NYU HEALTH SCIENCES LIBRARY

Computational Reproducibility

NYU HEALTH SCIENCES LIBRARY

Review

In the context of scholarly literature, what is "bias"?

NYU HEALTH SCIENCES LIBRARY

0

What is error?

NYU HEALTH SCIENCES LIBRARY

0

When publishing literature, how can you demonstrate rigor and make your research more replicable?

NYU HEALTH SCIENCES LIBRARY

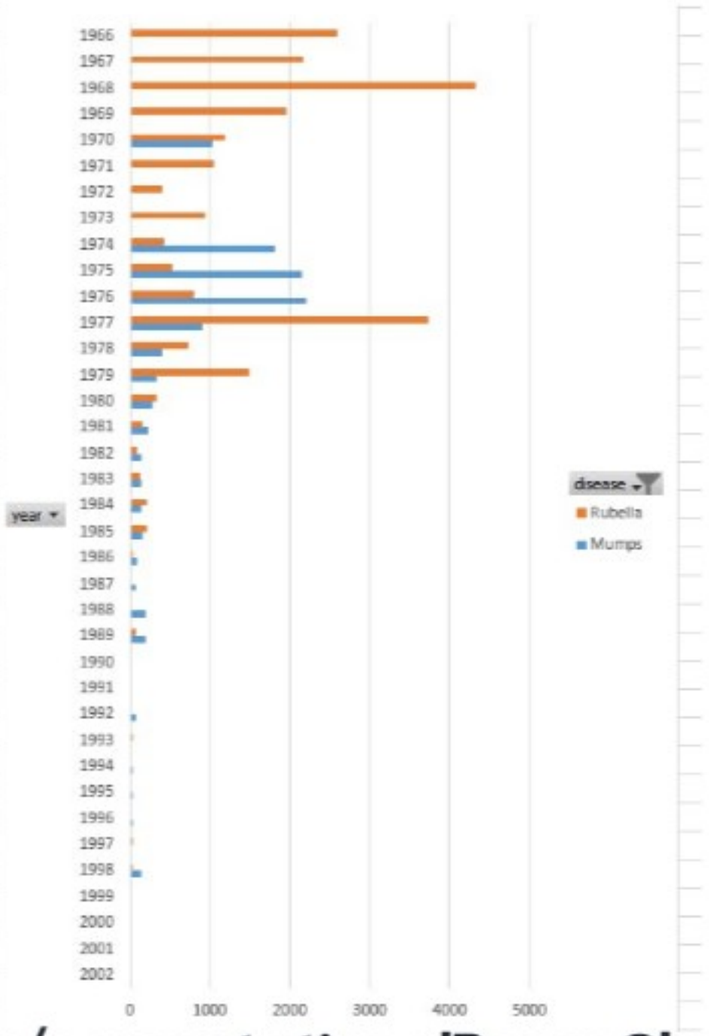
0

Outline

- Case Study with Excel
- Computational Reproducibility Defined
- Intro to R and R Studio
- Intro to R Projects and how they relate to reproducibility

?

- https://raw.githubusercontent.com/fredlapolla/computationalReproClass/master/NYDiseases.csv
- Right click/CTRL Click and save the file to you computer, open it in Excel
- Create a chart of your choosing and document what you did
- Provide a partner with your steps but not your chart, see if it matches



https://github.com/fredlapolla/computationalReproClass/blob/master/NYDiseases.csv

What would have made that easier?

NYU HEALTH SCIENCES LIBRARY

0

What would have made that easier

- Steps to take, in order
- Explanation of what the variables mean
- Instructions (or maybe even automatic instructions)

?

Save your raw data and document steps

?

Save your programs and scripts, and automate your workflow

?

Make it easy to find files

?

NYU HEALTH SCIENCES LIBRARY

Make life easier for collaborators to understand

Including yourself in the future

?

Make it easy to find the files and know the analysis order

?

Track your changes

Explain it clearly in the manuscript

?

NYU HEALTH SCIENCES LIBRARY

R & R Studio

What is R?

R is a free, open source coding language designed for doing data analysis

?

What is RStudio?

RStudio is an "Integrated Development Environment" for R.

?

What does that actually mean?

R is a "language" for giving instructions to the computer on what to do with a file of data.

|

|

R Studio is a program that makes R easier to work with, because you can save commands and see files.

?

Why R and R Studio?

- Free
- Open source
- Makes your analysis more reproducible
- You can run the same analysis on new sets of data

Concerns?

NYU HEALTH SCIENCES LIBRARY

0

Who has analysis experience?

NYU HEALTH SCIENCES LIBRARY

0

What tools have you used for analysis?

NYU HEALTH SCIENCES LIBRARY

0

Analyzing your data in R let's you:

- → Leave raw data unchanged
- → Document your analysis steps in your code
- → Automate analysis and save it for later
- → Share with other R users
- → Explicitly state which files you are using for analysis
- → Present your order of operations
- → Maintain a version file (if you want)
- → Publish code

?

NYU HEALTH SCIENCES LIBRARY

R Vocabulary

R Script

A set of instructions that tells R what to do.

A script can be saved for later.

The same set of instructions can be run on new data.

?

NYU HEALTH SCIENCES LIBRARY

R Console

An area for running code that does not save.

Good for trying out code to see if it works.

?

Environment

A "place" where "objects" are stored. This can mean a data-frame (think: spreadsheet), a list of numbers/words etc, and functions (sets of instructions for R).

?

R Projects

- A file that lets you store analysis code, data files, a history file and versioning tools all in one place
- Automatically keeps everything together
- Makes it easier to automate by pointing directly to data in the project folder

Let's try it!

On your own, create a new R Project in a new directory. A directory is just a file folder. I recommend doing this in a file folder on your computer rather than in Google Drive, as sometimes working in Drive will give you repeated annoying warnings about access. Save this somewhere that will make sense for you and name it something that will make sense later, like compReproClass.

NYU HEALTH SCIENCES LIBRARY

R & GitHub

Git is a tool for tracking changes

Version control is basically like the "track changes" feature in Word.
Git is a tool for tracking changes, which you *can* use in the "command line."
If you are interested in learning about this:
*https://www.codecademy.com/learn/learn-the-command-line
*https://www.datacamp.com/courses/introduction-to-git-for-data-science

?

GitHub is a Website

Lets you store and share code, track changes

?

GitHub

You can create a "repository" which is a place where code and files are stored for access later.

?

Linking Projects to GitHub

You can link your project to a GitHub repository by setting up a new project and selecting Version Control.

|

You can then provide the URL of a GitHub page. This lets you store a project online, update it and track changes. You can also work with others projects this way as well. To save your new versions you must save locally and then commit the changes.

For demo purposes: https://github.com/fredwillie/RigorReproducibilityClassNYU