ADD IN MEANINGFUL TITLE HERE

STA304 - Winter 2025 - Assignment 2

Group 24: You Yu Fu, Eric Guo, Finn Tran, Lina Jin

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.

2 Data

For this project, we will be analyzing the phone survey data set to investigate if education level and age affect voting party preference. To do this, we will be stratifying our data into based on education level. To do this, we will be using the International Standard Classification of Education (ISCED) levels to group our participants based on their response to q61 in the Phone Data Dictionary which asks "What is the highest level of education that you have completed?"

Table 1: Our mapping of Phone Data Dictionary Q61 to ISCED_Level and ISCED_Level meaning

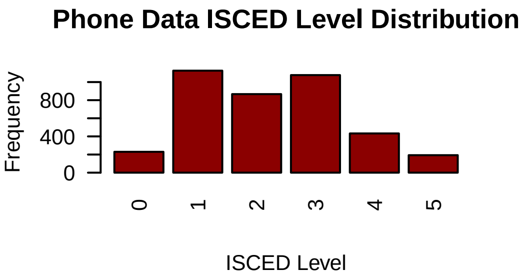
Response	Corresponding Education Level	ISCED Level Definition
No schooling	0	Early childhood education
Some elementary school	0	Early childhood education
Completed elementary school	0	Early childhood education
Some secondary/high school	0	Lower secondary education
Completed secondary/high school	1	High school diploma
Some technical, community college	1	High school diploma

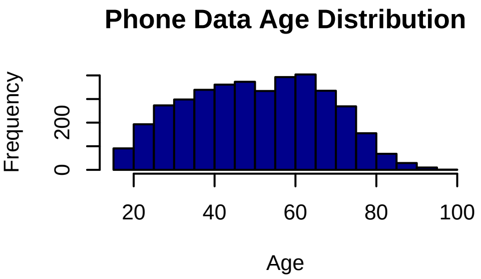
Response	Corresponding Education Level	ISCED Level Definition
Some university	1	High school diploma
Completed technical, community college	2	Short-cycle tertiary education
Bachelor's degree	3	Bachelor or equivalent level
Master's degree	4	Master/PhD or equivalent level
Professional degree or doctorate	5	PhD or equivalent level

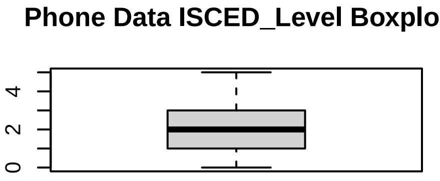
We used the above mapping to map respondent answers to ISCED Levels using the ISCED_Level variable. Respondents who picked the option ("Don't know", "Refused to answer" or "Skipped") were removed from the data set, with only 11 out of the over 4000 participants choosing this response. The other variables of interest were cleaned, where we removed all NA values and invalid voter ages (age < 18). We used q11 to determine surveyor voting choice, where we grouped them into three categories: "Liberal" for those who said they would vote Liberal, "Not Liberal" for those who said they would vote for a party that was not the Liberal party, and "Unknown" for those who skipped the question or chose not to answer.

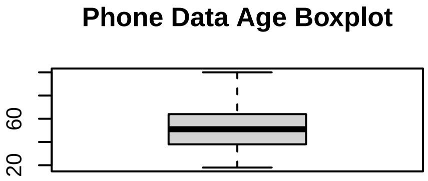
Given that over 25% of phone survey participants chose not to say which party they were going to vote for in the upcoming election, we will generalize this category (q11) into 3 responses, 'Unknown', 'Liberal', and 'Not Liberal'. Removing all these observations could introduce bias into our analysis and decrease the significance of our results.

