Assignment 1

Statistical Thinking 2023

Due 4:30pm Monday 9th October 2023.

Instructions

This assignment is a group assignment. Only one submission for each group is required. Your Group Number, and all group member names and ID numbers must be stated in the YAML section of the Rmarkdown file and on any other files submitted.

Write your answers in the RMarkdown file so that it compiles to produce the desired results.

You will need to upload **two (2)** separate files to the link on Moodle for your group. These files will have names such as:

1. **GroupNumber_A1.Rmd**
2. **GroupNumber_A1.pdf**

This assignment has small sub-tasks which should be answered using small paragraphs. Think of these as mini-reports. Your written answers should be non-technical. You must make recommendations and support these with evidence.

Show and evaluate all code chunks using the 'echo=TRUE' and 'eval=TRUE' chunk options, and format the output so that it does not run off the page when printed. You can also suppress all other messages and warnings - as per the following command (to be included in the first code chunk of your .Rmd file):

```
knitr::opts_chunk$set(echo = TRUE, eval = TRUE, warning = FALSE,
    message = FALSE, error = FALSE, tidy.opts = list(width.cutoff = 60),
    tidy = TRUE)
```

Anything that is not part of your answer should not be included in the .Rmd, even if it is not evaluated or does not appear in the rendered file.

Assessment marks

There are three (3) broad tasks, each with several sub-tasks with mark allocations provided.

The final mark for this assessment task will be based on
i. The completeness and clarity of your analysis for each sub-task (note the amount of marks allocated to each sub-task);
ii. Overall presentation [**5 marks**] and;
iii. The ability of the submitted .Rmd files to compile immediately as submitted and without error (reproducibility) [**2 marks**].

Peer Evaluation

You will also complete a peer evaluation survey (Teammates) to rate the contribution of each team member.

- You will be penalised for poor contribution scores.
- Scores will range between 1 (poor) to 4 (excellent).
- This survey will be opened on October 6th
- The survey will close on **October 12th.**
- **If you do not fill out the survey, you are assumed to give your group members a score of 3 and you will be penalised 10% of your group grade.**
- The survey is external and the due date cannot be extended.

General comments

- All Groups are to do all questions in this Assignment, regardless of your unit code.
- Refer to the **R Markdown Quick Reference** available under the **RStudio** *Help* menu (located at the very top of the **RStudio** environment) for help with formatting your report. Note the section on *LaTeX Equations* for how to insert mathematical symbols into your document. A useful resource for the LaTeX math symbols can be found at https://www.caam.rice.edu/~heinken/latex/symbols.pdf.
- You may discuss your ideas about the Assignment with your classmates, including on the Discussion Forum. Groups may even work together, however every group must complete their own submission which should accurately reflect the work and efforts of all students in your Group. Any questions about the Assignment should be asked in consultation or via the student Forum in Moodle.
- Experience suggests it is a poor strategy for a group to allocate questions to individuals and then just to present them together as if they were each undertaken collectively by the group. *Some tasks that you complete in this assignment will appear on the exam.*

The best strategy is to have each group member attempt each question on their own, with the group meeting together after a few days to review and decide whether the questions can be answered directly, or if additional work is required. Be sure to leave enough time to combine your efforts and review the final report before submitting the assignment.

- Organise your submission carefully with well defined headings. Separate each section using a sub-sub-section heading (three hashtags) in your document (i.e., ###) to ensure structure to your report.
- All graphs must be properly labelled

Introduction

A lecturer in management I met recently devised a new quiz which their class recently completed. The lecturer was hoping that the average mark would be around 70% with most students within about 10% of that grade. She attempted to make the quiz challenging, but not too difficult, so a "Pass" would be about 60% with only a few students getting a grade below this.

The lecturer has some understanding of descriptive statistics, but has no idea as to how to approach modelling the distribution of quiz grades to check if the quiz went to plan. She has also heard about randomisation and Bayesian techniques and would prefer using these rather than relying upon the Central Limit Theorem.

Luckily for the lecturer, I know that you can help. There are several problems she wishes you to report on. I have made a list of tasks for you to complete, however each requires a non-technical summary.

The Dataset

The lecturer has supplied a dataset of grades named *GradesData*. This has quiz scores and cohort information on 200 students. Be sure to familiarise yourself with the data before you analyse it.

