Project proposal

Runshi Yang

In this project, I will provide insights about voter opinions about the Liberal Party and party leader, with a focus on voter demographics.

Research question 1:

-Research question

What is a range of plausible (i.e., reasonable) value for the average rating of the Liberal Party rated by eligible voters who are certain or likely to vote?

-Population

The population is all Canadian citizens who are 18 to 99 years old and certain or likely to vote. The parameter of the population is the population's average rating of the Liberal Party.

-Data visualization

I want to use **histogram** to do the data visualization. I will set the x-axis to be the rating of the Liberal Party rated by eligible voter who is certain or likely to vote, then set y-axis to be the number of voters who give a corresponding rating. I choose to use histogram because the rating is a numerical variable and the plot can show you the distribution of the ratings intuitively.

-Methods

First, I need to do the **data wrangling** to extract all observations we are interest in. I will use the filter( ) function to pick out the observations that variable 'citizenship' has the value 'Canadian citizen', and variable 'v_likely' has the value 'Certain to vote' or 'Likely to vote'.

Then I will use the **bootstrap method** to come up with a range of the average rating that would be plausible for the true parameter value. I will draw many bootstrap samples from the data I filtered, then calculate the average value of the variable 'party_rating_23' for each bootstrap sample. After that I can get the bootstrap confidence interval by figuring out the percentiles.

Research question 2:

-Research question

Is the average rating of Justin Trudeau similar between voters aged 18 to 60 and voters aged 60 to 99?

-Population

The populations are all Canada citizens who are 18 to 60 years old and all Canada citizens who are 60 to 99 years old. The parameter of the populations is the population's average rating of Justin Trudeau.

-Data visualization

I want to use **barplots** to do the data visualization. I will set the x-axis to be people aged 18 to 60 and people aged 60 to 99, then set the y-axis to be the average rating of Justin Trudeau. I choose barplots because the age group is a categorical variable and we can compare the difference of two groups directly from the plot.

-Methods

First, I need to do the **data wrangling** to create new variables and extract all observations we are interest in. I will use mutate( ) to create a variable called 'age_group', and set its value to 'young' for the observations with between 18 and 60 and 'old' for others using the case_when( ) function.

Then I will use **hypothesis test** to compare the average rating of Justin Trudeau between people aged 18 to 60 and people aged 60 to 99. My **null hypothesis** $\boldsymbol{H_0}$: there is no difference in the average rating of Justin Trudeau between people aged 18 to 60 and people aged 60 to 99. And **alternative hypothesis** $\boldsymbol{H_1}$: the average rating of Justin Trudeau is different between people aged 18 to 60 and people aged 60 to 99. To get the test statistic, I will calculate the difference between the two averages using the variable 'lead_rating_25'. After that I will simulate under the null hypothesis to get the distribution of the two averages' difference. And finally, I can evaluate the evidence against the null hypothesis and make a conclusion based on the p-value.

Research question 3:

-Research question

Is the interest value in this federal election similar between men and women eligible voters who rates the Liberal Party lower than 50 in Canada?

-Population

The populations are all Canadian men citizens who are 18 to 99 years old and rates the Liberal Party less than 50 and all Canadian women citizens who are 18 to 99 years old and rates the Liberal Party less than 50. The parameter of the population is the population's average interest value in this federal election.

-Data visualization

I want to use **boxplots** to do the data visualization. I will set the x-axis to be men and women and set the y-axis to be the interest value of the voters. I choose boxplots because it can show the distribution of interest value of both men and women on the same plot, which is convenient to compare. And we can see median values of interest of two groups of people directly from the plots.

-Methods

First, I need to do the data wrangling to extract all observations we are interest in. I will
use the filter( ) function to pick out the observations that variable 'citizenship' has the value 'Canadian citizen', variable 'gender' has the value 'man' or 'women', variable 'party_rating_23' has a value lower than 50.

Then I will use hypothesis test to compare the means between men and women voters. My **null hypothesis** will be there is no difference in interest value in this federal election between men and women eligible voters who rates the Liberal Party lower than 50. And **alternative hypothesis** is the average interest values in this federal election are different between men and women eligible voters who rates the Liberal Party lower than 50. To get the test statistic, I will calculate the difference between these two averages using the variable 'interest_elxn_1'. After that I will simulate under the null hypothesis to get the distribution of the two averages' difference. And finally, I can evaluate the evidence against the null hypothesis and make a conclusion based on the p-value.