The age-old debate: analyzing the underepresentation of younger voter in surveys*

STA304 - Winter 2025 - Assignment 1

Ahnaf Alam - 1006805076

January 28, 2025

1 Introduction

Election surveys are essential tools for understanding public opinion leading up to an election. They provide valuable insights into voter preferences, opinions on candidates, and key issues, allowing surveyors to estimate and predict election outcomes before voting takes place. Surveys are commonly conducted via phone or online questionnaires. In this report, I will compare and analyze similarities between phone and web survey results, using the 2019 Canadian election data.

In this report, I focus on two variables: age and likelihood of voting. Examining age allows me to compare the number of respondents across different age groups in phone and web survey data. This helps identify discrepancies in responses by age group between the two methods. Older individuals are more likely to use landlines, while younger people typically have better access to the internet. Any such discrepancies could suggest that one survey method is more effective for specific demographics, enabling better sampling strategies based on age. I also analyze the proportion of respondents who say they are likely to vote. In Canada, voter turnout is historically low, with only 66% participation in the 2015 general election. I aim to determine whether advanced surveys can predict voter turnout. By comparing turnout rates from phone and web surveys with actual turnout, I assess which method aligns more closely with real-world data and explore potential reasons for any differences.

In Section 3, I discuss the data source and outline the cleaning process to prepare it for analysis. In Section 3, I present a bar plot comparing the number of respondents in phone and web survey data across different age groups, followed by a discussion of why some age groups have fewer respondents. Next, Section 4 examines potential reasons for the high proportion

[*] For replication, the data and code can be found at https://github.com/AhnafAlam1/analyzing-2019-Canadian-elections-data

of respondents reporting a likelihood to vote. In Section 5, I address response bias in phone data and coverage bias in web data. The report concludes with a discussion on the role of generative AI in this analysis.

2 Data

The report uses phone data (Stephenson et al. 2020b) and web data (Stephenson et al. 2020a) from the 2019 Canadian Federal Election. Despite different collection methods, both surveys included comparable questions on election-related topics such as vote intention, political engagement, and partisanship, as well as demographic questions on age, province, education level, and gender.

This analysis focuses on two variables: respondents' age and likely voter turnout. For age, I examine whether there are differences in age-group representation between phone and web surveys. Identifying such discrepancies can help surveyors address underrepresentation of certain demographics in future data collection. For likely voter turnout, I analyze whether survey results align with historical turnout data. Discrepancies between reported and actual turnout can provide policymakers with insights into why some individuals who express intent to vote ultimately do not, enabling the design of policies to address this gap.

In the phone survey, the variable `age` represented respondents' age, and `q10` represented likelihood of voting, with values ranging from 1 to 5 ("Certain" to "Already voted in advance"). In the web survey, `cps19_yob` represented age, while `cps19_v_likely` measured likelihood of voting, with values from 1 to 7 ("Certain" to "I voted in an advance poll"). I created an `age_group` variable for both datasets, categorizing respondents into 18-34, 35-54, and 55+ age groups. Further, phone and web data were merged into a single dataset to analyze trends. To estimate voting likelihood, a binary variable was created, named `voted_or_not`. Respondents who selected "Certain," "Likely," or "I already voted in an advanced poll" were grouped as 1 (likely to vote). Those who answered "Unlikely" or "Certain not to vote" were grouped as 0 (unlikely to vote). Responses like "I am not eligible to vote" and "Don't know/prefer not to answer" were excluded as irrelevant.The complete data and code can be found here.

3 Demographic Variables

Figure 1 shows the number of respondents in each age group, separated by survey method. Approximately 45% of respondents in both the phone and web surveys were aged 55 or older, while around 20% were younger individuals. Both surveys effectively captured older populations but struggled to reach younger respondents.

In the phone survey, the target population included all adult Canadians with access to a landline or cell phone. Constructing the sampling frame was challenging, so Stephenson et al.

(2020b) used external survey agencies. Advanis provided wireless phone numbers, and ASDE provided landline numbers. Only in-service numbers were included in the frame. Wireless numbers were randomly selected, while for landline samples, if multiple eligible respondents were present, the individual with the next birthday was chosen to ensure randomization.

For the web survey, Qualtrics, an external platform, created the sampling frame. Respondents had previously signed up for surveys, and Qualtrics drew from multiple panels to match the Canadian population's distribution. The target population included adult Canadians, including permanent residents, with internet access. Participants were randomly selected from this frame.