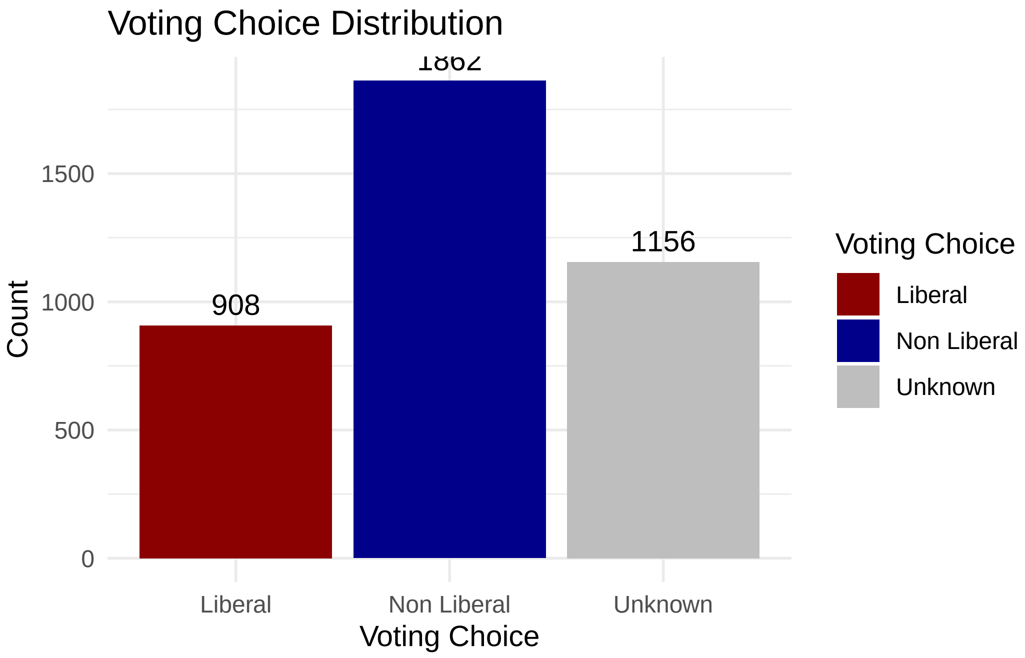
Plotting ISCED_Level (based on q61), and Voting preference (based on q11):











```
[1] "ISCED LEVEL"
```

