ADD IN MEANINGFUL TITLE HERE

STA304 - Winter 2025 - Assignment 1

Ahnaf Alam - 1006805076

January 23, 2025

1 Introduction

Prior to elections, polling is a common way to understand how the electorate if feeling leading up to an election. It allows to estimate and predict the results before the election date. Elections surveys revolve around question of how people are feeling about each of the candidates/parties and their voting intentions in the upcoming election. The most common way collecting such data is through either phone questionnaire or online surveys (web). The target population for pollsters is the entire Canadian electorate who are residing in Canada and is eligible to vote. The frame population a caller list for web or phone which contains phone number for everyone. The sample population are those who chose to do the survey, either through phone or web and therefore, is in the data. The phone purpose of this exercise to infer how Canadians may vote in the upcoming election.

The current surveys collects data of seven variables across Canada. It asks a persons their gender, and age, along with their current province, highest level of education attainment, interest in the upcoming election, likelihood of voting and who they are most likely to vote for. I focus on two variable in this report: age and likelihood of voting. Historically, there has been discrepancy between voting intention and the actual voter turnout. For example, in the 2015 elections, the overall voter turnout was 66.1% despite voting intention being high. While data on actual voter turnout can only be found after the election, this report want to visualize the electorate voting intention and compare that with actual turnour rate to see if there any discrepancy between the two.

The following report will comment

2 Data

The data comes from 2019 Canadian Election Study. The surveys were either conducted through phone or through an online survey. Both versions of the surveys were comparable

and the questions focused on election-related issues, like vote intention, political engagement, and partisanship, among other variables. The surveys also contained demography question, pertaning to respondents age, province, education attainment, and gender.

The following report focuses on two variables: respondents' age and voter turnout. Particularly, we are interested in knowing whether the distribution of respondents differs between the two sources, phone or web. We do this to understand whether the patterns for those

In the phone survey, the variable name for age category was `age` and certainty over voting was called `q10`. The `age` variable was numerical number, corresponding to respondents age. The question likelihood of voting took values 1 though 5, representing categories "Certain", "Likely, Unlikely", "Certain not to vote", and "Already voted in advanced polling", respectively. Similarly, for the web data, `cps19_yob` represented respondants age and likelyhood of voting variable is called `cps19_v_likely`. It takes values from from 1 through 7, representing categories, "Certain", "Likely Unlikely", "Certain to not vote", "I am not eligible to vote", "Don't know/prefer not to answer", and "I voted in an advance poll". For both phone and web data, we create variable `age_group`, which divides the repsondents into different age group, and then combine phone data and web data together to create one dataframe, which we use to visualize the demography variable. To estimate the likelihood of voting, we create a binary variable `voted_or_not`. We grouped respondents together who answered "Certain", "Likely" and "I already voted in an advanced poll" in the survey. The variable took 1, indicating these respondents have either voted in the election or is more likely to vote. Similarly, we grouped respondents together who answered "Unlikely" and "Certain not to vote" together, assigning 0 to the variable. Additionally, we got rid of observations who answered "I am not eligible to vote" and "Don't know/prefer not to answer" as their answer is inpplicable to the question being asked. The complete data and code can be found here.

3 Demographic Variables

Table 1 shows number of respondents between for each of the age group, separated by survey method. The graph highlights few key points. Firstly, for both phone data and web data, most of the respondents are between the ages of 30 and 49.In the case that it signals that that respondents at this age group ae more likely to respond to survey, demographers could look for databases that separates respondents by age group, and create a frame population using only those in that age group. Sampling from this database will increase the popability of survey response and with a large enough sample population, the resulting analysis will allow deeper understanding of how the true population of this age group may vote in the election.

Similarly, we see that number of respondents is lower for 18-30 age group. This could imply coupld things. Firstly, it could be that people in that age group are just less likely to respond, irrespective whether they are reached through phone or web. Moreover, frame population is not represetive of the true population as database they are using to survey people, includes inadequete amount of data on that age group, which means sapling population is also lower

Table 1: A comparison of the number of people interviewed by phone or web across different age groups

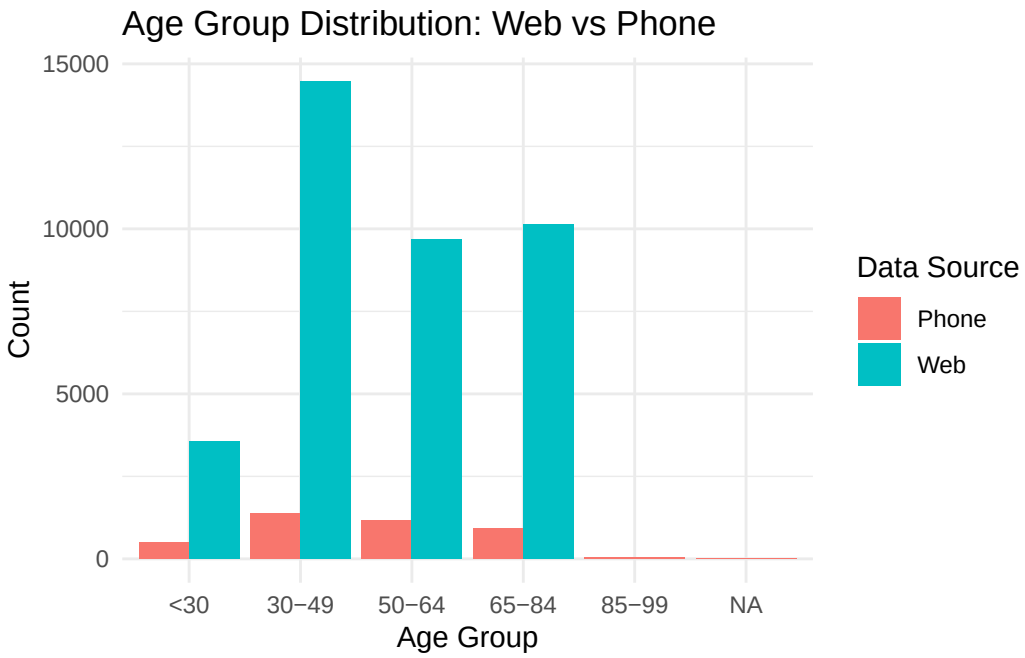

Table 2: 95% Confidence Interval showing who likely have voted or is certain to vote in the 2019 elections

Outcome Variable	Proportions	95% Confidence Interval
Phone Survey	0.951	(0.944, 0.958)
Web Survey	0.913	(0.910, 0.916)

than other age demography and any inference related to that age group must be evaluated further as the current sample data may not be good presentation of the true population.

4 Outcome of Interest

Clearly state what your outcome variable is, and give a brief explanation of why you chose it. You will analyze this outcome in both datasets. For each survey (phone and web) the formula used for the confidence interval should also be presented and referenced [2]:

$$\hat{p} \pm z \cdot SE$$

5 Comparative Analysis

Here you will write a few paragraphs with a general reflection commenting on: demographic differences, biases/errors, and implications for analysis.

6 Generative AI Statement

Here is where you can explain your usage of Generative AI tool(s). Be sure to reference it.

Alternatively, if you did not use Generative AI, please include a brief statement outlining your workflow for completing this assignment.

7 Bibliography