Center for Open Science virtual unconference

Collaboratively Developing Guidelines for Analyzing Complex Educational Data Using R

CYNTHIA D'ANGELO, JOSHUA ROSENBERG, AND DANIEL ANDERSON

Who are we?

- **Cynthia D'Angelo, Ph.D.**
  - Assistant Professor at University of Illinois at Urbana-Champaign
  - Twitter: @dapostrophe
- **Joshua Rosenberg, Ph.D.**
  - Assistant Professor at University of Tennessee, Knoxville
  - Twitter: @jrosenberg6432
- **Daniel Anderson, Ph.D.**
  - Research Assistant Professor at University of Oregon
  - Twitter: @datalorax_

Plan for today's session

Introduction and goals

What is complex educational data?

Starter guidelines and tips

Some example data sets and approaches

Breakout sessions to develop and refine

Share and discuss as a group

Logistics

- Our slides and other documents are linked here:
  - https://osf.io/fqcmk/
- Please put questions into the chat
- We will have lots of time for discussion towards the end

Goals of our session

- Present initial guidelines and tips that we have developed
  - use RMarkdown and R Notebooks for individual and collaborative research and sharing and communicating results;
  - structure data in consistent forms to ease analytic work;
  - visualize your complex data to make sense of it;
  - common educational data sets that are available to provide context for your data (i.e., NCES data);
  - document (using R Notebooks) your analytic work (including codes as well as writing explanatory text); and
  - re-use code (including writing functions to speed up analysis)

What is complex data?

What is complex educational data?

- Multiple data sets
  - Different types of data
- Multiple measurements within one dataset
- Huge data sets
- Unstructured data
- Complexity in nuance across sites and/or participants

Common issues with complex data

- Missing or incomplete data
- Missing context
- Incompatible scales
- Multiple levels (i.e., individual and group level data)

Stakeholders

- Multiple stakeholders sometimes: funding agency, researchers, teachers, schools, districts
- Each group might have different perspectives on what is important or what to focus on
- Who voices are heard in the data analysis process?

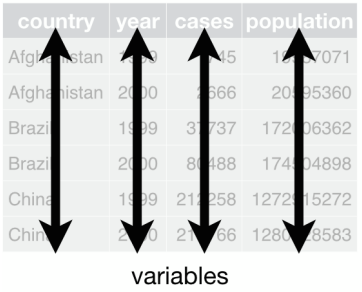
Why R for open science?

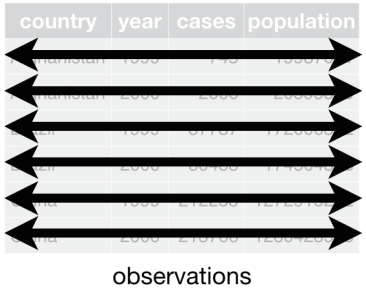
- The Open Science paradigm requires transparency with your data and analysis
- R is open source, free, and accessible
- Huge community of people documenting and helping others
- Continually improving

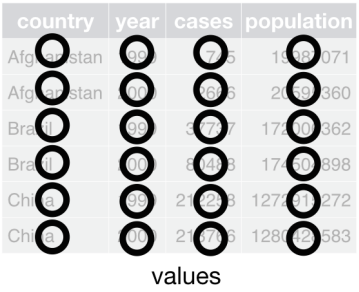
Initial guidelines and tips

Data Structures

- Tidy data paradigm
  - https://www.tidyverse.org
  - https://r4ds.had.co.nz
- xAPI
- Relational databases
- Creating subsets of your data







Data Dictionaries

- Place to describe the meaning of each of your variables
- Context and other info about observations
- Do this even if you think you know what they mean!
- Good way to document for your future self and for collaborators
- Can do within a R Notebook (more on that later)

Data Infrastructure

- Where is your data?
- Data storage issues
- Databases
- FERPA
- Other privacy considerations
- Check with your institution and/or IRB to see what options are available to you

Version Control

- Most important if you're collaborating with others, but also good practice if you're the only one working on your code
- Git
- SVN

Available educational data

- National Center for Education Statistics
  - https://nces.ed.gov
  - ElSi
  - DataLab
- Info about schools from yearly surveys (e.g., size, location, FRPL, demographics)