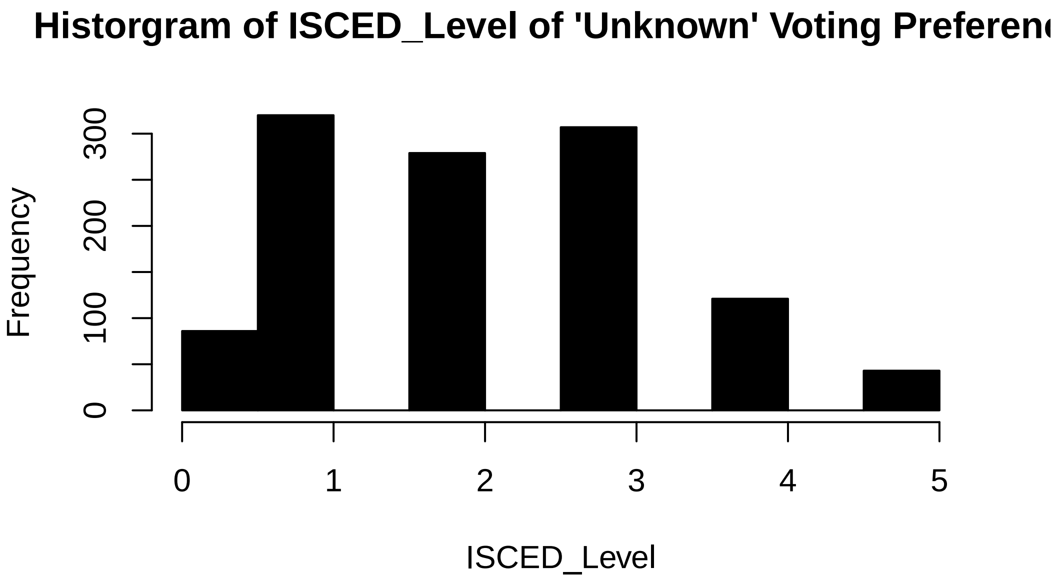
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.000   2.000   2.238   3.000   5.000 
```

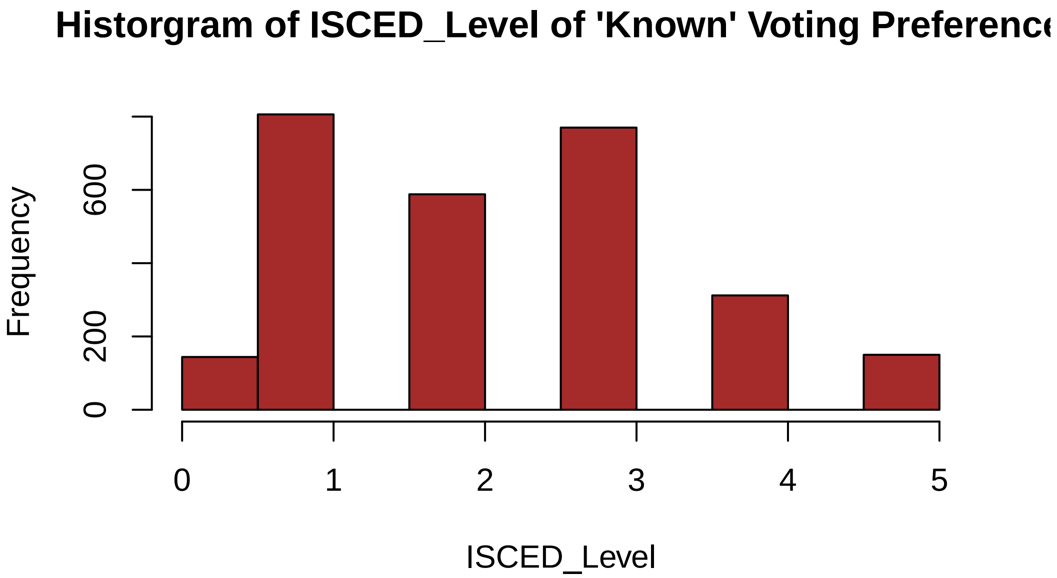
```
[1] "AGE"
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   38.00   51.00   50.89   64.00  100.00 
```

From the summary statistics and graph, we can see that the survey population is decently well educated, with a mean and median ISCED score of around 2. This means that the average survey respondent has at least completed community college. The ISCED_Level distribution appears to be right skewed, with the interquartile range from 1 to 3 from the box plot and histogram. The age seems slightly right-skewed, with the interquartile range from late 30's to early 60's from the box plot, but has a higher range of ages from 18 to late 90's. Among the web survey respondents, it shows a slight bi-modal distribution on the histogram, with the interquartile range from from late 30's to early 60's.

We can make further comparisons on phone survey participants in the 'Unknown' category with those who provided an answer.





These distributions are very similar, therefore, they could be removed without introducing bias into our data.

3 Methods

To calculate the Confidence Interval for a proportion and stratified sampling, the following formula will be used:

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

Where $\hat{p} = \sum_{h=1}^{H} W_h \hat{p}_h$ and $W_h = \frac{N_h}{N}$

In the formula,

- H is the number of Education Level strata.
- h denotes a specific stratum.
- $\hat{p}_h$ specifies the sample proportion of Liberal voters from the strata h.
- N_h is the population size of the stratum h, with numbers from Statistics Canada (Statistics Canada, 2021).
- n_h is the sample size of the stratum h.
- s_h^2 indicates the sample variance of the stratum h.
- $Z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to a 95% confidence level ($\alpha = 0.05$)

Below is the constructed logistic regression model:

$$\log\left(\frac{1-P(\text{Vote Liberal})}{P(\text{Vote Liberal})}\right) = \beta_0 + \beta_1(\text{Education Level}) + \beta_2(\text{Age})$$

- Education Level, the ordinal variable that the population is stratified by, defining the respondent's education level
- Age, discrete variable for the respondent's age
- $P(VoteLiberal)$ describes the probability of the person voting for the Liberal party. The intercept value β_0 references the expected log-odds of a person with Education Level = 0 and age 0 to vote for the Liberal party. The interpretation is not relevant as a person of age 0 cannot vote.
- The coefficient β_1 identifies the expected effect of Education Level in log-odds of voting Liberal compared to the baseline education level (ISCED Level = 0) when Age is kept constant.
- β_2 is the expected effect of an unit change in Age on the log-odds of voting Liberal, when Education Level is kept constant.

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.

In Table 1 I present the confidence intervals of phone surveys:

Table 2: The proportions and 95% confidence intervals of outcome variable of interest calculated for the Canadian Election Study 2019 phone survey data.

	Proportion of Outcome Variable	95% Confidence Interval of Outcome Variable
Phone Survey	0.20914	(0.1950576, 0.2232315)

```
Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

Below is the estimate regression model:

	Estimate (β)	Standard Error	p-value
Intercept	-2.373643	0.163359	< 2e-16 ***
ISCED Level	0.208224	0.038918	9.28e-08 ***
Age	0.012804	0.002728	2.77e-06 ***

Accounting for stratification, -2.373643 is the log-odds intercept for a person with no school on the ISCED scale (ISCED level = 0) and zero years old. This is just a baseline variable, since anyone who is 0 years old cannot vote.

β_1 is positive and p-value is < 0.05, meaning that higher education is associated with higher odds of voting Liberal at significance level. The log odds is 0.208224, meaning that for each one-unit increase in education level, the log-odds of voting Liberal increase by around 20.8%.

β_2 is positive with a log-odds value of 0.012804 and p-value is < 0.05, meaning that older age is slightly associated with higher log-odds of voting Liberal.

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.

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

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.

8 Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (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. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)

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.