STA304 - Fall 2021

Assignment 1 Instructions

Samantha-Jo Caetano

Instructions

This is an individual assignment. You are expected to work on this independently. You are more than welcome to discuss ideas, code, concepts, etc. regarding this assignment with your class mates. Please do not share your code or your written text with your peers. It is expected that all code and written work should be written by yourself (unless they are taken from the materials provided in this course or are from a credible source which you have cited). Please note, this assignment is fairly open, so the context of most of the work completed here should not match your peers.

There is a starter Rmd file (called Assignment1.Rmd) available for you to use to start your code.

Submission Due: Friday October 1st at 11:59pm ET

Your complete .Rmd file that you create for this assignment 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/236142/assignments/712397) by 11:59PM ET, on Friday, October 1st, 2021. Late assignments are not accepted. Please consult the course syllabus for other inquiries. **If you do NOT submit the pdf in your LATEST submission you will receive a grade of 0.**

Assignment grading

There are two parts to this assignment. One is theory-based and the other is data analysis and communication/writing focused. We recommend you spellcheck and proofread your written work. We will be directly marking the pdf files, thus 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 reproducible. **If you do NOT submit the pdf in your LATEST 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/236142/assignments/712397). TAs will look over each section and select the appropriate grade for that section based off a coarse overview (one-time read over) of that section. Your assignment should be well understood to the average university level student after reading it once. I 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 scaling of the class and marking).

Part 1

Description

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

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

Goal

Explain the goal/topic of your survey. Why is it (objectively) interesting (you will likely need citations for this)? Be sure to explicitly link 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 draw backs of each question.

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. I.e., no slang or emojis should appear.

Part 2

Description:

In this question you will write up a "Data" section, a "Methods" section, a "Results" section and an "Appendix" of a report, based on data collected (or simulated) from the survey in Part 1. Your data analysis will consist of a hypothesis test and a confidence interval.

Here you have two options: (i) You can implement your survey to collect data; or (ii) you can simulate your data.

Minimum requirements:

- Data needs to have a sample size of at least 20 (but the more the better - if you are simulating you should have much larger n).
- There is no expectation on the simulation.
- Methods 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.
- Data section should include at least one plot/figure.

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 data collection/simulation process.
- If you simulated the data you should explain the entire simulation process so one can reproduce it (based solely on the writing).
- If you collected data via the survey you must explain your collection process (again as one may wish to reproduce it) and comment on any drawbacks/obstacles.
- A summary of the cleaning process (if you cleaned the data). Again, one (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 important variables.
- Some appropriate numerical summaries (at minimum center and spread, but something else may be more appropriate). If there are a lot, please put them in a well formatted and labelled 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).
- Text explaining/highlighting each table or figure.
- Some text (and perhaps graphical summaries) of the variables you will perform the HT and/or CI on. This should help prep the reader in understanding why the HT/CI is important/interesting and whether it is appropriate.
- In line referencing/text 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 each methodology you are using entails. So an explanation of the hypothesis test and confidence interval.
- Explain/justify any assumptions.
- An explanation of the parameters of interest (i.e., mean/variance/percentile/etc.).
- An explanation of the method for a general science reader (i.e., not a statistician).
- A description of why the method is appropriate (based off assumptions, variable types and practical rationale).
- In line referencing.

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 (HT and CI) included in the report.
- An explanation/interpretation of the results.
- Some commentary on whether or not the results seem reasonable.
- Text explaining/highlighting any tables or figures.
- In line referencing.
- In line R code to produce output in text (E.g. The mean is `` `r mean(x)` ``.).

Bibliography

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

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 data (for us to check the variables match that of the survey and the sample size is sufficient). - 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 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 and clean.
- Everything in Part 2 should be written in full sentences/paragraphs.
- With the exception of the glimpse output in the Appendix, there should be no evidence that Part 2 is an assignment, I 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 will also need a reference/bibliography section. 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. It is important that you communicate in a clear and professional manner. I.e., no slang or emojis should appear.
- 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 reiterating the key points from your writing.