STAT100 Written Assignment 1

Due: 11:55pm, Sunday 7 August 2022

Overview

Influenza ('the flu') is a respiratory virus that circulates globally each year. There are a number of strains and it evolves rapidly, so new vaccines are produced and distributed each year. In this assignment you will explore some data from a historical influenza study, and propose a design for a hypothetical future study.

Some useful science knowledge

One way that we study influenza is by measuring how your immune system responds to the virus. If we take a sample of your blood, we can test diluted samples of the virus against that blood sample, and the final dilution that reacts is recorded as the 'titre' (sometimes spelled 'titer'). Because of how titre is measured, it typically takes very discrete values like 2, 4, 8, 16, … or 5, 10, 20, 40, 80, …, with a higher titre indicating that your immune system responds more strongly to the virus. Typically (and in this study) you take before and after samples and look at the change in titre (as a ratio, i.e., after/before). So, if you are infected with the virus between the before and after samples, your immune system will learn to respond to it, and the titre will be higher: we generally say that if the "after" titre is 4x or more the "before" titre it is evidence that you may have been infected (this value is indicative only; you should not use it in your analysis).

You will be provided with a dataset that contains the following columns:

- subjectID
- age
- sex
- blood_date_1
- blood_date_2
- vaccination_status
- fold_change_titre

Each row is a different person, with blood samples taken on two separate dates. Fold_change_titre holds the base-2 log of the ratio of the two titres from those two samples, i.e.,

- if it is 0, they have the same titre,
- if it is 1, the second titre is double the first titre,
- if it is 2, the second titre is 4x the first titre,
- if it is -3, the second titre is 1/8 of the first titre,
- and so on.

Your data are a subset of the data collected for this study:

Wei, V. W., Wong, J. Y., Perera, R. A., Kwok, K. O., Fang, V. J., Barr, I. G., ... & Cowling, B. J. (2018). Incidence of influenza A (H3N2) virus infections in Hong Kong in a longitudinal sero-epidemiological study, 2009-2015. *PLoS One*, *13*(5), e0197504.

(you don't have to read it)

Obtain your dataset by selecting the "Get your assignment 1 data here" entry in the Assessments section of the STAT100 moodle site, and following the directions to download a feedback file (that file is your data).

You will explore and analyse the data using R, and present your findings in a written report. You must present the following 3 sections:

Section 1: Preliminaries (20 marks)

Your task is to determine if the vaccine was effective in reducing influenza infection in the community during the year in which your data were sampled. Based on your understanding of this data and the useful science knowledge above, describe:

1. The response variable, including what type of variable it is (i.e., nominal, ordinal, continuous or discrete; include a justification for your answer)
2. The explanatory variable, including what type of variable it is (i.e., nominal, ordinal, continuous or discrete; include a justification for your answer)
3. The research question, stated in terms of the response and explanatory variables (your research question should include the word 'difference')
4. The Null and Alternative hypotheses.
5. Describe the other variables that are present in the data and the type of variables they are.

Section 2: Data analysis (50 marks)

1. Produce a table or tables presenting appropriate sample statistics for age, sex, fold_change_titre, and vaccination status. The values you enter in this table should be computed in Rstudio, but you should enter them in your word processing software with appropriate formatting (i.e., the output should not be a screenshot or a cut-and-paste). You will also need to supply the code that you used to compute these summary statistics.

2. Produce a professional-looking, coloured, well-labelled boxplot that presents the key pattern in this data pertaining to your research question.

3. Describe in your own words the pattern that you observe in this plot. You might like to describe aspects such as relative scatter, amount of variability, differences between groups, symmetry, outliers, and general trends. Then, describe what this plot suggests to you about the research question.

4. Produce one additional graphic (again professional-looking, coloured and well-labelled) that shows a pattern in the data that you believe to be interesting or worth possible further investigation. This figure should not be a boxplot. You may consider including the age or sex variables, or even the dates the samples were taken. Include the R code you used to generate this plot.

5. Describe in your own words the pattern that you observe in this plot, and why you believe it is interesting or worth further investigation.

6. Perform a two-sample t-test (without assuming equal variance) to assess the outcome of your research question. Include the R code you used to perform this test, and the paste the full R output in your report.

7. Provide an interpretation of the output of your t-test in the context of the research question. Include specific references to the p-value and confidence intervals. Refer to your figure to provide additional context.

8. Provide a brief description of **one possible confounding variable** with an explanation of why you believe that the given variable may have had an impact on your data.

Section 3: future proposal (20 marks)

You are tasked with designing a study that will assess the efficacy of the influenza vaccine in a future season (say, in 2023). From what you have learned in the first part of this assignment, describe in detail how you would like to design this study. You should include how many participants you would like to include, how they should be selected, and what information you wish to record for each participant. Justify every decision you make.

Presentation (10 marks)

Produce a report in your choice of word processing software (e.g., Microsoft Word or Google Docs) that includes all of the above information. Ensure that your report includes your name, is logically structured, and the sections and questions within those sections are clearly indicated with headings. Include all the R code that you wrote, in the relevant sections (e.g., include the code that produced a plot below that plot) – please show all code in a separate font (a fixed width font like `Courier New` is ideal). Save your report as a single pdf file, and submit that file on moodle.