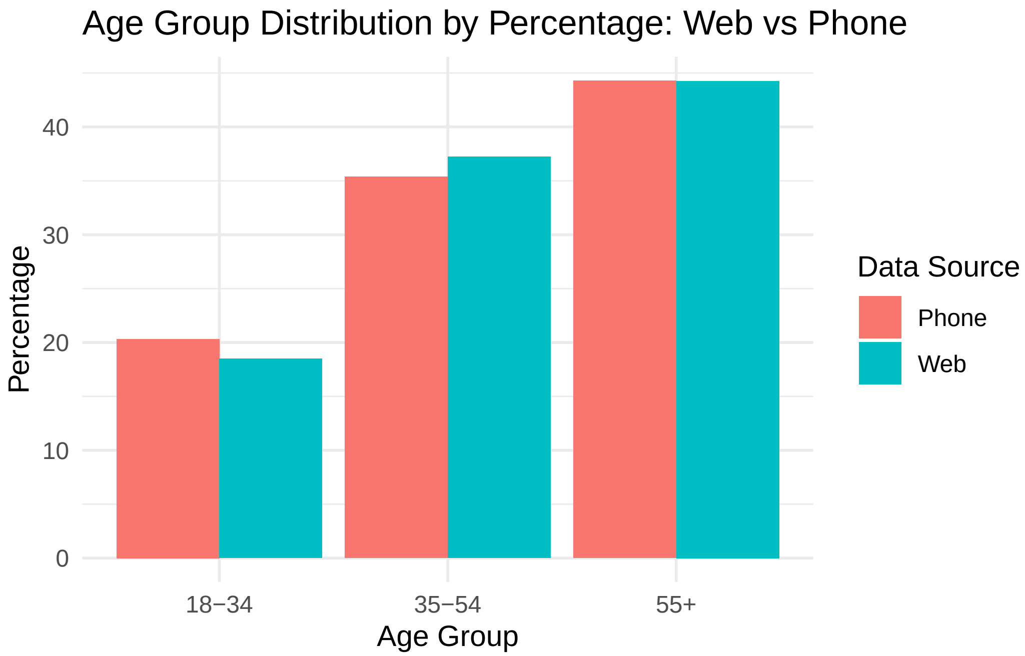

Figure 1: A comparison of the number of people interviewed by phone or web across different age groups

4 Outcome of Interest

The outcome of interest is the likelihood of voting in the upcoming election. As outlined in Section 2, I created a binary variable for both phone and web datasets, and then calculated the proportion of respondents who reported that they were likely to vote or had already voted. For phone data, 95.1% of respondents reported already voted or strong intention to vote in advanced polling, compared to 91.3% for web data. Additionally, we calculated the 95%

Table 1: 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)

confidence interval (CI) for the probable turnout rates in both datasets using the formula:

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

where $\hat{p}$ is the proportion, z is the z-score for a 95% CI, and SE is the standard error. The computed 95% confidence intervals are provided in Table 1.

Both datasets show a high proportion of respondents who reported either voting or intending to vote, with estimates in the 90th percentile. However, this contrasts with the actual voter turnout rate of 68.3% in the 2015 Canadian election (Canada n.d.). This discrepancy suggests a gap between self-reported voting intentions and actual voting behavior.

One explanation for this gap is that individuals more likely to vote may also be more likely to participate in surveys, whether conducted via phone or web. This introduces systematic differences between survey respondents and non-respondents, which could bias analysis if not accounted for. Another potential reason is that external barriers prevent people from voting despite their intentions. For example, access to voting for individuals with disabilities remains inadequate in Canada (Salvino 2024). Older adults, who are more likely to have disabilities, form a significant portion of the population. While they may intend to vote, insufficient accommodations in polling station may deter them. This highlights the need for policy interventions to address these barriers and enhance voter turnout, ensuring that everyone can exercise their right to vote.

5 Comparative Analysis

Stephenson et al. (2020a) targeted a sample consisting of 28% respondents aged 18-34, 33% aged 35-54, and 39% aged 55 and older. However, as shown in **?@tbl-1**, only 18% of web survey respondents were aged 18-34, while 45% were aged 55 and older. These trend in underrepresentation of young people is also present in the phone survey data, biasing the data toward older respondents. This bias affects the large proportion observed in Table 1, as older demographics historically exhibit higher voter turnout. In the 2015 Canadian federal election, turnout for 18-34-year-olds was approximately 57%, compared to 75% for individuals aged 55 and older (Canada (n.d.)). Thus, the sample composition skews the proportion of likely voters toward the 90th percentile, while historical data suggests an overall turnout closer to 66%.

Creating a dataset more representative of young people is challenging. Young individuals are harder to reach for surveyors, and many choose not to participate due to political apathy. This detachment from the political process stems from growing skepticism about government efficacy and a sense of exclusion from decision-making process that impacts their future (Berthin 2023). This means that younger people are less likely to respond to election surveys. Additionally, young people are less likely to respond to phone surveys, and while online polls are more accessible, competing priorities such as work and social activities reduce participation.

Phone data also suffers from response bias, as individuals who avoid or refuse surveys may differ systematically from participants. This nonresponse leads to a sample that does not fully represent the adult Canadian population. On the other hand, web surveys are subject to coverage bias. Reliable internet access is limited in rural areas, and older populations and low-income groups are less likely to have internet access. These biases result in underrepresentation of certain population segments, affecting the generalizability of findings.

6 Generative AI Statement

Large Language Models (LLMs) were used for two purposes in this report. First, ChatGPT was employed to generate code for Figure 1. Specifically, the LLM was used to transform the graph into a percentage format. I adjusted all labels and column names to align with previous code chunks. Second, while all the text in this report is my own, I used ChatGPT to identify and correct grammatical, spelling, and punctuation errors. It served effectively as a proofreader.

Bibliography

Berthin, Gerardo. 2023. "Why Are Youth Dissatisfied with Democracy?" *Freedom House*, October. https://freedomhouse.org/article/why-are-youth-dissatisfied-democracy.

Canada, Elections. n.d. "Voter Turnout by Sex and Age." Elections Canada. https://www.elections.ca/content.aspx?section=res&dir=rec%2Feval%2Fpes2019%2Fvtsa&document=index&lang=e.

Salvino, Caitlin. 2024. "The Case for a Constitutional Right to Barrier-Free Voting for Electors with Disabilities." *Canadian Journal of Disability Studies* 13 (3): 102–40.

Stephenson, Laura B., Allison Harell, Daniel Rubenson, and Peter John Loewen. 2020a. "2019 Canadian Election Study – Online Survey." Harvard Dataverse. https://doi.org/10.7910/DVN/DUS88V.

———. 2020b. "2019 Canadian Election Study – Phone Survey." Harvard Dataverse. https://doi.org/10.7910/DVN/8RHLG1.