R and Reproducible Research

Daniel Anderson

A couple caveats

- Much of what I'm going to be discussing represents an ideal that I have only recently begun working towards.
- None of what I will talk about should be taken as a referendum on you or your current practices. However, I hope to convince you that you should be working toward the reproducible research ideal, and that, as a field, we should be moving toward reproducible research being the *minimal standard*.
- I will be focusing on reproducible research with R (obviously). Other options are available but, in my view, none are as clear, comprehensive, and easy to implement as the tools at your disposal through R.

What is reproducible research?

- **Replicability** is the gold standard for research. Ideally, most research would be verified through replication.
- Reproducibility represents a minimal standard, which itself can aid replication (tremendously), by conducting and documenting the research sufficiently that **an independent researcher could reproduce all the results from a study**, provided the data were available

Why should we care?

- Reproducibility as an ethical standard
  - More transparency
  - More potential for results to be verified (and errors found/corrected)
- If your work **is not** reproducible, it is usually not truly replicable.
- If your work **is** reproducible, then others have a "recipe" for replication

Are journal articles research?

- Initially, we may think of journal articles as research, but really the research is everything that went into the article, not the article itself.
- Some (Buckheit & Donoho, 2015) conceive of the article as the "advertisement".
- If all we have is the advertisement, can we really fully understand the steps and decisions made during the research?
    - In large-scale data analysis, the answer is generally "no".

The Duke Reproducibility Crisis

- 2006 investigation into "personalized" medicine.
  - Can genetic makeup be used to identify therapeutic regimes?
- Study claimed success, published in *Nature Medicine*
  - Deemed a "remarkable breakthrough"
  - Named one of the top publications of 2006 by *Discover*
  - Provided hope for cancer patients
- Original research article was followed by others in high-tier high-impact journals
- Other excited researchers asked for their data and code
- Analysis was found to be almost entirely not reproducible

Is this a big deal? Is it isolated?

- Between 2000 and 2010, a *conservative* estimate of 80,000 patients participated in clinical trials based on research that was incorrect, with papers retracted. (Ince, 2011)
- We have an ethical responsibility to be transparent in our process and ensure that any reported findings are reproducible
- Reproducibility does not imply "correctness", it implies transparency

Tangential benefits

Striving toward reproducible research will:

- Make your own code more efficient/easily interpretable
    - Can help with collaboration on a project
- Reduce errors
- Increase efficiency by not having to redo tables and figures with each tweak to a model.

What does the process actually look like?

- Start with a basic text document (not Word, text)
- Use the text document to write your article
- Embed code within the text document that corresponds to your analysis. Note this is not just copying the code in. The code should be live and what you're working with while conducting your research.
- Render the document into a different format (pdf, html, etc.).
    - Select which code (if any) will be displayed
    - Build tables of results and plots to be produced
- Readers can then read the "advertisement", but if they are interested in reproducing your results (maybe because they disagree with you, or they think your results are weird and want to clearly see all the steps you took), they can access the text file that contains the computer code.
- The end result is a single product that has the advertisement and the research process embedded.

Other reasons dynamic documents are useful

Outside of reproducibility, you may want to use R Markdown to:

- Produce slides
    - Just be careful, I have a horror story
- Keep track of your analysis (notes, essentially), even if you end up using something like Word
- Share code with others
- Quickly share results with others
- etc... ideas?

Demo

PDF output

You will also need to install a TeX distribution.

- Macs: MacTeX (http://tug.org/mactex/)

- Windows: MikTeX (http://miktex.org)

Summarizing

- R Markdown is relatively simple and easy to learn (R is the hard part, R Markdown is easy in comparison).
- Tables are probably the most difficult piece.
- Lots of options to get it to do what you want.
- Great for sharing and documenting your work.

but...

- The more you ask from it, the more difficult it will become.
- The *papaja* package is pretty incredible for producing APA output (https://github.com/crsh/papaja)

Final remarks on R Markdown

- Make sure to look at the documentation
    - http://RMarkdown.rstudio.com
    - http://RMarkdown.rstudio.com/authoring_basics.html
    - http://RMarkdown.rstudio.com/authoring_rcodechunks.html
- The more you ask from it, the more complicated it becomes.
- Challenges
    - Word is the industry standard (frustratingly so, to me)
        - Word output is less than ideal
    - Can be difficult when collaborating with others
    - Some journal articles *require* papers submitted in Word
        - Potentially get a pdf to word converter, but still less than ideal
    - Advanced features have a relatively steep learning curve

Take home message

- Fairly straightforward as a method to produce reports/keep track of your analysis ("lab" notes)
- Start small and work your way up; don't get discouraged too easily
- R Notebooks may be a good place to start
- Writing *manuscripts* used to be a substantial challenge but that's starting to change with the *papaja* package.
- I recommend Yihui's book, it's quite good.