Voter Insights For The Liberal Party

With a Focus On Demographics

Project group 37

December 7, 2020

Introduction

Goal

- Give the members of the Liberal Party insights about voter opinions about the party and its leader.

Our questions will focus on voter demographics so we can better understand the opinions of different subgroups within eligible voters.

1. Given that a voter has the Liberal Party as their first choice, is the probability that he/she is female 0.586?
2. Is there a difference in the average rating of the Liberal Party between those who are 18-40 years old and those who are 40+ years old?
3. What is the range of plausible values for the average rating of Trudeau's government among eligible voters with at least a high school education?

Data Summary

Overall Summary

The dataset that we will use for all analyses and tests is taken from the Canada Election survey 2019

Relevant Variables

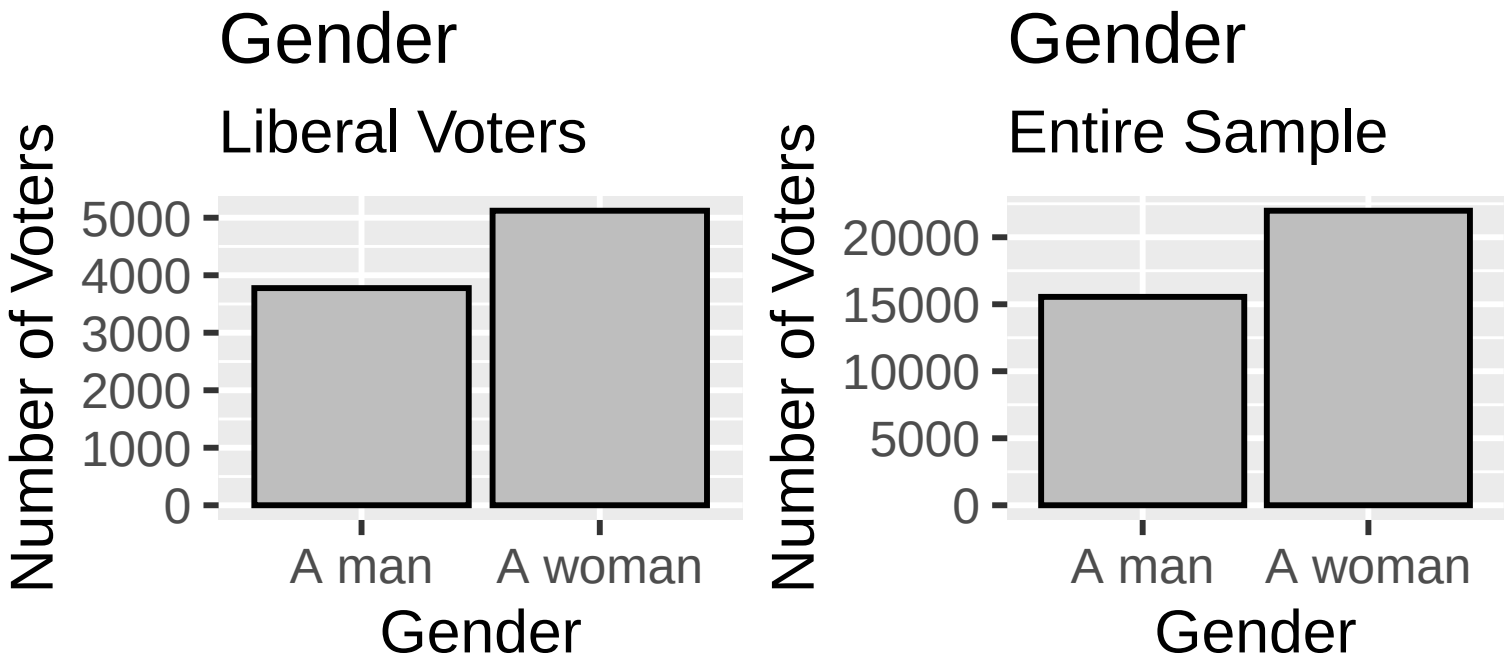
- votechoice: A person's first vote choice
- lead_rating_23: A person's rating of Justin Trudeau
- party_rating_23: A person's rating of the Liberal Party
- Demographic variables: gender, education, age

How will we process variables?

