Reproducibility

Dr. Richard W. Evans

March 29, 2017

Why Reproducibility?

- What is reproducibility?

- What are problems with lack of reproducibility?

- What are key steps to reproducible research?

Quote:

"Non-reproducible single occurrences are of no significance to science."
(Karl Popper, "The Logic of Scientific Discovery," Hutchinson, 1959.)

Math slide

- Let's do some math

$$\frac{x}{y} = \int 2\sin(\theta)\frac{x^2 + 3x + 5}{42xy}dx \tag{1}$$

Open Source vs. Open Science

Open Source

- Data and code are openly available to public on some platform
- Encourages collaboration on projects
- Different open source licenses

Open Science

- Methods, equations, algorithms are openly available
- NOT necessarily exact code and data availability
- Focus on independent verification

How do we verify results?

- Replication
  - Seems trivial: hit go on someone else's code and data and see if get same result
  - Necessary condition
  - Only learn whether researcher copied results correctly from computer to paper
  - This is more about Open Source

- Verification, validation, and uncertainty quantification (VVUQ)
  - Independent replication and verification of results
  - Emphasis on sensitivity analysis and model uncertainty (UQ)
  - This is more about Open Science
  - *Journal of Verification, Validation and Uncertainty Quantification* (*JVVUQ*)

Casadeball and Fang paper

Casadevall and Fang, "Reproducible Science," *Infection and Immunity*, 78:12, pp. 4972-4975, (2010)

Abstract: The reproducibility of an experimental result is a fundamental assumption in science. Yet, results that are merely confirmatory of previous findings are given low priority and can be difficult to publish. Furthermore, the complex and chaotic nature of biological systems imposes limitations on the replicability of scientific experiments. This essay explores the importance and limits of reproducibility in scientific manuscripts.

Camerer, et al, paper

Camerer, et al, "Promoting Transparency in Social Science Research," *Science* (Magazine), (2014)

- "Social science journals do not mandate adherence to reporting standards or study registration, and few require data-sharing."

- "In this context, researchers have incentives to analyze and present data to make them more 'publishable', even at the expense of accuracy."

- "These practices, coupled with limited accountability for researcher error, have the cumulative effect of producing a distorted body of evidence with too few null effects and many false-positives, exaggerating the effectiveness of programs and policies"

Camerer, et al, paper (continued)

Camerer, et al, "Promoting Transparency in Social Science Research," *Science* (Magazine), (2014)

Proposals

1. Require **disclosure** of key details of data collection and analysis

2. Registration and **pre-analysis plans**

3. **Open** data and materials

American Economic Review policy

- *AER* Data Availability Policy

  - Must submit conflict of interest statement

  - Must submit data, scripts, and README file

  - Experiments require even more reporting

Organizations for Open Research

- Berkeley Institute for Transparency in the Social Sciences (BITSS)

- The Replication Network

- *Journal of Verification, Validation and Uncertainty Quantification* (*JVVUQ*)

- American Psychological Association, *Journal of Experimental Psychology* replication articles

- Harvard Political Science Replication Initiative

- Open Source Policy Center (OSPC)

Make your code and data open source

- Use version control and transparency of `Git` and GitHub
- GitHub allows open access to contributions and attribution
- Your GitHub page can become a central part of your CV/resume
- Open source through GitHub can encourage collaboration, professional network
- Open source code requires great docstrings, commenting, efficiency, and modularity
- If you want people to contribute to your code, you need to offer something in return
    - Contribute to their code
    - Offer collaboration, coauthorship, consulting

Work arounds for proprietary data

- Synthetic datasets

- John Abowd (Census)

- IRS/OSPC Public Use File-Current Population Survey

Use open software

- Ince, et al, "The Case for Open Computer Programs," *Nature*, 482, pp. 485-488 (23 Feb. 2012)

- Editorial about *Nature* paper, Kyle Niemeyer "Nature Editorial: If you want reproducible science, the software needs to be open source," Arstechnica.com (2012/02)

- Case for Python, R, Julia

- Case against MatLab, Stata, Eviews, SPSS

Publication strategy

1. Replication is good first step

   - You will make sure you have the tools necessary to extend research
   - You will learn more deeply why they used certain tools and approaches
   - You might find an error (**publication**)
   - You might find a better way to do the same thing (**publication**)

2. You will now be able to better contextualize your extension

3. You will be in good position to invite original author(s) to collaborate

4. By open sourcing your work, you are less threatening, more inviting