STA304 - Fall 2023

Assignment 1 Instructions

Samantha-Jo Caetano & Emily Somerset

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, you have a lot of flexibility to choose your subject for 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 September 28th at 11:59pm ET

Your complete .Rmd file AND the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/317099/assignments/1151557). 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 two attempts to submit this assignment. If you submit prior to the September 28 11:59pm ET deadline, then we will the latest submission that came in prior to Sept 28 11:59pm ET.

If you wish to use the grace period please do NOT submit prior to September 28 11:59pm ET. Note:if you used 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. Note: there is an Ouriginal draft submission page available (not for grades) for you to practice submitting your assignment if needed.

Assignment grading

There are three parts to this assignment. The first is theory-based where you will talk about the design of a survey. The second part is data analysis and communication/writing focused. And the third part is for referencing and including supplementary materials for the first two parts.

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 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/317099/assignments/1151557) - 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).

Part 1: Designing a survey

Description

In this question you will create a survey and describe and critique aspects of your survey. Create a survey about any topic with at least 6 questions, using any of the following platforms: SurveyMonkey, Google Forms, Microsoft Forms, Typeform, Qualtrics, etc.

You should touch on the following points (we suggest at least one paragraph for each of the following):

Goal

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.

Procedure

Propose how you would implement the survey (ie. provide a procedural outline of how you will collect data - be realistic). Here you should identify your target population, frame population, and sample population. Identify any drawbacks in your procedure. Justify the strengths of your proposed sampling procedure.

Showcasing the survey.

Provide a link to your survey.

Choose 3 questions in your survey to showcase in this submission. Copy and paste them (and make sure they are formatted nicely) into this part of the document. Explain why you chose these three questions and the benefits and drawbacks of each question. For each drawback you list, justify why you couldn't avoid/address it in the design of your survey.

General Notes (for Part 1):

- You should have a bibliography and you should use inline citations. If it is clear that you (or the reader) looked up something that is not common knowledge, then you will lose points if it is not cited.
- Please have a look at our course Resources page if you need help with Writing, LaTex, Tidyverse, etc.
- Grammar is *not* the main focus of the assessment, but it is important that you communicate in a clear and professional manner. Therefore, no slang or emojis should appear.
- If you introduce an acronym, be sure to define it before subsequent uses. For example: The University of Toronto (UofT) is a large school. At UofT we have a lot of students.

Part 2: Data Analysis

Description:

In this part you will write up a "Data" section, a "Methods" section, a "Results" section and an "Appendix" of a report, using *data simulated to mimic data collected from the survey you showcased in Part 1.* Your data analysis will consist of one hypothesis test and one confidence interval.

Minimum requirements:

- Data must be simulated to mimic the survey showcased in Part 1.
- Data needs to have a sample size that reflects the procedure of your survey showcased in Part 1.
- Data section should include at least one plot/figure.
- Data section should include at least one summary table.
- Your analysis must include: (i) at least one hypothesis test on some parameter of interest; and (ii) at least one confidence interval on a different parameter of interest. Thus, your "Methods" section will describe one hypothesis test and confidence interval to be calculated, and your "Results" section will showcase the output of the hypothesis test and confidence interval.

Data

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.

Your **Data** section should include the following:

- A detailed description of the simulation process, you should explain the entire simulation process so the reader can reproduce it (based solely on the writing).
- A summary of the cleaning process (if you cleaned the data). 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 variables in the main analysis.
- Some appropriate numerical summaries (at minimum center and spread, but something else may be more appropriate). Please put them in a well-formatted and labelled table. When possible, numerical summaries should be presented in the same table. For example: the center and spread can be shown in the same table.
- At least 1 aesthetically pleasing plot/graph/figure (No more than 4 plots - if you have more than 4 then create an Appendix section at the end of the report and add any supplementary figures and tables there).
- Text explaining each table or figure in the Data section and a some brief text referring to any tables and figures in the Appendix.
- 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.
- In line referencing if needed.
- Reference the programming language/software used to complete this section.

Methods

The goal of the Methods section is to introduce the reader to the statistical methods that you will be using to analyze the data.

Your **Methods** section should include the following:

- A complete explanation of what the methodologies: an explanation of the hypothesis test and confidence interval.
- Explain and justify any assumptions.
- An explanation of the parameters of interest (i.e., mean/variance/percentile/etc.).
- Explanations are for a general science reader, not a statistician.
- A description of why the method is appropriate based on assumptions, variable types and practical rationale.
- In-line referencing.
- Note: if you want to show derivations of your hypothesis test statistic and/or confidence interval, your derivations should be placed in an Appendix.

Results

The goal of the Results section is to present the results of the statistical analyses to the reader.

Your **Results** section should include the following:

- The results of the methodologies (parameter estimates, hypothesis test, and confidence interval) included in the report.
- An explanation and interpretation of the results.
- Some commentary on whether or not the results seem reasonable.
- Text explaining any tables or figures.
- In-line referencing.
- In-line R code to produce output in text (E.g. The mean is `` `r mean(x)` ``.).

General Notes (for Part 2):

- It is expected that your methods be a hypothesis test and a confidence interval. It is recommended that you use standard approaches, but this is not required.
- All tables/figures should be well-labelled, clean and captioned.
- With the exception of the glimpse output (in Part 3 - the Appendix), there should be no evidence that Part 2 is an assignment, we should be able to take a screenshot of this section and paste it into a newspaper/blog.
- There should be no raw code. Any output should be nicely formatted.
- You should reference the data, any outside code/documentation and any ideas/concepts that are taken outside of the course.
- Note, we are not marking grammar, but we are looking for clarity. If you need help with writing there are resources posted on the Course Info>Resources page of Quercus.
- Use full sentences.
- Be specific. Remember, you are selecting this topic/data and the reader/marker may not be familiar with it. A good principle is to assume that your audience is not aware of the subject matter.
- Remember to end each section with a concluding sentence. This means summarizing the key points from your writing.

Part 3: Referencing

Description:

In this part you include all references to external and supplementary materials that were used in Parts 1 and 2.

Generative AI Statement

Your **Generative AI Statement** should include a detailed description of any generative artificial intelligence tools used along with a reference(s)/citation(s). Please include a list of prompts, and explain in detail how the result of the prompts were used in this assignment (which sections of the assignment were constructed via the usage of generative AI and how).

It is expected that this section will likely be about 100-200 words, but maybe longer depending on your usage of different tools.

If you did not use any generative AI then please still include this section and explain that you did not use any generative AI tools. If this is the case, you do not need to include any references/citations to generative AI tools.

Bibliography

Your **Bibliography** section should include a list of well-formatted, consistent citations. No specific reference style is required, but you must use some formal reference style in order to ensure that the references are easy to understand.

Appendix

The goal of the Appendix is to include any secondary information to the reader.

Your **Appendix** section should include the following:

- A `glimpse()` of the raw data (for us to check the variables match that of the survey and the sample size is sufficient). You may also include glimpses of any cleaned data, but this is optional.
- Any additional plots or calculations that are not of primary necessity to the report, but should be included for completion-sake.

General Notes (for Part 3):

- Again, any external code/documentation and any ideas/concepts that are taken outside of the course need to be included in the bibliography and should be referenced within the text of Parts 1 and 2.
- Use full sentences.
- Be specific in the Generative AI section. Remember, we are aiming for reproducibility throughout the assessment. A good principle is to assume that your audience is not aware of the subject matter so be as detailed as possible when recanting your steps/processes.