- Filter out missing values, select relevant variables to the test
- More details on the individual tests.

Research Question 1

Q1: Given that a voter has the Liberal Party as their first choice, is the probability that he/she is female 58.6%?



- Removed those who do not identify as male or female, and also those who did not answer the question for simplicity.
- The proportion of males and females among those with the Liberal Party as their first vote choice is similar to the proportion of males and females in the entire sample.

In our sample of 8898 males and females with the Liberal Party as their first vote choice, 57.6% are women. 58.6% of all voters in the dataset are women.

We cannot conclude anything with just this test statistic. We will run a 1-variable hypothesis test.

The population is all people eligible to vote in Canada who have the Liberal party as their first choice, the parameter being measured is the probability that a given voter is a woman.

Hypotheses:

- Our null hypothesis is that the probability is 0.586, suggesting that chance alone is altering the statistic we obtained.
- Our alternative hypothesis is the alternative, which is that the probability is not 0.586

Question 1: Procedure

Simulations

- We simulated 1000 samples under the null hypothesis.
  - Each sample contains 8898 observations, each with a 0.586 chance of being female.
- We then calculate the proportion of women in each sample.

Evaluation

- We will calculate the proportion of simulated statistics that were more extreme than our test statistic. This number is called the p-value
- We will evaluate all hypothesis tests at the 0.05 significance level.
  - If the p-value we calculate is less than 0.05, we will reject the null hypothesis.

Question 1: Conclusion

Why do we need a p-value?

- We want to collect evidence against our null hypothesis.
  - The simulated samples roughly show what *could* happen due to chance alone.
  - If our test statistic was very unlikely to occur due to chance alone, we have evidence that something other than chance is influencing the results.

