Are Women More Likely To Vote Liberal?

STA304 - Winter 2025 - Assignment 2

GROUP NUMBER: 77

1 Introduction

In this section you will briefly describe your report. Explain the importance of the subsequent analysis and prepare the reader for what they will read in the subsequent sections. Provide an overview of the research question. Briefly describe the 2019 Canadian Federal Election Study and its relevance. State the purpose and goals/hypotheses of the report.

Elections are a cornerstone of democracy, and understanding voter behavior is crucial for predicting outcomes and interpreting political trends. Election polls not only reflect public opinion but can also influence voter turnout and vote choice (Dahlgaard, 2017). However, the validity of these polls depends on how data is collected. Phone surveys may oversample certain demographics due to selection bias, while web surveys are prone to coverage errors and self-selection biases (Harrison, 2023). These methodological differences can shape demographic distributions and affect the conclusions drawn from the data.

Knowing the liberals won the election, we would like to gain some insgihts on the demographic breakdown of their votes This study investigates the effect of **gender** on the **intention to vote Liberal**, using data from the 2019 Canadian Election Study (CES). Females tend to have higher intention to vote for liberal parties, given that these parties are generally characterized by support for … . Investigating the relationship between gender and liberal voting intention shows us how socioeconomic factors influence political preferences, particularly in the context of growing ideological polarization.

Overall, by exploring how the distribution in the study's participants' education level and the proportion of people intending to vote for the liberal party can help us assess potential biases introduced by survey methodology. Hence this study contributes to a deeper understanding of voter behavior and the importance of survey design in political research.

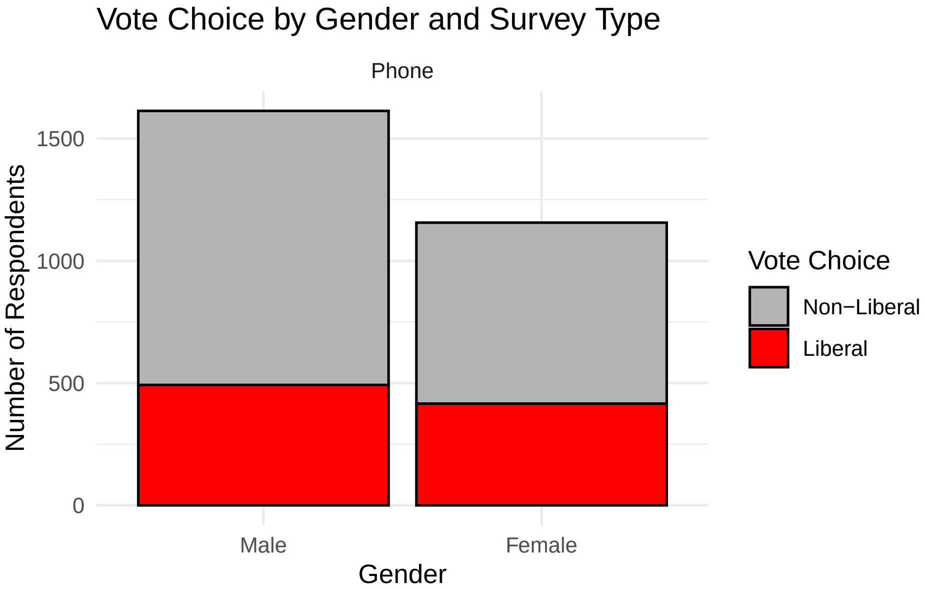
2 Data

Briefly introduce the data and key variables of interest. If you do any general data cleaning or data processing you should describe it (in a reproducible manner) here. Identify the stratification variable used. Include at least one plot displaying the distribution of the strata variable. If you do any data cleaning or data processing to the you should describe it (in a reproducible manner) in this section.

The analysis draws from the 2019 Canadian Federal Election Study (CES), which was collected using stratitifed random sampling by gender. Total male population and total female population were retrieved from the 2021 Census of Population Statistics Canada. In 2021, 50.7% of the Canadian population were women (18.77 million out of 37.6 million). For the CES, data was obtained via a phone survey where 2,769 responses were obtained with only 41.7% of the respondents being female. The participants were asked a series of questions regarding their sociodemographic background and their intent to vote in the upcoming election, including interest in the election, likelihood to vote, and intended vote choice.

In this study, the data cleaning involved two main steps. First, we dichotomized the primary outcome variable, voting intention for the people's party. Second, entries with missing or invalid responses for intended party or gender were excluded. Additionally, only male and female responders were considered given the small number of people in the other gender categories. The 'other' category had a very small sample size, which could limit meaningful statistical analysis and may have led to unreliable estimates. Given this, we chose to exclude this category to ensure more stable and interpretable results. We acknowledge that this decision reduces the inclusivity of our analysis and may overlook important differences in experiences. Future research with a larger and more representative sample should aim to better capture gender diversity.

Be sure to have text describing any plots or tables included.



3 Methods

This study investigates whether gender influences the likelihood of voting for the Liberal Party, using data from the 2019 Canadian Federal Election Study (CES), which was collected through stratified random sampling. Since stratified sampling ensures representation across key subgroups, all statistical estimates—including the proportion of Liberal voters and the logistic regression model predicting voting likelihood—account for survey weighting and finite population correction (FPC) (Lohr, 2019).

