Assignment 1 Instructions

STA304 - Winter 2025

Samantha-Jo Caetano

Instructions

Please read all instructions carefully.

This is an individual assignment. While you are expected to work on this independently, you are more than welcome to discuss ideas, code, concepts, etc. regarding this assignment with your classmates.

Please do not share your code or your written text with your peers. It is expected that you write all your code and written work (unless they are taken from the materials provided in this course or are from a credible source which you have cited).

Please note, there is flexibility in this assignment, so the context of most of the work completed here should not match your peers. You are allowed to use Generative Artificial Intelligence to support your completion of the work, but it is recommended that you perform your own proofreading and editing following the usage of Generative AI. Please read through the "Generative AI" policy on the course syllabus and in the instructions of this assignment to ensure that your usage is inline with the requirements of this assessment.

There is a starter Rmd file (called Assignment1-starter_code.Rmd) available for you to use to start your code. We suggest you read the entire assignment before starting.

Submission Due: Thursday January 30th at 11:59pm ET

Your complete .Rmd file (or .qmd file) AND the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file or 'Render' from your .qmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/374323/assignments/1465255). We will only be accepting submissions through this Quercus page (i.e., we *not* be accepting email submissions). Please consult the course syllabus for other inquiries. **If you do NOT submit the pdf in your submission you will receive a grade of 0.**

There are three attempts to submit this assignment, to account for the possibility of an error in your first attempt/submission. If you submit prior to the January 30 11:59pm ET deadline, then we will grade the latest submission that came in prior to January 30 11:59pm ET.

If you wish to use the grace period please do NOT submit prior to January 30 11:59pm ET. Note: if you use the grace period we will grade the *latest* submission, so please ensure that you are including BOTH the pdf and Rmd in your upload/submission.

Assignment grading

This assignment is to be a report. The page limit is 5 pages in total (this includes the bibliography and Generative AI statement).

In this report you will perform a comparative data analysis and describe your insights/findings. Thus, the assignment requires coding, analysis and written communication. We recommend you spellcheck and proofread your written work.

We will be directly marking the pdf files, so please ensure that your final submission looks as you want it to look before submitting it.

As mentioned above, this assignment will be marked based on the output in the pdf submission. You must submit both the Rmd (or qmd) and pdf files for this assignment to receive full marks in terms of reproduciblity. **If you do NOT submit both the pdf AND Rmd in your submission you will receive a grade of 0.**

This assignment will be graded based off the rubric available on the Assignment Quercus page (link: https://q.utoronto.ca/courses/374323/assignments/1465255) - the rubric will be available at least one week in advance of the due date. TAs will look over each section and select the appropriate grade for that section based off a brief overview (one-time read over) of that section. Your assignment should be well understood to the average university level student after reading it once.

We would suggest you make sure your document looks clean, aesthetically pleasing, and has been proofread. You will be able to see the rubric grade for each section. There may be some comments/feedback provided (by the TAs) if the same issue seems to be arising in multiple sections, but you will likely receive no comments/feedback (due to the size of the class and limited time for marking).

Assignment 1: Report

Survey Design & Analysis: Comparing Phone and Web Surveys

Objective

The objective of this assignment is to explore how different survey methodologies (phone surveys vs. web surveys) can produce different demographic distributions and, in turn, affect the inferences made about public opinion. You will be using two datasets from the 2019 Canadian Election Study (CES), one collected via phone surveys and the other via web surveys. The starter code has cleaned up versions of the data, but more documentation and information regarding accessing the entire raw files is available here in the cesR documentation.

Your task is to analyze the two datasets and reflect on the potential biases in each survey method, their impact on demographic representation, and the implications for drawing conclusions about Canadian voters.

