Applied Statistical Inference Coursework 2025-26

Chris Hallsworth

2024-11-14

Background

This task concerns data from the International Stroke Trial, a large randomized open trial to evaluate the safety and efficacy of two different therapies, heparin and aspirin. The study followed 19 435 patients with suspected acute ischaemic stroke entering 467 hospitals in 36 countries. Patients were randomised within 48 hours of symptom onset to one of six different treatment regimes.

You are not expected to read the article describing the study, however you may find it useful for understanding what the variables mean, and checking that any numerical answers are of the right order of magnitude.

We will focus the evidence for an effect of aspirin on the probability of death within 6 months of randomization.

Suggested Workflow

The workflow below guides you through the task. The deliverable that will be marked is the report. You should also submit code structured according to the points in the workflow, as an appendix. There is no need to include any code within the body of your report. You are welcome to use existing R functions as you need them.

Reading the data

The data file is available on Blackboard and can be read in to R as follows.

```
library(tidyverse)
IST_data <- read_csv("IST-data.CSV")
```

The meaning of the variables can be found in Table 2 of this article. Many of the variables will not feature in this analysis.

Exploratory Data Analysis

Produce plots and tables summarizing the data at baseline, by aspirin treatment group. Variables you might want to consider are age, sex, systolic blood pressure and conscious level.

Generalized linear model to predict death at 6 months

1. Compute the proportion of deaths at six months for each distinct combination of the variables month of randomization, age and country. Investigate the relationship between the the sample mean and variance of the proportion of deaths when you average over month of randomization.
2. Fit a binomial generalized linear model with logistic link function, taking deaths at six months as the response variable and age, sex, systolic blood pressure, conscious state at randomization and aspirin treatment as covariates. Interpret the point estimates of the effect of age, blood pressure and conscious state on the odds of death at six months.
3. Use bootstrap to obtain approximate 95% confidence intervals for the parameters in your model. How well do they match the confidence intervals obtained by assuming the estimators are normally distributed?
4. Illustrate the output of your model e.g. by plotting the predicted effect of age within different covariate groups.

Power analysis

The variable EXPD6 gives a predicted probability of death at six months from randomization.

1. Evaluate whether the predicted probability is well-calibrated, i.e. for $x \in (0, 1)$, is the proportion of individuals with a predicted probability of death in $(x, x + \delta)$ close to x?
2. Verify that the distribution of the predicted probability of death does not differ between treatment groups.
3. Conduct a simulation study to evaluate the power of the trial to detect a given treatment effect, as follows.

- Choose a range of treatment effects θ.
- Transform the predicted probability to get a predicted odds of death and, for each value of θ, multiply the predicted odds of death by θ for each patient in the aspirin group and transform back to get a modified probability of death.
- For each value of θ, perform many simulations:
  - For each patient, simulate whether or not death occurs by drawing a binary random variable, using the modified death probability,
  - Fit a binomial GLM to the simulated data,
  - Determine whether or not the treatment effect is significant.
- Estimate the power of the trial for a given θ as the proportion of simulations where the treatment effect is significant at 5%.
- Make a plot of power as a function of treatment effect.

Deliverable

Write a short report aimed at a statistician who is not familiar with the data but is interested is understanding how the trial evaluated the evidence for an effect of aspirin on the probability of death at six months. Your submission should include:

- A brief introduction to the data, including tables and plots for any variables of interest. [5 marks]

- A discussion of the effect of age, sex and treatment with aspirin on the odds of death within six months. Give estimates, with confidence intervals. [5 marks]
- A discussion of the power of the trial to detect a treatment effect. [5 marks]
- A separate summary aimed at the clinicians who collected the data. This should explain your main conclusions in plain language, in the context of the original data. Comment on strengths of the design and provide recommendations for how the trial might be improved. [5 marks]

Specifications

Your report should be clearly written, with plots that are easy to read. Axes should be labelled, with units where appropriate. Numerical values in the text and tables should be appropriately rounded.

When evaluating your work, markers will bear in mind the following aspects:

1. **Completeness**: has the task specification been met? Roughly, in the language of Imperial exam mark schemes, these are most straightforward 'A' marks.
2. **Accuracy**: are the tools from the module used correctly, and are the conclusions drawn reasonable? Accurate work is a fundamental requirement; these are marks for solving straightforward problems, i.e. B marks.
3. **Clarity and Accessibility**: are the conclusions described clearly, in plain language, appropriate to the audience? This is a higher level skill, and corresponds to the C marks in an exam.
4. **Reproducibility**: could your analysis be repeated according to the summary you have provided? Where decisions are made, e.g. in removing suspect data points, are these justified and their consequences considered? This is a higher-level skill, which requires good insight and independent thought. These marks roughly correspond to D marks.

Your report should not exceed 5 sides of A4. It does not need to follow the suggested workflow explicitly. It should be 10 point font or larger, with normal margins. Include complete and carefully commented R code as an appendix. The appendix does not count towards the page limit. The report should be submitted as a single PDF file on Blackboard.

Submission

Please submit your report through TurnItIn on Blackboard. The deadline is **13:00 UK time Friday 28 November**.

Academic integrity

You are welcome to make use of any sources, so long as you cite the sources that you have used. You are welcome to discuss this work with other students on the module, but you should write your own submission, including all code.

It is reasonable to use large language models to make routine coding tasks easier, to the extent that they are not the focus of what you are meant to be learning. However, responsibility for errors in submitted work is entirely your own. Any use of large language models must be clearly and fully acknowledged. This is a judgment call - the gold standard would be a reflective paragraph on how you used large language models, what errors you identified in their response etc, with prompts and responses included as an appendix.

Mastery material (Year 4 only)

By design this is an open-ended task. You are recommended to structure your response as a short report (1 A4 side is fine; at most 2) for another statistician to read. Include carefully commented code as an appendix.

The task is based on this study of nerinetide for the treatment of stroke. You do not need access to data for this task; it is based on simulation.

This is a more modern trial design that incorporates an *interim analysis* - this means that the study statisticians evaluate the data part-way through the trial, but according to a more stringent significance threshold, such that the overall probability of a false positive is well-controlled.

Your task is to write a report containing the following information:

1. A brief overview of what the trial aimed to achieve.
2. An explanation of how the investigators may have determined the sample size (see the Statistical Analysis section of the article). The R function `power.prop.test` from the `stats` library may be helpful.
3. An explanation of the idea of 'alpha spending' for an interim analysis for efficacy.
4. A discussion, based on a simulation study, of the power of the trial to detect treatment effects of different sizes. What is the conditional power, given an unsuccessful interim analysis?

This section will count for 25% of the available marks.