STA130 Final Project

Analysing Voter Demographics - Liberal Party

Chetanya Saxena (1007320533), Maryam Ansari (1006917204), Raazia Hashim (1006819454) and Suchi Aidasani (1006753229) - Group 132

December 7, 2020

Introduction

The questions we chose to cover in our final project provide insight to the Liberal Party about their current standing in voter opinions. For this, we use past data and manipulate it to make predictions. The findings of this project will contain information about the gender split among the past votes in favor of Liberal Party as well as if Liberal Party is the first choice of voters. For the research questions entailed in this presentation, we are using the data from the 2019 Online Canadian Election survey. We have treated the data we have of 25850 (after removing missing values) as a sample and the population is all the people politically participating in voting; all Canadians over the age of 18.

Data Summary

Each of our research questions involve different variables from the Election Survey results. To familiarize the audience with them, their descriptions are attached as follows: - *votechoice*: Which party is your first choice to vote for? - *gender*: Which gender does the voter belong to? (e.g. man, woman or other) - *bornin_canada*: Were you born in Canada?

The methods for data wrangling that we will be using are filtering, grouping and mutating variables.

Research Question 1

Introduction

What percentage of votes should the Liberal party expect from people born in Canada during the election at this point in time?

The problem explored in this question is the proportion of people who would vote for the "Liberal Party" as their first choice as opposed to the other parties collectively. I have used our sample to try and predict a confidence interval for the Liberal party so that they know what percentage of vote they should be expecting.

Statistical Methods

The statistical method we will use to approach our question is bootstrapping. Bootstrapping is a method used to estimate the sampling distribution of the population of eligible voters in Canada. We draw many bootstrap samples of the same size (n = 22146) from the sample that we have (in this case 1000 bootstrap samples) with replacement which allows our bootstrap sample to have duplicate values. Next, for each bootstrap sample, we filtered the data by vote choice of Liberal and by people born in canada. We are not creating new data, rather we are exploring variability from the original sample to create a range of plausible values for the difference in proportions of votes.

Visualization

We will visualize the proportion of people who would vote for the liberal party as their first choice by creating a histogram which will allow us to observe the shape and distribution of the proportions.This distribution shows that the values range from around 0.295 to 0.310, and the center is around 0.3025. Some extremities can be observed but this provides an overview of the number of people who would vote for the liberal party.

95% confidence level

```
##       2.5%      97.5%
## 0.3295821 0.3413709
```

Results

If we repeated this process many times, 95% of those confidence intervals would include the true proportion of people born in Canada who would vote for the Liberal party. It is always good to know where you stand and identify what needs to be improved, therefore this data is useful for the liberal party. This provides a good starting point when deciding the approach to this election in that the results of this data provides a reason for the liberal party to modify their policies so that they can appeal to immigrants and greater demographics so that they could increase the proportions of people who would vote for the liberal party as their first choice.

Conclusion

Moreover, since this is a proportion in comparison to all the parties collectively, this data tells the Liberal party that they have an unspoken lead in the election because approximately 30% is a large percentage and therefore they should strengthen their weak policies and further strengthen their stronger policies and consider different strategies to market themselves more to the general voting population.