Common file naming convention

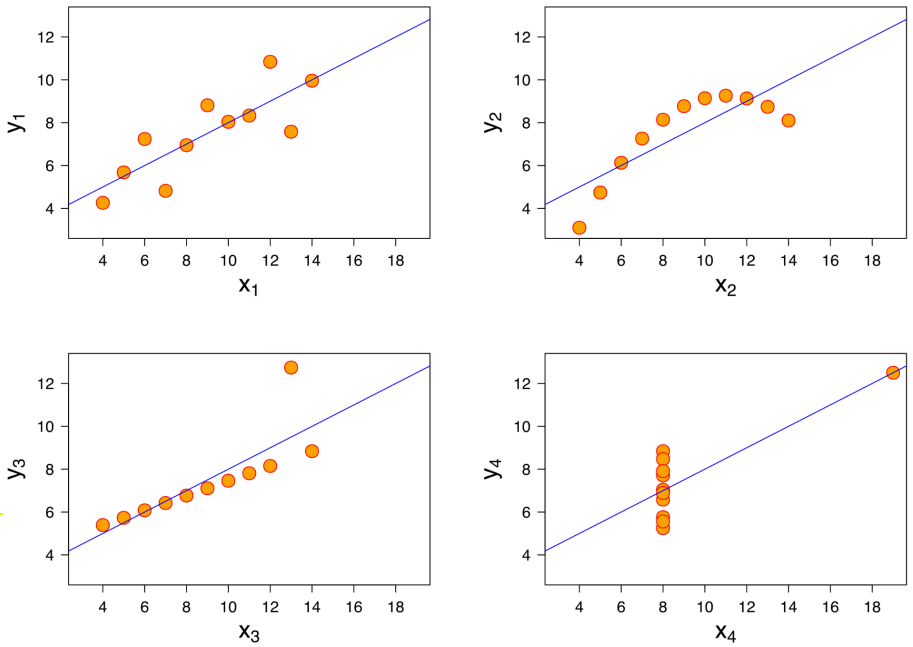
- Do this at the very beginning of your project!
- Document conventions and make sure collaborators adhere to them
- Keep it flexible but also have it make sense
- Think about how the file names will be sorted
  - Date formats -> YYYYMMDD
- Think about unit of analysis you will care about later (e.g., teacher, class, unit, student, group) and build that into your file name
  - Can later subset your data just by filtering for certain parts of the filename

Writing functions

- Are you copying and pasting your code a lot?
- Functions can speed up the process
- Functions can provide more consistency

Visualizing your data

- Important process step, especially for complex data
- Descriptives statistics can be misleading (e.g., Anscombe's quartet)
- Always important to look at your data!
- Visualizing is more difficult with complex data
- Histograms are very helpful
- Facets



RMarkdown & R Notebooks

RMarkdown and R Notebooks

- Probably one of our favorite things about RStudio and the R ecosystem
- Plain text files (work well with version control)
- Produce a html version of your file that is accessible via a browser by anyone, even if they don't have R installed or know anything about R
- Easy to share
- Can mix together explanatory text, code, and figures
- Not limited to R; can also execute code in Python, SQL, JavaScript and others

Important features

- Persistent code execution
- Can see multiple plots at a time
- Interactive tables
- Formatting
- Multiple output options

Output formats: R Notebooks

- Many options available
  - Different kinds of documents (interactive R notebooks, pdf, Word, rich text, etc.)
  - Presentation slide decks (html, pdf, PowerPoint)
  - Interactive dashboards
  - GitHub documents
  - Books
  - Websites

Example data sets

- Online learning systems (e.g., LMSs, online curricula)
- Game-based learning
- Audiovisual data
- Data from collaborative activities
- Social media posts

Collaborative Activities and MMLA

D'ANGELO, C. M., SMITH, J., ALOZIE, N., TSIARTAS, A., RICHEY, C., & BRATT, H. (2019). MAPPING INDIVIDUAL TO GROUP LEVEL COLLABORATION INDICATORS USING SPEECH DATA. 13TH INTERNATIONAL CONFERENCE ON COMPUTER SUPPORTED COLLABORATIVE LEARNING. LYON, FRANCE: INTERNATIONAL SOCIETY OF THE LEARNING SCIENCES.

Talkativeness by group and item

Talkativeness by group and item

What are your tips for analyzing complex educational data?

What questions do you have?