Reproducible research 1

Reproducible research

What is reproducible research?

Reproducible: Could someone else re-do your entire analysis?

- Data available
- All code for cleaning raw data
- All code and software (specific versions, packages) for analysis

The *Methods* section of the future. . .

Why is it important?

Some advantages of making your research reproducible are:

- You can (easily) figure out what you did six months from now.
- You can (easily) make adjustments to code or data, even early in the process, and re-run all analysis.
- When you're ready to publish, you can (easily) do a last double-check of your full analysis, from cleaning the raw data through generating figures and tables for the paper.
- You can pass along or share a project with others.
- You can give useful code examples to people who want to extend your research.

Why is it important?

An example of how reproducibility can help check scientific results (Source: The New York Times—link below):

> Dr. Baggerly and Dr. Coombes found errors almost immediately. Some seemed careless — moving a row or a column over by one in a giant spreadsheet — while others seemed inexplicable. The Duke team shrugged them off as "clerical errors."
>
> And the Duke researchers continued to publish papers on their genomic signatures in prestigious journals. Meanwhile, they started three trials using the work to decide which drugs to give patients.

Read more from:

- The Economist
- The New York Times
- Simply Statistics

Some steps to making research reproducible

- All your raw data should be saved in the project directory. You should have clear documentation on the source of all this data.
- Scripts should be included with all the code used to clean this data into the data set(s) used for final analyses and to create any figures and tables.
- You should include details on the versions of any software used in analysis (for R, this includes the version of R as well as versions of all packages used).
- If possible, there so be no "by hand" steps used in the analysis; instead, all steps should be done using code saved in scripts. For example, you should use a script to clean data, rather than cleaning it by hand in Excel. If any "non-scriptable" steps are unavoidable, you should very clearly document those steps.

Tools for research reproducible

There are several software tools that can help you improve the reproducibility of your research:

- **`knitr`**: Create files that include both your code and text. These can be compiled to create final reports and papers. They keep code within the final file for the report.
- **`knitr` complements**: Create fancier tables and figures within RMarkdown documents. Packages include `tikzDevice`, `animate`, `xtables`, `pander`, and `kableExtra`.
- **`packrat`** and **renv**: Save versions of each package used for the analysis, then load those package versions when code is run again in the future.

Today we will focus on using `knitr` and RMarkdown files.