Research Question 2

Introduction

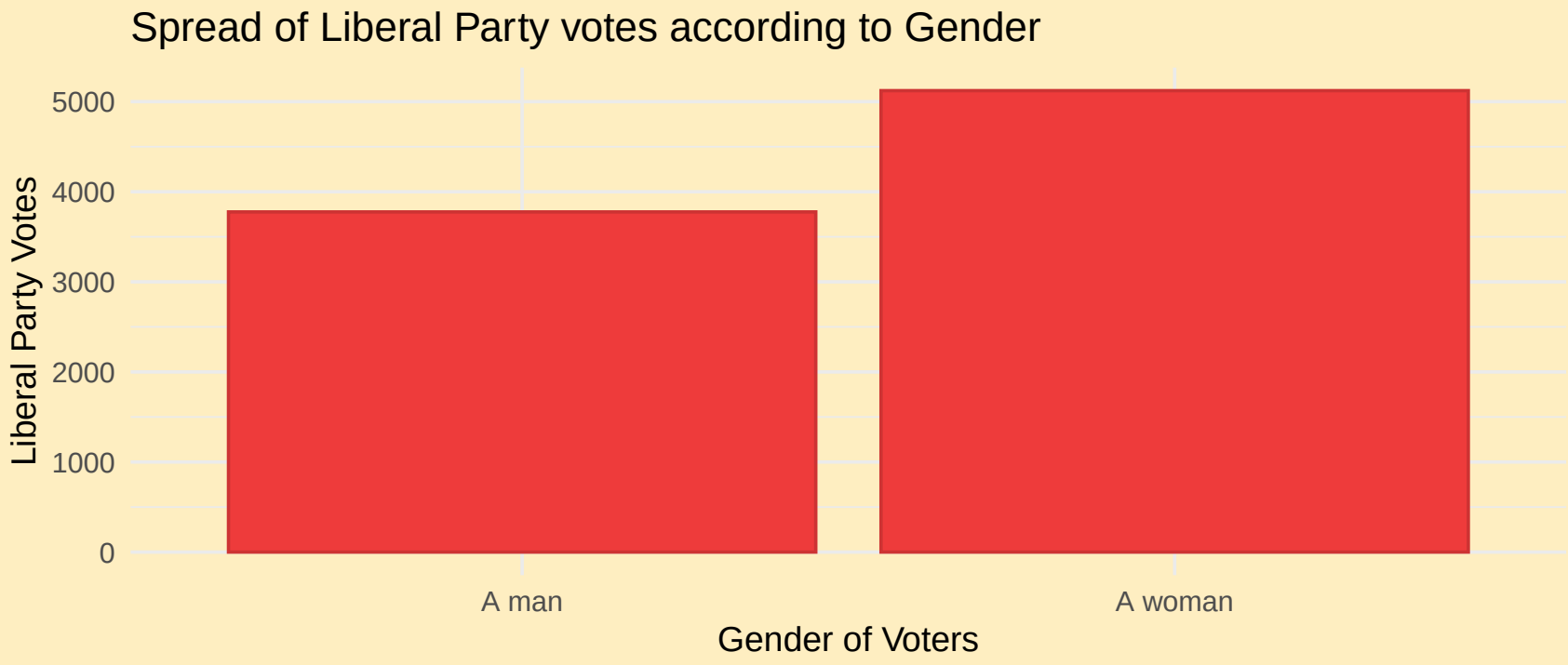
Is the proportion of male people who voted for the Liberal Party 50%?

This research question can help determine the gender split among the votes, and whether a certain gender preferred the Liberal Party.

Our population for this specific research question would be the people that filled out the survey, voted for the Liberal Party and are within the ages of 18 and 99 inclusive.

Visualization

The data, for all the eligible votes whose first choice vote is for the Liberal Party, is seen on the Bar plot below and is divided by gender.



From this data we can already conclude that more women had Liberal Party as their first choice for their vote, compared to men.

Visualization (contd.)

This bar plot was created in steps. First the data was filtered by removing missing (or NA) values and then a new data set called 'liberalvotes' was created which stores all the data for the individuals that had listed the Liberal Party as their first choice of vote. The reason a bar plot was chosen is because the variable gender has 2 different categorical levels which makes this plot the most suitable. Furthermore, bar plots are more easily interpreted by those who lack statistical knowledge.

