Assignment 1 Instructions

STA304 - Fall 2025

Samantha-Jo Caetano

Instructions

Please read all instructions carefully.

This is a group assignment. You are expected to work on this either independently or in a group of up to 4. You are expected to work exclusively with your group-mates and not other groups. You are more than welcome to discuss ideas, code, concepts, etc. regarding this assignment with your class mates, but only share your writing and your code with your groupmates. Do not share your code or your written text with peers outside of your group. It is expected that all code and written work should be written by members of your group (unless they are taken from the materials provided in this course or are from a credible source which you have cited).

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-startercode.qmd) available for you to use to start your code. We suggest you read the entire assignment before starting.

Submission Due: Thursday September 25th at 11:59pm ET

Your submission will consist of three components:

1. .qmd file (submitted as a Group)
2. .pdf file (submitted as a Group)
3. Assignment 1 - Group Work Survey (completed as an individual. This is to be completed even if you worked alone)

Group Work Submission

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/404574/assignments/1591844).

Please note that only one group member needs to submit the .qmd and .pdf files onto Quercus in ONE submission. We will be directly marking on the LATEST submission of the .pdf (submitted on/before the due date/time). All group members will receive the same grade. 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. Both the .qmd (or .Rmd) file and the resulting PDF must be included in your submission. Failure to submit both files will result in a 20% grade deduction.

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 September 25 11:59pm ET deadline, then we will grade the latest submission that came in prior to September 25 11:59pm ET.

If you wish to use the grace period please do NOT submit prior to September 25 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/qmd in your upload/submission.

Assignment grading

This assignment is to be a report. The page limit is 7 pages in total (this does not include the Generative AI statement, Bibliography or Appendix).

In this report you will design a survey, perform a short data analysis and describe your insights. 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 20% grade deduction.**

This assignment will be graded based off the rubric available on the Assignment Quercus page (link: https://q.utoronto.ca/courses/404574/assignments/1591844) - the rubric will be available at least one week in advance of the due date. TAs will evaluate each section based on a single careful read-through. Your report should therefore be clear, concise, and understandable to a university-level reader on the first read.

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: STEM Buddies Handbook Design

Objective

This assignment is in partnership with STEMBuddies (https://stembuddies.ca/). STEMBuddies is a nonprofit organization that provides free, accessible resources to help underserved students navigate high school and pursue post-secondary education. Their work focuses on developing mentorship supports, workshops, and post-secondary-focused programming. They have helped over 700 students, from many underrepresented communities, build confidence and develop their skills to create clearer path forward in STEM.

The objective of this assignment is to help STEMBuddies (a nonprofit which help with resource support for highschool students) design a survey to assess the needs, preferences, and feedback of high school students regarding the development of the STEMBuddies handbook. They are looking for a way to gauge what is missing from the handbook (i.e., what are the needs of their population) as well as ways to improve the handbook (i.e., feedback from the population about the current draft). The purpose of the handbook is to support students through their academic and personal journeys.

In Sections 2-3 you will design a develop a survey to gain feedback from the students regarding their needs and/or their relations to the handbook. In Sections 4-6 you will perform an exploratory data analysis using *data simulated to mimic data collected from the survey you showcased in Section 3*. Your data analysis will consist of one table, one graph and one confidence interval.

Deliveribles:

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 7 pages (for the pdf - not including Generative AI statement, Bibliography and Appendix)
- 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.
- Your final PDF must not contain visible code. Only results, figures, and written explanations should appear.

Note: Any content from Sections 1-6 beyond the first 7 pages will not be read by the grader.

1 Introduction (1-3 paragraphs)

Explain the goal/topic of your survey. Why is this topic relevant? Note: you will likely need citations for this. Be sure to explicitly describe how the survey will contribute to the overall goal/topic.

2 Survey Showcasing (3-4 paragraphs)

Provide a link to your survey. The survey should include: (i) Introduction section, (ii) at least 2 demographic survey questions, (iii) At least 2 needs assessments survey questions, and (iv) Conclusion/closing message. The survey could be built using any of the following platforms: SurveyMonkey, Google Forms, Microsoft Forms, Typeform, Qualtrics, etc. All questions of your survey should be included (and nicely formatted) in the Appendix of your report.