Deliverables:

You will produce a report (pdf) that is completely digestible to a university level student (who understands what a confidence interval generally is, but may not be familiar with mathematical theory or R code/output). The restrictions are as follows:

- Maximum page length is 5 pages (for the pdf)
- Standard margin sizes (i.e., 2.5cm)
- Standard font sizes (i.e., 12 pt font)
- Any plots, tables, and output are neatly presented and organized.
- No visible code in the pdf.

Note: Anything beyond 5 pages will not be read by the grader. I.e., we will read pages 1, 2, 3, 4, and 5, but will stop reading on page 6.

Assignment Components:

1 Introduction (1-3 paragraphs)

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.

2 Data (1-3 paragraphs)

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. It might be helpful to explain the variables of interest (i.e., the ones you will present in the subsequent sections) here and give rationale as to why investigating these variables is important to the study/inference. If you do any data cleaning or data processing to the demographic variable or outcome variable that you select (in sections 3 and 4) you should describe it (in a reproducible manner), so that the variables are clearly defined.

3 Choose a Demographic Variable (1-3 paragraphs)

You will choose ONE of the following demographic variables to showcase:

- Age
- Gender
- Province
- Education

Your first task will be to compare the distribution of this demographic variable across the two surveys (phone vs. web). To do this, you will create side-by-side visualizations (e.g., two bar plots, two boxplots, or two histograms, etc.) showing the distribution of your chosen demographic variable for both the phone and web survey datasets. Discuss how these distributions compare and what they imply about the populations reached by each survey method. You should describe the target, frame and sample populations of each data set.

4 Choose an Outcome of Interest (1-3 paragraphs)

Next, choose one outcome of interest from the following list:

- Interest in the Canadian federal election
- Likelihood of voting in the next Canadian federal election
- Voting intentions (i.e., who they would vote for in the next election)

You will analyze this outcome in both datasets. For each survey (phone and web), calculate the proportion of respondents indicating a particular response to the chosen outcome and compute a 95% confidence interval for that proportion. Present both confidence intervals in a neatly organized, and labelled/captioned table. Compare the results between the two surveys to explore whether different methodologies result in different findings about the Canadian electorate's behavior or opinions. Be sure to also write out the formula for the confidence interval which was used (and reference the textbook or course which it is from).

Note: you may need to refine your chosen outcome further (e.g., proportion of voters who will vote for the Liberal party or proportion of people who are interested in the election (i.e., a score of 7 out of 10 or above)). It should be clear what the outcome is, and why it is relevant.

5 Comparative Analysis (3-5 paragraphs)

Write a reflection based on the following prompts:

- Demographic Differences: What are the key differences (or similarities) between the two surveys in terms of the demographic variable you chose? Are certain groups over- or underrepresented in either the phone survey or the web survey? How might this impact the findings of the survey?
- Biases/Errors: Reflect on potential biases/errors in both survey methodologies. How do these biases affect the population sampled and the generalizability of the results?
- Implications for Analysis: How might the different demographics captured by each survey method affect the analysis of the outcome of interest? Consider whether certain demographic groups are more likely to hold particular views, and how their representation or lack thereof in the samples may influence your conclusions about the Canadian electorate.

6 Generative AI or Workflow Statement (1-2 paragraphs)

If you have used generative AI tools (e.g., ChatGPT, other writing assistants) to help write your report, please include a brief reflection (one to two paragraphs) on how you used these tools. This should include:

- What specific tasks you used the AI for (e.g., writing assistance, literature review, generating code, etc.).
- How you ensured that the final report was your own work and aligned with the assignment's requirements.

Please note: The use of AI tools should supplement, not replace, your own critical thinking and analysis. Ensure that you cite and properly attribute any content generated by AI.

If you did not use generative AI tools on this assessment, please include a brief statement outlining your workflow for completing this assignment. This statement should include timelines and a general description of any resources you used (e.g., websites, textbooks, writing centers, etc.).

7 Bibliography

Include a bibliography with at least three academic sources or relevant references that help contextualize your analysis (e.g., references on survey methodology, biases in data collection, etc.). These sources should be cited in the lines of your report.