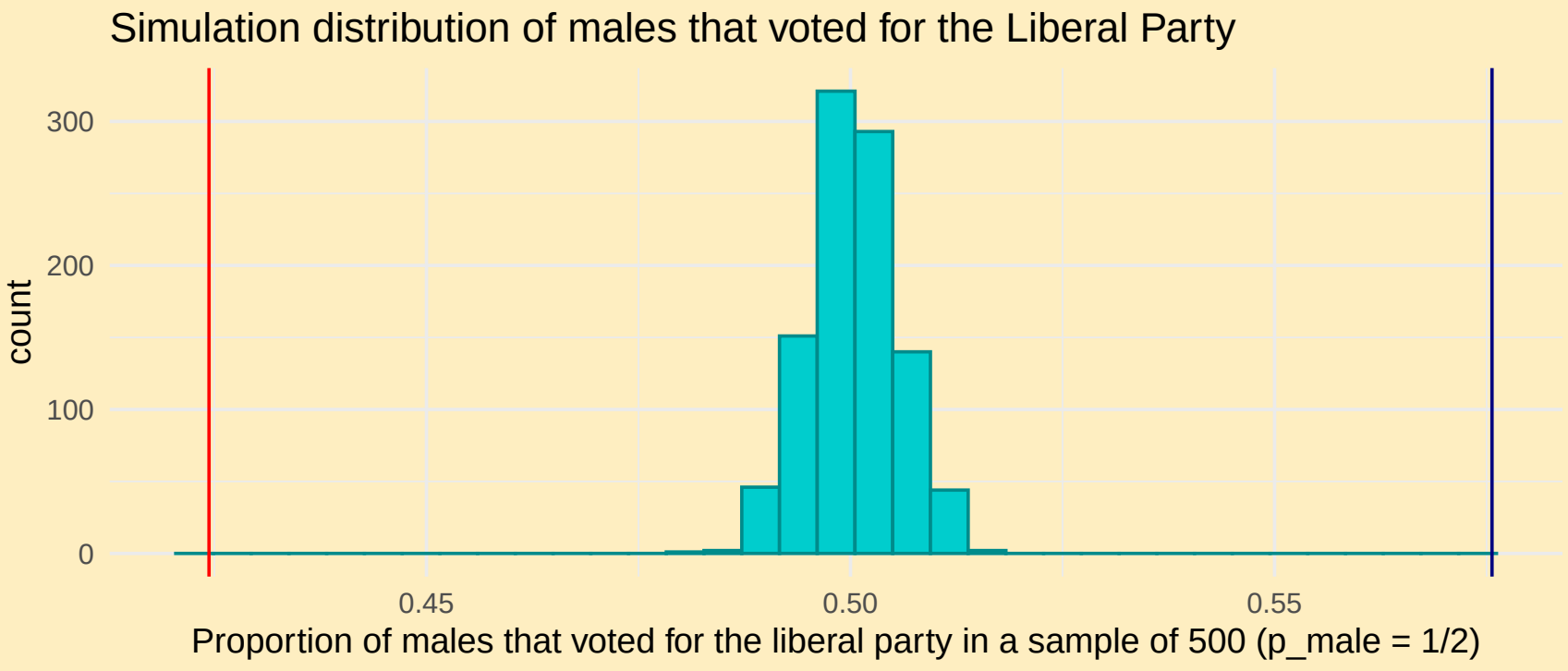
Statistical Methods

I can test our original research question by carrying out a one-sample hypothesis test. Our hypotheses for the test are listed as follows with H_0 being the null hypothesis and H_1 being the alternative hypothesis, with p being the proportion of men.

$$H_0 : p_{male} = 0.5$$

$$H_1 : p_{male} \neq 0.5$$

Hypothesis Test Stimulation



Results

```
##      2.5%     97.5%
## 0.4896578 0.5106766
```

Above is the result from a 95% confidence interval (it is used to calculate how confident we are with our data). The results from this calculation state that "We are 95% confident that between 49% and 51.1% of people that voted for the Liberal party are male". A narrow confidence interval means that there is less variability in our data and may explain why there is a large gap between our vertical lines and histogram.

```
## [1] 0
```

Our histogram visualization is symmetrical, centered at 0.5 (mean proportion) and values range between 0.49 and 0.52. We use our test statistic and original proportion to find our p-value (the probability of obtaining test results in the least extreme scenario, under the null hypothesis). In this case, the p-value is 0.

Conclusion

Since our p-value 0, we have very strong evidence against the null hypothesis that the prevalence of males among those whose first choice vote was the Liberal Party, is 0.5.

Introduction

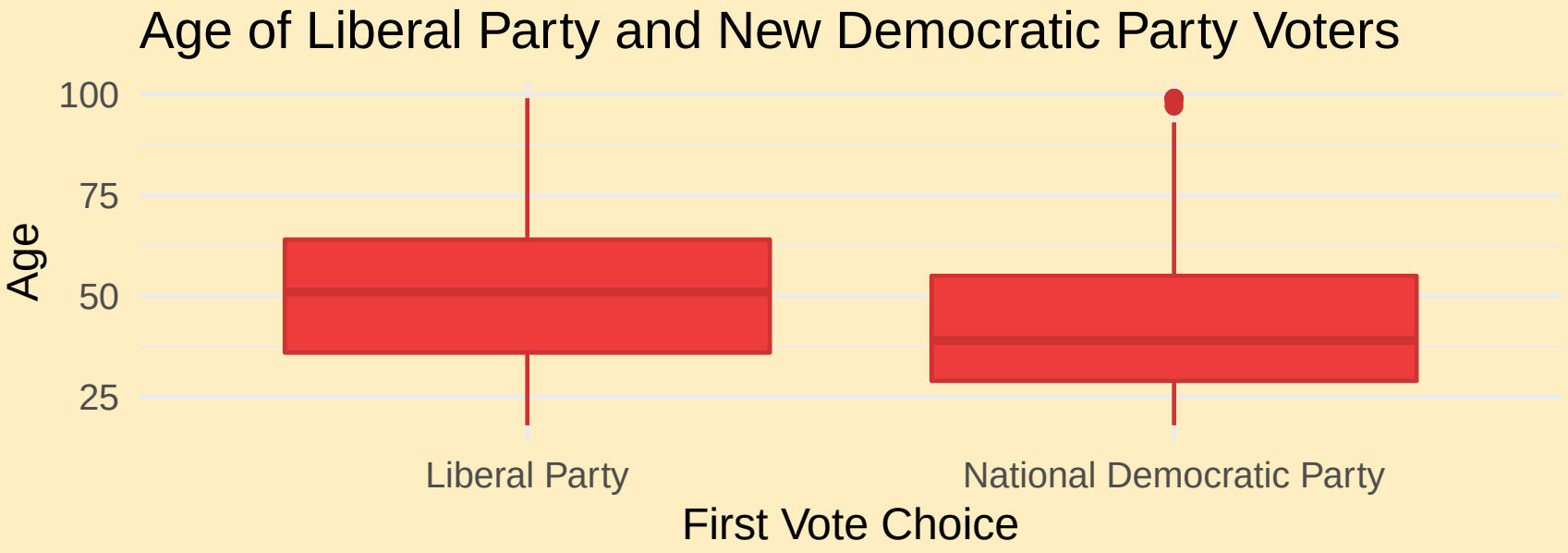
What is the range of plausible values of the difference in median age of Canadians who would vote for the Liberal party verses the National Democratic Party as their first choice?