To estimate the proportion of Liberal voters, we use a weighted mean across strata. The estimated proportion is calculated as:

$$\hat{p}_{st} = \sum_{h=1}^{H} W_h \hat{p}_h$$

where H represents the number of strata (e.g., provinces or education levels), $W_h = N_h/N$ is the stratum weight based on its share of the total population, and $\hat{p}_h$ is the proportion of Liberal voters within each stratum. The 95% confidence interval (CI) is given by:

$$CI = \hat{p}_{st} \pm z_{\alpha/2}\sqrt{\sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(\frac{\hat{p}_h(1-\hat{p}_h)}{n_h}\right)}$$

where n_h is the sample size within each stratum, N_h is the population size in that stratum, and $(1-n_h/N_h)$ is the finite population correction (FPC), which accounts for cases where the sample represents a substantial fraction of the population. This formula is referenced from Ziegel, E. R., & Lohr, S. L. (2000). Sampling: Design and Analysis. *Technometrics.* Without this adjustment, confidence intervals could be overly wide, leading to inflated uncertainty.

To analyze the relationship between gender and voting preference, we fit a survey-weighted logistic regression model, which adjusted for stratification effects and unequal selection probabilities. Since the dependent variable (voting for the Liberal Party) was binary (1 = Yes, 0 = No), logistic regression was the appropriate modeling choice (Lumley, 2010). The model is specified as:

$$\log\left(\frac{P(VoteLiberal=1)}{1-P(VoteLiberal=1)}\right)=\beta_0+\beta_1\times\text{Gender}_{\text{female}}+\beta_2\times\text{Age}$$

Here, β_1 represents the effect of gender. If the corresponding odds ratio e^{β_1} is greater than 1, it suggests that women are more likely to vote for the Liberal Party compared to men. If the odds ratio is less than 1, it suggests women are less likely to vote Liberal compared to men. Similarly, β_2 captures the influence of age. An odds ratio greater than 1 indicates that an increase in age is associated with a higher likelihood of voting for the Liberal Party, while an odds ratio less than 1 suggests the opposite.

Since the dataset was collected using stratified sampling, failing to account for this in the regression model would lead to biased coefficient estimates and incorrect standard errors. We apply survey-weighted logistic regression (svyglm()) from the survey package in R, incorporating design weights and finite population correction (Lumley, 2010).

4 Results

Present a table showing the estimated proportion of votes for the selected party along with the 95% confidence interval, and include text describing this table and the key takeaways.

Table 1: Stratified Confidence Interval for Liberal Voters (Phone Survey).

Survey Group	Proportion Voting Liberal	95% Confidence Interval
Overall	**0.333**	**(0.315, 0.350)**
Male	**0.305**	**(0.283, 0.327)**
Female	**0.360**	**(0.332, 0.388)**

The proportion of respondents in the phone survey who reported voting Liberal was 0.333, with a 95% confidence interval of (0.315, 0.350), indicating that the true population proportion is likely within this range.

	Dependent variable		
Predictors	Odds Ratios	CI	p
(Intercept)	0.26	0.20 – 0.34	**<0.001**
gender [2]	1.27	1.08 – 1.49	**0.004**
age	1.01	1.01 – 1.02	**<0.001**
Observations	2769		
R^2 / R^2 adjusted	0.008 / 0.007		

The results from the survey-weighted logistic regression indicate that gender and age are both significant predictors of voting for the Liberal Party. The odds ratio for gender (female) is 1.27 (95% CI: 1.08–1.49, p = 0.004), suggesting that women have 27% higher odds of voting Liberal compared to men. Since the odds ratio is greater than 1, this indicates a positive association between being female and voting Liberal. Similarly, the odds ratio for age is 1.01 (95% CI: 1.01–1.02, $p < 0.001$), implying that older individuals are slightly more likely to vote Liberal. While the effect size for age is small, it is statistically significant.

5 Discussion

Summarize key findings. Discuss limitations of the analysis (e.g., potential biases, missing variables, survey errors). Provide recommendations for future research or improvements.

6 Generative AI Statement

Here is where you can explain your usage of Generative AI tool(s). Be sure to reference any tools with inline citations.

In the completion of this assignment, generative AI (OpenAI, 2025) was used to assist in struc- turing the analysis and generating explanations related to the influence of survey methodology on political preferences. Specifically, AI tools helped draft and refine sections on the potential biases introduced by phone and web surveys, ensuring clarity and conciseness in presenting the comparative analysis. Additionally, AI was used to fix the formatting of visualizations and synthesize my initial draft of the paper before a subsequent re-write from my end, this was done to improve the overall coherence and readability of the final report. The use of gen- erative AI ensured a more streamlined process and provided valuable insights into improving the presentation of data. # 7 Ethics Statement

Explain how you ensured that your analysis is reproducible (e.g., documenting code, using proper statistical methods).

Since the CES 2019 data is publicly available, describe whether or not this the work completed in your report needs Research Ethics Board approval for the report the be made publicly available. Be sure to specifically discuss the privacy of human participants in this study.

Our analysis ensures reproducibility by documenting all experimental decisions and employing appropriate statistical methods. We used well-established techniques and provided clear methodological descriptions to facilitate replication.

Since CES 2019 data is publicly available, this report does not require Research Ethics Board approval for public release. The dataset is anonymized, ensuring no identifiable information about human participants is disclosed, thereby maintaining privacy and confidentiality.

8 Bibliography

1. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 4, 1991)

4. Vaughn, B. K. (2008). Data analysis using regression and multilevel/hierarchical models, by Gelman, A., & Hill, J [Review of *Data analysis using regression and multilevel/hierarchical models, by Gelman, A., & Hill, J*]. *Journal of Educational Measurement*, *45*(1), 94–97. Blackwell Publishing Inc. https://doi.org/10.1111/j.1745-3984.2007.00053_2.x

5. Ziegel, E. R., & Lohr, S. L. (2000). Sampling: Design and Analysis. *Technometrics*, *42*(2), 223-. https://doi.org/10.2307/1271491

6.

9 Appendix

Any additional notes/derivations that are supplementary to the report can be added in an appendix. This section will not be directly graded, but may be included for completion-sake.