Practical 13: Planning a Data Analysis Pipeline

ADS2

Semester 1, 2023/24

Scenario

Your collaborator Dr. Hu has approached you: They would like your help with data analysis for a clinical study. The study is about a new surgical procedure for hip replacement. The question is whether the new procedure improves recovery post-operation and specifically, whether it reduces pain.

In order to assess this, the doctors at Dr. Hu's hospital collect information from patients at two time points: Once the day after the surgery (baseline), and once during a checkup 3 weeks later.

Pain is measured by asking patients to rate their pain on the 11-point NRS-11 scale (https://www.painconsortium.nih.gov/pain_scales/NumericRatingScale.pdf). This assigns numerical values to the intensity of pain as follows:

Rating	Pain_Level
0	No Pain
1-3	Mild Pain (nagging, annoying, interfering little with activities of daily life)
4-6	Moderate Pain (interferes significantly with activities of daily life)
7-10	Severe Pain (disabling; unable to perform activities of daily life)

Dr. Hu tells you that the average reported pain immediately after hip replacement surgery is 6, with a standard deviation of 1. After 3 weeks, they would expect pain to be done to around 2 or 3, but it is very dependent on the patient.

Aside from the pain rating, the following information is collected from each patient:

- the patient's name
- the time point (immediately post-op or 3-week check-up)
- the patient's phone number
- the patient's gender
- the patient's home province

Dr. Hu's plan is to generate a spreadsheet with all pain assessments that were done in one week at the end of the week and send it to you. The time frame of the study is a year, so you can expect to receive around 50 .csv files.

Your task

- Create a synthetic dataset that you can use to test your data analysis pipeline (note that although you will have 50 .csv files in your final study, you don't need that many to test your pipeline. But it may make sense to have several .csv files, just to see how your pipeline handles this in principle.)
- Do you have enough information to create your synthetic dataset, or are there things you need to check with Dr. Hu?

- Plan out a data analysis pipeline. You don't have to code it all (that would be way too much work for a practical), but you can either sketch one, or write some bullet points in an R Markdown document - whatever helps you get an overview of the entire process and what needs to be done at each step.
- Think in particular about the following questions:
  - How will you handle having the data in several spreadsheets?
  - How can you protect a patient's privacy, while also making sure that you can match a person's pain level at their 3-week checkup to their original pain level?
  - How are you going to deal with ambiguous or missing data? What kinds of data problems are likely to occur?
  - Of all the information recorded in the spreadsheet, what information does the data analyst need? What information should be hidden from the data analyst? Should somebody else have that information?
  - What would be a good way of visualising the data?
  - What kind of data analysis will you likely be using?
  - Are there any other problems you notice or things you may need to pay attention to?

Last update by DJ MacGregor in 2023