We will determine whether a Canadian who would vote for the Liberal Party, is on average younger or older than a Canadian who would vote for the National Democratic Party. This is an important question to answer, to determine if voters of the two parties belong to a similar age demographic.

Data Summary

To explore this question, we shall look at Canadians 18 years or older, that is who are eligible to vote in the election and who report that they will vote for the Liberal Party or the National Democratic Party as their first choice, and we will filter our data to only contain this subset of observations.

Visualization



We will visualize the age of Liberal and the NDP voters by creating two box plots, which will give us a way to compare the center and distribution of ages for the two groups. The center, that is the median, which is the middle values is seen by the position of the line in the center of the boxes. We can see that median age of NDP voters is much lower than the median of Liberal voters. The distribution of NDP voters is very skewed to the right (that is there is a longer right tail), meaning that most voters tend to be younger. The distribution of Liberal voters is also slightly right skewed, but not as much as the other group. Finally, we see there are a few outliers in the NPD voter ages, where some voters are a bit older than the rest in the group.

Statistical Method

The statistical method we will be using to answer our question is bootstrapping. Bootstrapping is a method used to estimate the sampling distribution of the population of eligible voters in Canada. We draw many bootstrap samples of the same size (n = 200) from the sample that we have (in this case 5000 bootstrap samples) with replacement. Meaning that our bootstrap samples may have duplicate values in them. Next, for each bootstrap sample, we filtered the data by vote choice, calculated the median age and found the difference between the median age of Liberal voters and the median age of NDP voters.

Doing this, we are not creating new data, rather we are exploring variability from the original sample to create a range of plausible values for the difference in median age. This is a confidence interval, we found it by taking the middle 80% of the bootstrap distribution (wider and narrower intervals could have been taken). This confidence interval tells us that if we repeated this procedure many times, 80% of those times, the true difference in median age would fall inside the confidence interval.

```
##  10%  90%
##    6   16
```

Results

The confidence interval we found through our bootstrap sampling for the difference in median age for Liberal and NDP voters in between 6 and 16. Since the interval is positive, and recall that we subtracted the median age of Liberal voters from the median age of NDP voters, we can conclude that on average, NDP voters tend to be younger than Liberal voters.

Conclusions

Since this is the case, the Liberal Party should be focusing their campaigning efforts and new party policies towards a younger demographic of Canadians, as this is something that the National Demographic Party is succeeding in. The NDP is able to attract a younger voter base through the fresh ideas they propose and the Liberal Party should be taking a page out of their book. Finally, some limitations in answering this question using the bootstrapping method is that the sample data that we begin with may be biased and not fully representative of all Canadian voters and since bootstrapping only reuses our sample data, the interval we came up with may be biased.

Limitations

Some limitations to this data include the sample data may not be completely representative of the population due to bias. In addition to this, since samples are randomly drawn, this can lead to some uncertainty. Another limitation that cannot be removed is that some of the results might be biased due to confounding in variables.

Summary

The findings contained in this project serve to advise the Liberal Party on how their campaign can be improved by targeting the right audiences. They are as follows:

- The Liberal Party only has 30% of the population sure about voting for them and they need to improve their campaign generally.
- In general, the proportion of men to women in Liberal voters is not 50%. The Liberal Party needs to target their policies towards men.
- They also need to target the upcoming new voters more, maybe by easier student loans or similar policies to earn the favor of this age group.