Reproducible research

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kbroman@rstats.me
Slides: kbroman.org/Talk_JAXsymp

Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce

The results in Table 1 don’t seem to correspond to those in Figure 2.

Where did we get this data file?

Why did I omit those samples?

Which image goes with which experiment?

How did I make that figure?

In what order do I run these scripts?

“Your script is now giving an error.”

“The attached is similar to the code we used.”

Reproducible research

organize the data and code in a way
that you can hand them to someone else
and they can re-run the code
and get the same results
(the same figures and tables)

Steps to reproducible research

- Organize your data and code
- Everything with a script
- Automate the process
- Turn scripts into reproducible reports
- Turn repeated code into functions
- Package functions for reuse
- Use version control
- License your software

kbroman.org/steps2rr

Additional considerations

- Arranging data within files
- Metadata
- Sharing data and code
- Software testing
- Capturing the software environment
- Containers
- Handling large-scale computations
- Coordinating with collaborators

Challenges in collaborations

- Shared vision
- Compromise
- Coordination
- Communication
- Sharing code and data
- Synchronization

Tools

- R, python
- R Markdown, Quarto, Jupyter notebooks
- GNU make, snakemake, targets
- functions and packages
- git and GitHub
- renv, conda
- docker

The most important tool is the mindset,
when starting, that the end product
will be reproducible.

– Keith Baggerly

The most important tool is the mindset,
when starting, that the end product
will be reproducible.

– Keith Baggerly

The second-most important tool is training.

Slides: kbroman.org/Talk_JAXsymp

kbroman.org

github.com/kbroman

@kbroman@rstats.me