Open, transparent, and reproducible data analysis

Why and how

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

18 May, 2023

How we usually do data analysis

When carrying out data-analysis, the following is the common *modus operandi*:

- We manually and interactively process, explore, visualise, analyse the data.
- We then manually copy-and-paste or insert results into reports.
- The report, and not the data and code that led to all the results, is then made public.

Problems with the traditional approach

- Working interactively and then manually copying-and-pasting or inserting results into reports is inherently error prone.
- It is also highly inefficient; making even small changes become prohibitively expensive.
- The workflow is not reproducible; the details of the pipeline from raw data to reported results are not recorded.
- The reported results are not transparent; the public views only a carefully selected facade.
- Data and code are separated from the report and remain hidden; data and code are the second class citizens of scientific communication.

Doing open, reproducible, and transparent analysis

The following are some tools[1] that can greatly facilitate open, reproducible, and transparent data analysis:

- *R*: Write code for all steps of the analysis from raw data to final results. Record all code in scripts or code packages.
- *RMarkdown*: Write all reports using a framework for writing reproducible reports.
- *Git*: Use version control software for tracking and managing all versions of the code.
- *Docker*: Run all code in a virtual operating system that contains all the required software stack.
- *GNU Make*: Automate complex workflows using build automation tools.

[1] There are many alternatives and equivalents to each of these tools.