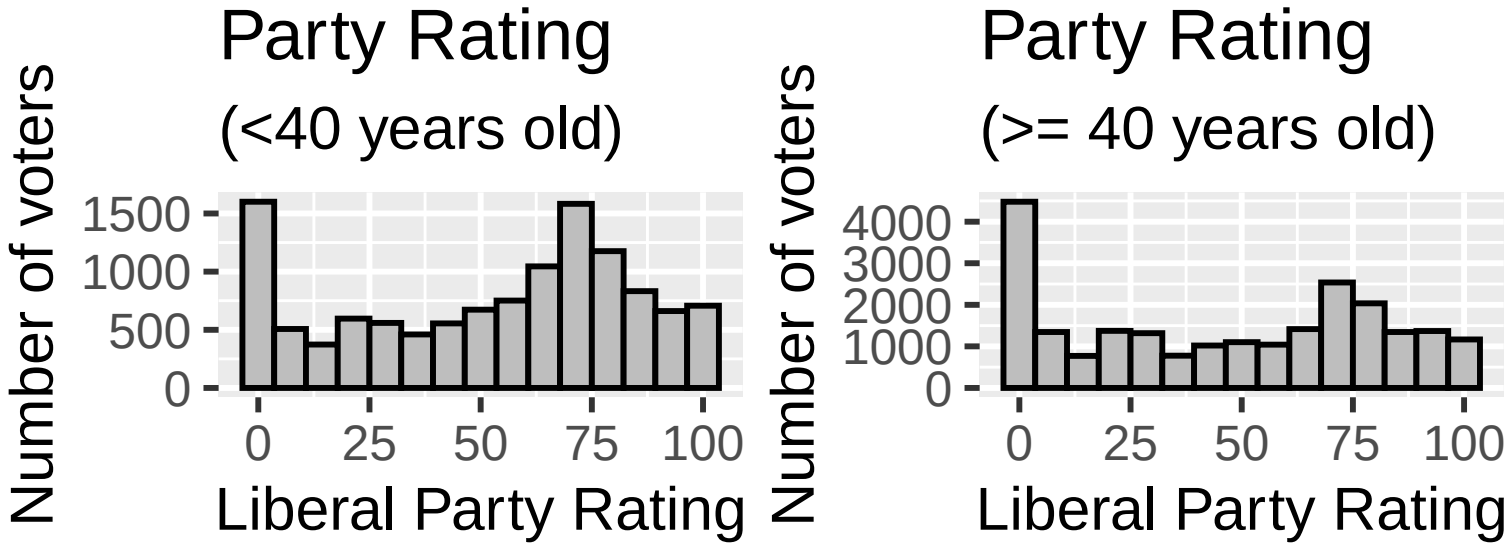
Calculating a p-value

The p-value from this test is 0.076. Basically, 7.6% of our simulated statistics were more extreme than the test statistic.

- *We conclude that the result of this test was not unusual to the point that we would claim some unknown factor is influencing a female's likelihood of voting for the Liberal Party, but the test statistic is certainly unusual.*

Question 2

Is there a difference in the average rating of the Liberal Party between those who are 18-40 years old and those who are 40+ years old?



- Created a new variable "age_group," that separated voters into age groups.
- There is a larger "bump" on the graph for the younger group centered at 75, compared to the older group.

- The older group's average party rating is 6.11 lower than younger group.
- We will establish the null hypothesis to run a hypothesis test.

The population is all eligible voters. Our null hypothesis for this test is that the average Liberal Party rating of those below 40 years old and those above 40 years old will be the same.

Procedure

- We simulate samples under the null hypothesis, where the difference in means is 0. We achieve this by redistributing the age group variable randomly to observed rating values.
  - We then calculate the statistic(difference in means) for each sample.
- Get a p-value by calculating how many samples had a difference in means more extreme than our test statistic.

Question 2: Results

Calculating the P-value

- Our p-value is 0, meaning that none of the simulated samples had a difference in mean party rating that was more extreme than our test statistic.