Briefly describe your survey and the general questions (all survey questions (along with any category selection options) should be neatly typed up in an Appendix). Be sure to describe any decisions you made and justify them accordingly (e.g., if you are categorizing something explain why the categories chosen were selected). Briefly describe how you tested your survey.

Choose at least 1 question in your survey to showcase in this section of the submission (Note: in this section you should be showcasing any relevant questions to subsequent parts of this report). Copy and paste the analysis relevant question(s) (and make sure they are formatted nicely and easy to find/read - you could italicize this question(s) or put it/them into a well-formatted labelled and numbered table) into this section of the document. Explain why you chose this question(s) and the benefits and drawbacks of the design/wording of the question(s). For each drawback you list, justify why you couldn't avoid/address it in the design of your survey.

3 Procedure (2-4 paragraphs)

Propose how you would suggest implementing the sampling process that STEMBuddies should use for collecting survey feedback (ie. provide a procedural outline of how you propose STEMBuddies should collect the data from the population - try to be realistic). Here you should identify a target population, a reasonable sampling frame (if you are suggesting using one), and the (proposed) sample units. Identify the strengths of your proposed sampling procedure and highlight any potential errors/biases.

Since we cannot collect data on this population, you will simulate your survey data to perform the analysis in the subsequent sections. Here you will describe how you simulated the data for the relevant questions in your analysis (i.e., describe this in a reproducible manner for the question(s) you showcased in the previous section). Note: You in your R code you must simulate data for ALL of the survey questions you have developed and show a glimpse of this simulated data in your Appendix, but you do not need to describe how you simulated all of the data. Be sure that the simulated data matches the expected inputs of your survey questions.

4 Data (3-6 paragraphs)

The goal of the Data section is to introduce the reader to the data, showcase some meaningful aspects of the data, and get them thinking about potential hypotheses/findings. Including:

- A reproducible summary of any data cleaning. Note: Someone (who is NOT necessarily familiar with Tidyverse functions) should be able to read this section and reproduce your cleaning process based off reading your description.
- A description of the variable(s) in the main analysis. Note: this should align with the question(s) highlighted in Section 2.
- At least 1 aesthetically pleasing plot/graph/figure of the variable(s) of interest and this should be labelled and numbered.
- Text explaining all tables and figures in this section.
- Some text (and perhaps graphical summaries) of the variables you will perform the hypothesis test and/or confidence interval on. This should help prep the reader in understanding why the test or interval is important and whether it is appropriate.

5 Methods (1-2 paragraphs)

The goal of the Methods section is to introduce the reader to the statistical methods that you will be using to analyze the data. Include the formula for calculating the confidence interval (do not include specific numbers yet) and provide a description of the confidence interval's components. In this section you are preparing the reader for how to interpret the numbers displayed in the next section (Results). Be sure to explain and justify any assumptions. Reminder that explanations are for a general science reader, not a statistician.

Note: if you want to show derivations of your hypothesis test statistic and/or confidence interval, your derivations should be placed in an Appendix.

6 Results (1-2 paragraphs)

The goal of the Results section is to present the results of the statistical analyses to the reader. Here you will present a (well formatted, labelled and numbered) **table** showing the estimated confidence interval results, and this section will include text describing this table and the key takeaways.

Be sure to provide commentary on whether the results appear reasonable and consistent with expectations. Remember there is the option to use in-line R code to produce output in text (e.g. The mean is `` ` r mean(x) ` ``.).

7 Generative AI or Workflow Statement

If you have used generative AI tools (e.g., ChatGPT, CoPilot, etc.) to help write or code this assignment, please include a reproducible reflection 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.

8 Bibliography

Include a bibliography with at least five academic sources (or relevant references) that help contextualize your motivation, design, and analysis for this report (e.g., references on survey methodology, biases in data collection, STEMBuddies website/handbook, etc.). These sources should be cited in the lines of your report.

If you are not sure what bibliography style to use you can use APA as a default (but you can use any well known bibliography style: APA, MLA, Chicago, etc.). Just be sure whatever style you use, you stick to it and the references are in a consistent style and are listed in order of appearance throughout the report or in alphabetical order.

9 Appendix

This should include:

- a well formatted copy of your survey questions
- a glimpse of your simulated data
- any other supplementary items you wish to include (e.g., more plots, derivations, etc.)