STA238 - Winter 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 with your class mates. Please do not share your code or your 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 a from a credible source which you have cited). Please note, this assignment is fairly open, so the context of most of the work compeleted here should not match your peers.

How do I hand in this assignment for the January 29th deadline ?

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/204754/assignments/541878) by 11:59PM ET, on Friday, January 29th. Late assignments are not accepted. Please consult the course syllabus for other inquiries.

Problem set grading

There are two parts to your problem set. One is largely theory-based and the other is more data analysis and communication/writing focused. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

Description

Assume you are an STA238 TA and you need to give a short tutorial on the Central Limit Theorem. In this question you will develop an exercise/sample question and provide a sample solution (similar to the exercises in the CLT slides/videos).

Write out a sample question/excerise which requires usage of the Central Limit Theorem to calculate the probability associated with (or finding a meaningful quantile of) the sample mean (or sum/total). Be sure that you make note of any assumptions and make sure the question is clear/well understood.

Now create a model solution to your exercise above. Your solution should be thorough, making use of R codes, probability theory, should give some explanation as to what the Central Limit Theorem is and should be digestible to the average STA238 student.

General Notes (for Part 1):

- 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.
- Try to be creative with the context of the exercise, choose a topic that is interesting, important, relevant and appropriate.
- You may want to include a bibliography in this section (especially if you are looking up facts/background information regarding the topic).
- Be specific. A good principle is to assume that your audience is not aware of the subject matter.

Part 2

Description:

In this question you will create a "Data" Section of a report. This will consist of an exploratory data analysis, where you will summarize meaningful aspects about the data. Please find some data through either:

1. The Toronto Open Data Portal (https://open.toronto.ca/) and download it via the opendatatoronto R package (https://sharlagelfand.github.io/opendatatoronto/).
2. Some survey data from the 2019 Canadian Election Study (http://www.ces-eec.ca/) and download it via the cesR R package (https://hodgettsp.github.io/cesR/).

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

Your data section should include the following:

- A description of the data collection process.
- A summary of the cleaning process (if you cleaned the data).
- 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 asthetically pleasing plot/graph/figure (No more than 3 plots).
- Text explaining/highlighting each table or figure.
- In line referencing/text if needed.
- Reference the programming language/software used to complete this section.

General Notes (for Part 2):

- All tables/figures should be well labelled and clean.
- Everything in Part 2 should be written in full sentences/paragraphs.
- There should be no evidence that Part 2 is an assignment, I should be able to take a screenshot of this section and be able to paste it into a newspaper/blog.
- There should be no raw code; output should be nicely formatted.
- You will also need a reference 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.
- Use full sentences.
- 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.
- Be specific. Remember, you are selecting this 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 with a conclusion. This means reiterating the key points from your writing.