

1. What was the goal(s) of the Chancellor's office in commissioning this survey?
2. To which population does the Chancellor's office wish to generalize their findings?
3. Describe two parameters that the Chancellor's office is trying to estimate using the survey data.
4. What was initial sample size of students selected to take the survey? What was the final sample size that actually did? What was the response rate?
5. Describe one scenario wherein this response rate would lead to results that they could not generalize from the sample to the population. Describe a scenario where this same response rate have no effect on their ability to generalize. Be specific.
6. Consider the type of data collected in question 8, which is measured using the Likert Scale. Review the Wikipedia article on the Likert Scale (particularly the Scoring and Analysis section) to determine: Where does this type of data fall in the Data Taxonomy?
7. Sketch a data frame of what the first 5 rows of the data frame might look like that contains the responses from the first 5 students. Include columns showing what the data might look like that comes out of questions 1, 7, and 8. Note that in the data set, the data values are all translated from words into numbers. Speculate as to how this translation is done.
8. Sketch a plot of what the data might look like that is generated by each of the following survey questions. Note that this should be done without looking at the actual data frame.
    a. Question 9
    b. Question 10
    c. Question 18 and 21 (showing the change from before and after the information)

Part II: Computing on the Data

The following questions deal with a data set called `ppk` which can be loaded by running:

```
ppk <- read_csv("https://www.dropbox.com/s/zamyjzlclortmtu/ppk.csv?dl=1")
```

They represent a subset of questions that were asked in the questionnaire and have had random noise added to them. The results, in aggregate, share similar statistical properties to the raw data, but a given row no longer reflects an individual student's response completely.

9. Print the first few rows with the columns that correspond to the responses to survey questions 1, 7, and 8. Were they as you described them in question 7 here in the lab? If not, describe how they appear.

10. Return to your sketches from question 8 here in the lab. Create those visualizations (or more appropriate analogues) using the questionnaire data. For each, describe the distribution in words. For question 9 you're welcome select just three of the priorities to visualize.

    a. Question 9
    b. Question 10
    c. Question 18 and 21 (showing the change from before and after the information)

11. Create a new column called `support_before` that takes the response data from question 18 and returns `TRUE` for answers of "Very strongly support", "Strongly support", and "Somewhat strongly support" and FALSE otherwise. What proportion of the survey participants in each class (freshman, sophomore, etc) supported the People's Park Project before being presented with the information on the bottom of page 14?

12. What is the mean and median rating of the condition of People's Park (question 15 on the survey)?

13. Create a new column called `change_in_support` that measures the change in support from question 18 to 21. What is the average change in support of the survey participants in each class (freshman, sophomore, etc) for the People's Park Project after reading the information?

14. Construct one addition visualization that captures a variable or relationship between two variables that you are interested in. Describe the structure that you see in the plot.

Part III: Extensions

For the following confidence intervals, please use bootstrap percentile intervals.

15. In bootstrapping, the bootstrapped sample is usually of the same size of the actual sample collected. However, other sample size can also be used. In this questions, create two bootstrap distributions of the median of the rating of the condition, one with 25 reps, the other with 1000. How do they compare? What appears to be the affect of increasing the number of bootstrap samples you use?

16. Generate 9 bootstrap samples of size 1000 for the median rating of the condition of People's Park. Plot those 9 samples using three by three facet plots. Using one of them to construct a 95% confidence interval for the median rating of the condition of People's Park. Interpret the interval in the context of the problem: what does it tell us about the parameter we are trying to estimate?

Hint: For i from 1 to 1000, sample 1000 responses from the original sample about the rating of the condition of People's Park with replacement and compute the median of the 1000 responses. Now you will have 1000 medians, and they should follow a bell-shaped distribution when plotted on histogram.

17. Create a 95% confidence interval for the overall proportion of students that support the People's Park Project without having been exposed to the information on page 14. Interpret the interval in the context of the problem.

18. Create a 95% confidence interval for the overall change in support the Project before and after being exposed to the information on page 14. Does the interval contain 0? What are the implications of that for those working in the Chancellor's Office on the People's Park Project?