Instructions

Task 1 - fit a distribution [30 marks]
The lecturer wants to examine "genuine" (non-zero) attempts only. Check and modify the data (if necessary) to ensure that only genuine attempts are analysed. **[1 mark]**

- Traditionally Normal distributions have been used to model grade distributions. Plot your data and explain whether you think that this is good idea in this instance. **[3 marks]**
- An alternative is the beta distribution. Explain why this may be a valid alternative in this case. **[2 marks]**
- Use Maximum Likelihood to fit both a Normal and Beta distribution to the grades. Use a Bootstrap QQplot to assess the fit of each. Which distribution do you recommend? Be sure to briefly explain your reasoning. **[8 marks]**
- The lecturer is interested in trying the Beta distribution. Use your MLE's to report the mean and the median of the grade distribution (recall that the mean is a function of the shape and rate parameters). Be sure to interpret these values. **[4 marks]**
- Plot and interpret a 99% parametric bootstrap of the mean of the beta distribution (*HINT: set warning=FALSE in your code chunk*). Did the average quiz mark match the lecturer's goal? **[4 marks]**
- Using the MLE's, what is the estimated proportion of students within 15% of the average? According to the lecturer's benchmark, what proportion would have failed? How many would get HD's? **[4 marks]**
- Overall, do you think that the quiz achieved the lecturer's aims? **[4 marks]**

Task 2 - Are Postgrad students better? [30 marks]
Similar to our unit, the lecturer has a postgraduate (PG) and an undergraduate (UG) cohort. Using the lecturer's definition of a pass, we are interested in testing if the proportion of postgraduate students who "passed" the quiz is different from the proportion of undergraduates who "passed" the quiz.

- Create a new variable "Result" which contains the values "Pass" or "Fail" depending on the value of the variable `Total`, and add it to your dataframe. (**Hint: If you have the MASS package loaded, the dplyr command `select` will not work. Add `dplyr::`before any `select` command in your code.**) **[4 marks]**
- Report a table with the relevant proportions and discuss the result. **[10 marks]**
- Conduct a permutation test with 5000 replications. Plot the sampling distribution of your test. **[4 marks]**
- What is the result of this test? (Be sure to properly define the null and alternative hypothesis and discuss the result of the test). **[8 marks]**
- The lecturer may be confused as to why such a large difference in proportions does not have a p-value of almost 0. Explain why this may occur. **[4 marks]**

Task 3 - Bayesian Analysis [33 marks]
The lecturer is curious about using Bayesian analysis to compare the proportion of postgraduate students who "passed" with undergraduate students who "passed". The lecturer has no real opinion as to what the distribution of marks for each cohort would be.

- Suggest (and justify) an appropriate prior distribution for the proportion of postgraduate students who "passed" and one for the proportion of undergraduate students who "passed". **[2 marks]**
- State the posterior distributions for the proportion of postgraduate students who "passed" and one for the proportion of undergraduate students who "passed". **[2 marks]**
- Obtain 95% credible intervals for the proportion postgraduate students who "passed" and for the proportion of undergraduate students who "passed". Comment on how these two intervals compare to each other. **[4 marks]**
- Obtain and interpret the estimator that minimises the posterior expected squared error loss for the proportion of students who "passed" for a given cohort. Is this in a form that linearly combines a purely data-based estimator and some prior quantity? If so, quantify the contribution of the data-based estimator by deriving the credibility factor. Using the data, calculate the estimate and credibility factor for both cohorts of students. **[8 marks]**
- State the estimator that minimises the posterior expected absolute error loss for the proportion of students who "passed" for a given cohort. Using the data, calculate the estimate for both cohorts of students. **[4 marks]**
- For each cohort of students, visualise the posterior distribution the proportion of students who "passed". Discuss any interesting features. **[3 marks]**
- The lecturer wishes to do this again next semester. What priors do you suggest that she use? Explain your reasoning. **[2 marks]**

The lecturer wants to analyse the difference in proportions of "passing" for the two cohorts.

- Simulate the posterior distribution of the difference in the proportion of postgraduate students who "passed" and the proportion of undergraduate students who "passed". **[4 marks]**
- Give the lecturer an indication of the probability that the difference in the grades of the 2 cohorts is within 0.15. Be sure to provide a visualisation of the posterior to support your indication. **[4 marks]**

Conclusion

Remember that presentation is important, as is conveying your results/recommendations/findings in a succinct and informative way. Be sure to include only what you think is important to the lecturer - she does not want a textbook. She just wishes to know what you did, why you did it, and what it means.
Good luck!