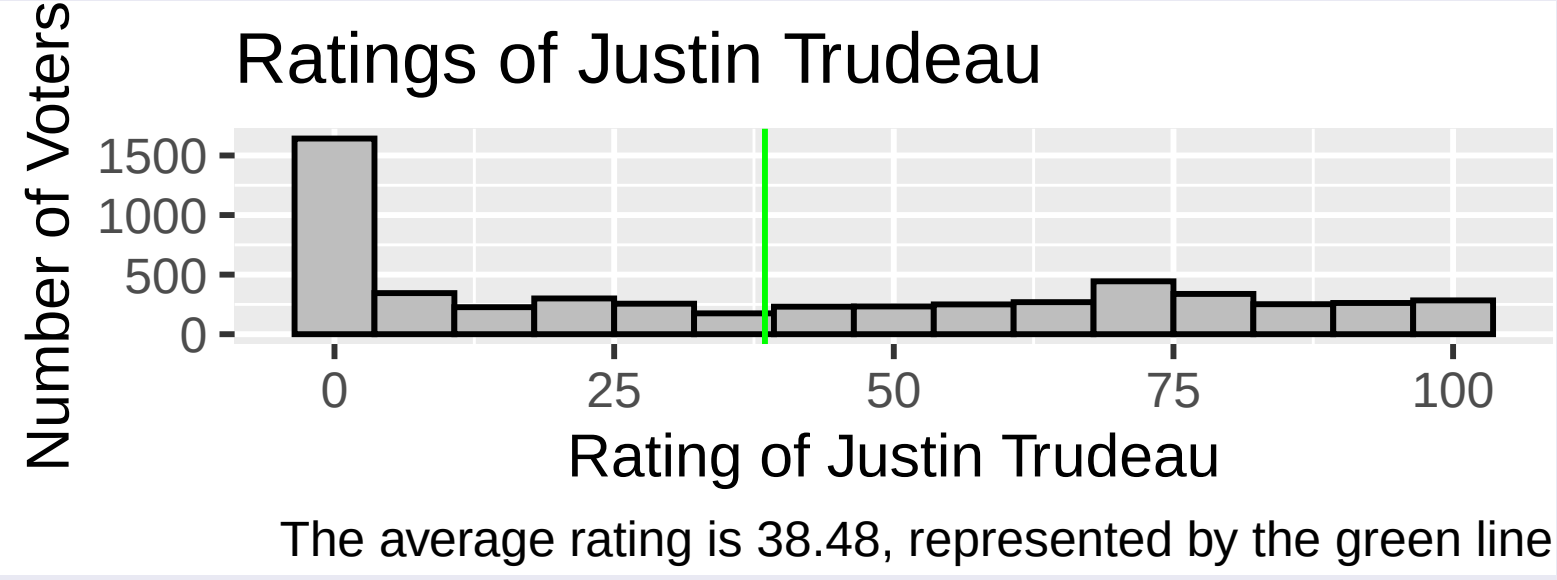
Evaluating the p-value

- We have very strong evidence against the null hypothesis, and would reject it at the standard 0.05 significance level that we previously set.
- We conclude that there is a difference in the mean rating of the Liberal Party between those who are 18-40 years old and those who are over 40 years old.

Question 3

Q3: What is the range of plausible values for the average rating of Trudeau's government among people eligible to vote in Canada who also have at least a high school education?

Visualization



Question 3: Procedure

- The average rating of Justin Trudeau based on a sample of 5508 voters is 38.48. Many people gave a rating of 0.
- We want a range of plausible values for the rating of Justin Trudeau among all eligible voters with a high school education. We will use the Bootstrap Method.

Boostrap Method

- We create simulated samples by reusing observations from the original sample. This will generate a variety of resamples similar to the original.
- For each resample, we can calculate the mean rating of Justin Trudeau, giving us a distribution of values.
- Using this distribution, we will generate a range of plausible values, called a Confidence Interval.

Question 3: Conclusion

The 95% confidence interval for the Rating of Justin Trudeau out of a score of 100 is between 37.53 and 39.44.

What does this mean?

- A confidence interval is a range of plausible values that contains 95% of the statistics generated by bootstrapping.
- We are 95% confident that the interval contains the actual average rating of Trudeau among all people who have at least a high school education.
- Further analysis is necessary to create inferences about this range of plausible values. We cannot say that this interval is good or bad without comparing with other groups.

Limitations

Limitations of the data

- The data from the Canadian Election survey may not be representative of the entire population(all eligible voters)
  - For example, the survey may attract people with stronger opinions about the government, a possible explanation for the large amount of zeroes under the party and leader rating variables

Limitations of the methods

- Potential errors in our conclusions
  - For example, we failed to reject the null hypothesis in Question 1, but the p-value was close to the significance level of 0.05. There is a possibility that we made a Type 2 error, where we fail to reject a null hypothesis that is false.

Overall Conclusions

Question 1

- We conclude that chance alone is influencing the gender of a person who's first vote choice is the Liberal party. The visualization shows that there is not a disproportionate amount of males and females who plan to vote the Liberal Party

Question 2

- We conclude that there is a difference in the average rating of the Liberal Party between those below 40 years of age and those above 40 years. In the sample data, the younger group had a higher average rating than the older group.
- Performing a similar test with subgroups within the younger group may help us understand why the younger group yielded higher ratings

Overall conclusions(continued)

Question 3

- We are 95% confident that the average rating of Trudeau's government among eligible voters with a high school education is between 37.46 and 39.46.
- Comparing to other parties and demographics would allow us to understand how to interpret this value, because we cannot judge whether it is good or bad with this information.

Further exploration and applications

- Investigate the cause of these results by analyzing different demographics, or subgroups within demographics used in this presentation.
- Perform the same analyses for other parties to better understand them.