

Emphasizing Reproducible Research by Teaching Data Analysis using R

Robin Donatello

Department of Mathematics and Statistics
CSU, Chico

CELT Conference: 2016

October 7th, 2016

- 1 *What?*
 - Participant Learning Outcomes
- 2 *Why?*
 - Analysis Process - in class vs real life
 - Teaching Reproducible Research
 - Why R + R Studio?
- 3 *How?*
 - Analysis Process
- 4 *What Next?*
 - How To Learn R

In an hour I hope you...

- Understand the importance of teaching reproducible research (RR) to undergraduates early in their career.
- Believe that this is an achievable outcome by undergraduates
- Walk away with a proof of concept reproducible data analysis document using R and Markdown.
- Start to think about ways to incorporate RR into your classroom (I can help!!)

Download slides and code files from my GitHub repository:

https://github.com/norcalbiostat/RR_CELT2016

Analysis process - as taught

- Statistical data analysis is a critical component in a scientific education program
- Often taught as a side topic
 - something a scientist does once
 - using some external program like Excel or Minitab
 - then simply writes the results into their manuscript.
 - Boom. Done.

You know it's not that straight forward.

Analysis process - Real life

- Data exploration and analysis often much more involved.
- No small amount of data preparation and sometimes high level computing processes to analyze the data (e.g. Genome Wide Association Studies (GWAS)).
- Analysis can morph and grow, or run into glitches
 - "Oh i'm sorry, that data you have is wrong. Here's the real data (version 5)"
 - "Let's add in data from this other study".
 - Your indicators got reversed: (e.g. control is marked as intervention)

Analysis process - Real life

- Reviewer comments
 - "You should use measure B instead of A. "
 - "How did you account for missing data here?"
 - "Did you look at the relationship between X & Y before making this conclusion?"
- 5 years post-manuscript questions
 - "How did you calculate this specific number?"
 - "Let's revisit this and...."

Teaching Reproducible Research

- Encourages best practices,
- Provides the skills and tools to integrate statistical data analysis into this research pipeline.
- Provides an explicit record of the data management and analysis process.
- Makes collaboration with other researchers as easy as passing them 2 files: report code and data.
 - They compile the report code on their machine using the data provided and produce the **exact** same report document as you did.

Additional references for motivation to conduct RR

- Special articles in Nature that discusses the need to share code and reproducibility in the sciences. <http://www.nature.com/news/reproducibility-1.17552>
- The spreadsheet error and austerity as discussed on The Colbert Report. Moral of this story: ask for the data and question results that look too good to be true. Even students can find serious errors!
<http://thecolbertreport.cc.com/videos/kbgnf0/austerity-s-spreadsheet-error---thomas-herndon>.
- The cancer research scandal at Duke. Simple and not so simple errors, combined with some fraudulent cover-up puts cancer patient's lives at risk. The 60 minutes story:
<https://www.youtube.com/watch?v=eV9dcAGaVU8>

Why R?

- Free (Not a trivial benefit) Gives students a skill they can use outside of school.
- It's one of the fastest growing statistical languages used in the Natural Sciences
- Everything is user-contributed "package" based. Don't reinvent the wheel, Google it!

Why R Studio?

- Integrated Development Environment (IDE), not "just" a GUI
- Integration of R with LaTeX (this presentation), Markdown, C++, javascript, d3, github....
- Specifically - automate the analysis into a single document that can be run with a click of a button.
- Four windows provides organization, syntax highlighting, tab-complete...
- More amazing things that make your workflow smoother and more organized.

This all sounds great, but *HOW* do we do it?

Demonstration time!

The SOUTH (my left, your right) side of the room go to this link
and answer a few questions.

<https://goo.gl/forms/y7V2lvC60fEujjRf1>

Analysis Pipeline

A typical analysis pipeline consists of

- 1 Reading the data into the statistical software program (getting the data)
- 2 Preparing your data for analysis (cleaning and transforming)
- 3 Univariate exploratory analysis including summary tables and visualizations (describing the data)
- 4 Analysis of the research question of interests using an appropriate statistical procedure (e.g. multiple regression, ANOVA, hierarchical modeling or logistic regression)
- 5 Interpreting and reporting the results. (summary table, model results)

Intro to R and Markdown via R Studio

Let's see how that analysis pipeline looks as a reproducible document.

- Code files all available at:

https://github.com/norcalbiostat/RR_CELT2016

COLLECT ALL THE DATA!



NORTH side, your turn!

<https://goo.gl/forms/y7V2lvC60fEujjRf1>

Cleanup

Woo. Fancy!

But but... that output looks kinda crappy.

At the very least, it's not **that** professional looking.

- Draft student reports.

Cleanup

Some nicer examples

- Chico State 2016 Data science Interest Survey
- Pretty Homework: <http://rpubs.com/gin>
- Lecture notes (Linear Regression): <http://rpubs.com/mdlama>

Where can I send my students (or myself) to learn R?

- Google: "How to learn R" – thousands of options!
- Good option: Data Camp (free Intro to R course)
- Campus Option: R Bootcamp - Department of Mathematics and Statistics.
 - 1 unit CR/NC - meets for 15 hours total, typically over the span of 3 weeks.
 - 3 Saturdays from 12-4 / one week over intersession
- Campus Option: Certificate in Data Science
 - Joint program MATH/CSCI
 - Intro to Data Science (CSCI 398 SP17)

Additional Resources to learn R and Markdown

- Coursera: <https://www.coursera.org/learn/reproducible-research>
- Reproducible Research Using RMarkdown and Git through RStudio <https://rpubs.com/marschmi/105639>
- R Markdown tutorial by Roger Peng
<https://www.youtube.com/watch?v=DNS7i2m4sB0>
- R Markdown tutorial by R Studio
<http://rmarkdown.rstudio.com/>
- Workshop for SUNY Geneseo MATH 341 students to learn how to write a professional report using Sweave, LaTeX, and R. https://www.youtube.com/watch?v=CNJ3ygl_xa0
- knitr: Elegant, flexible and fast dynamic report generation with R <http://yihui.name/knitr/>

Any Questions?

Thank you