Reproducible Research

R. Gentleman

Genentech

A narrow focus

- Given:
  - a data set (which is where most scientific papers start)
  - a description of the analysis that leads to the figures and tables
- Can a reasonably competent individual reproduce the results?

Why?

- *The Statistician (2003)* ***52, Part 4, pp. 423–438 Diversities of gifts, but the same spirit Peter J. Green***
  - Most statistics papers, as published, no longer satisfy the conventional scientific criterion of reproducibility: could a reasonably competent and adequately equipped reader obtain equivalent results if the experiment or analysis were repeated? Typically, the answer is no, both because there is not space to specify sufficient detail and because repetition could involve a huge cost in time and effort in developing computer code to parallel that of the author.

Broader Concepts

- methodological generalizability:
  - If I analyze the given data set in a similar way do I get a similar answer?
- generalizability:
  - If I generate a new data set, similar to the reported one, and analyze it in a similar way will I get the same answer?
- these questions are more central to scientific investigation; but we have not solved the simple problem

RR History

- Claerbout + Donohoe identified many of the issues
- it is important to realize that they were solving two problems
  1. first reproducibility of journal figures (and one presumes tables + other facts)
  2. the ability to hand a project from one person to another in a timely and efficient manner (extensibility)

RR applications

- publishing papers in the scientific literature
- facilitating multi-center analysis and scientific investigation
- enabling project integrity as personnel change
- analyses that are important to a community
  - clinical trials
  - studies of climate change
  - environmental studies (EPA)

RR Bottlenecks

- it is complicated
  - OS changes, component software changes
  - web services or cloud computing etc have all the same software and data provenance issues
- do we mean identical or just similar?
  - eg. not all machine learning algorithms are interchangeable
  - what if they made a mistake?

RR Infrastructure

- data provenance:
  - we can capture the data in an archive (package) that is versioned
  - query databases or on-line repositories – but they would need to have version numbers
- code provenance:
  - software could be captured in the same archive (package)
  - on-line repositories (Sourceforge) with version numbers could also be used
  - potentially problematic due to OS reliance

Authoring

- we need a decent authoring environment that supports a variety of underlying languages
  - funding agencies need to get on board and fund research that will ultimately lead to decent authoring tools
  - these need to be language agnostic
    - users can use R, Python, SAS, Matlab, Perl etc
  - they need to be OS agnostic
  - versioning
  - data storage

Authoring

- each version of the paper to be authored must
  - be generated automatically (with software tools)
  - in its entirety
  - from the component pieces
- we want to avoid
  - cut-and-paste errors
  - updating one part of the paper but not others when the data or code are updated

Authoring

- in the context of the R language Sweave (R + LaTeX) is a reasonable authoring environment
- the concept of compendia
  - R packages that reproduce a paper
- Gentleman and Temple Lang, *Statistical Analyses and Reproducible Research,* JCGS, 2007, outline a general architecture

User Interactions

- how could one interact with a piece of reproducible research
  - rerun it – do you get the same answers
  - dissect (debug?) it – can you see where the answer becomes surprising
  - tweak it – change their algorithms for yours
  - extend it – take some new direction that they did not explore
- if well designed, the output of an authoring tool would support these (and many other) interactions

A less narrow focus

- Given:
  - a data set (which is where most scientific papers start)
  - a description of the analysis that leads to the figures and tables
- Can a reasonably competent individual reproduce the results?
- Can a reasonably competent individual extend the results?

Acknowledgements

- Vincent Carey
- Duncan Temple Lang
- Wolfgang Huber
- Seth Falcon