Research Methods in AI – Assignment 1

Background

This assignment is based on data simulation. By simulating our own data, we can "know the truth". For example, if we simulate IQ scores for two groups of people, both with a mean IQ score of 100, we know there is no real difference in mean IQ between the two groups in our theoretical populations. However, a statistical test performed on this data can still give a significant result, because our data (sample) is random. In that case, we know that rejecting this null-hypothesis is a type 1 error: we say there is an effect, but in reality no such effect exists. The data was constructed in such a way that the null-hypothesis is true, but we reject it by coincidence when we do our significance test.

Frequentist significance testing is based on the p-value: the probability of finding a sample (statistic) as extreme as or more extreme than the one we have found, if (!) in reality the null-hypothesis is true. If this probability is low, we reason that our null-hypothesis is probably not true (since we know that our data is "true", because we have observed it). So when we do a test, we collect data and then see if that data seems plausible if "nothing is going on", i.e. the null-hypothesis is true. If the data is not plausible, the assumption that the null-hypothesis is true must have been wrong. Our threshold for "not plausible" is finding a p-value lower than alpha (often 0.05).

In the lectures and workgroup exercises, you have seen what the distribution of the p-value looks like, if in reality the null-hypothesis is true. In this case, for some proportion (alpha) of cases, the p-value will be below our significance threshold (that same alpha). In those cases, we commit a type 1 error. Researchers can make choices that inflate the proportion of type 1 errors, so it is no longer equal to alpha. This is problematic, because it leads to more published false positive findings.

In this assignment, you will simulate how this happens. You simulate a (large) number of datasets, for which we know that the null-hypothesis is true. On each of these datasets you perform the appropriate statistical test (for example an independent samples t-test), but you do so while committing questionable research practices (QRPs). In the lecture of week 3, we have discussed several of these QRPs and you can choose which ones you want to simulate.

The goal of the assignment is to show how committing a QRP impacts the type 1 error rate.

What to do

- Come up with a scientific study that requires performing one of the statistical tests discussed in this course. This can be based on a real study, or be imaginary, but in your simulations you have to make sure that the null-hypothesis of interest is true.
- Choose one or a few questionable research practice(s) that the researchers could use in this study. You can choose one or more from this list:
    - Sequential testing with optional stopping
    - Removing outliers with different criteria, depending on the results
    - Using multiple dependent variables and reporting only those giving desirable results
    - Reporting on specific levels (groups) of a nominal independent variable, depending on the results
    - Removing covariates (additional independent variables) or adding them to the model to get a lower p-value for the main independent variable
    - Rounding p-values down (e.g., p of .056 becomes $p \leq .05$)

- Simulate **a large number** of the studies in R, as if the researchers repeat their data collection and analysis many times (or in many parallel universes). Each time:
    - Make a fake dataset in line with the null-hypothesis.
    - Perform the appropriate statistical test.
    - Make the researchers commit the QRP(s) of your choice.
    - Keep track of the statistical conclusions (p-value and decision on rejection).
- Summarise and describe the results of the simulation in your report. What is the effect of conducting the QRP(s)?

What to hand in

Please hand in:

- The R code for your simulations (*.R* file), with comments to explain what the different parts of the code do (# like this).
- A written report of 1000-1500 words (*.pdf* file), containing:
    - Introduction: what is your (imaginary) scientific study about and what is the QRP that your researchers commit?
    - Methods: what did you do in the simulation? This should help us read your R code.
    - Results and discussion: what are the findings from your simulation? What happens to the p-values and type 1 error rate? Illustrate the outcomes, not only in numbers, but also with suitable graphs.
    - A short *Author contribution* section, stating what each group member has contributed to the project (in writing, analyses, etc.)

Grading and rules

When grading, we will take into account:

- Clarity of the problem description, both study context and QRP(s)
- Description of methods and choosing a suitable statistical test
- Description, presentation and interpretation of the results
- Reproducibility of your analyses (if we run your code, we should get the *exact same* results)
- Code quality and clarity
- Difficulty level of the QRP(s) you picked (rounding down p-values is easiest)
- **Most important**: did you actually simulate type 1 error and correctly implement the QRP(s)?

The educational goal of this assignment is to learn about the effects of QRPs, while getting some experience in conducting statistical tests in R. For that reason, using generative AI, such as ChatGPT, is not allowed in the production of code or text for this assignment. If you use these tools, as a source of inspiration for example, please indicate this in your report. Note that a substantial part of the work for the assignment will be done during the workgroup meeting of week 3.