Introduction to reproducible research

Noushin Nabavi & Monica Granados

2020-08-23

Contents

Preface

A repository to house materials for a reproducibility in research workshop using `R`.

The goal of this workshop is to teach concepts of `reproducible research` to new-to-programming professionals.

`Reproducible research` is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them.

The need for reproducibility is increasing dramatically as data analyses become more complex, involving larger datasets and more sophisticated computations.

Reproducibility allows for people to focus on the actual content of a data analysis, rather than on superficial details reported in a written summary. In addition, reproducibility makes an analysis more useful to others because the data and code that actually conducted the analysis are available.

This workshop focuses on the concepts and tools behind reporting modern data analyses in a reproducible manner. As part of this, we introduce tools that enable publishing data analyses in a single document that allows others to easily execute the same analysis to obtain the same results.

Additionally, as part of this workshop, we briefly introduce data structures, importing and wrangling data, exploring missingness, as well as data visualizations and reporting.

`R` is a popular statistical computing language, commonly used in many scientific disciplines for statistical analysis, generating production-quality graphics, and automating data workflow tasks. The workshop content will follow best practices for using `R`, giving attendees a foundation in the fundamentals of R and scientific computing.

Chapter 1

Introduction to reproducible research

- what is reproducible research?
- components of reproducible research

1.1   Motivation

Chapter 2

Tools for reproducible projects

- tools to set up a reproducible project
- the required tools: e.g. R, Rstudio, python, open-refine, good tables
- notebooks (Rmarkdown, Jupyter)
- demo

Chapter 3

Reproducible research projects

- Tidy data principles
- tidyr and dplyr
- what are data structures?
- what are data frames?
- tidy principles
- missingness
- demo with data

Chapter 4

Introduction to markdown

- how does rmarkdown work? Advantages, disadvantages
- rmarkdown syntax (examples)

Chapter 5

Rmarkdown and usage

- reproducible reporting with rmarkdown

Chapter 6

markdown capabilities

- plots and graphs
- reproducible code to visualize data
- ggplot library in R
- demo with data

Chapter 7

demo

- demo one example

Chapter 8

Reproducible research for github

Chapter 9

Bibliography

Chapter 10

Resources

- Reproducible research with R and RStudio: http://christophergandrud.github.io/RepResR-RStudio/
- Tools for reproducible research: https://kbroman.org/Tools4RR/
- Data privacy and security: https://dataprivacymanager.net/security-vs-privacy/
- BC-Gov framework for github https://github.com/bcgov/BC-Policy-Framework-For-GitHub
- Making slides with Xaringan package in RMarkdown: https://arm.rbind.io/slides/xaringan.html
- Data wrangling with R: https://cengel.github.io/R-data-wrangling/
- Data cleaning with R and tidyverse: https://towardsdatascience.com/data-cleaning-with-r-and-the-tidyverse-detecting-missing-values-ea23c519bc62
- Gallery of missing data visualization: https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html
- How does R handle missing values: https://stats.idre.ucla.edu/r/faq/how-does-r-handle-missing-values/