The Reproducible Research Movement in Statistics

Victoria Stodden
Department of Statistics
Columbia University

59th ISI World Statistics Congress
Sharing Data, Code and Publications - Making Research Reproducible
Hong Kong
Aug 30, 2013
The Credibility Crisis

Standards and Scientific Communication

Support from Funders, Journals, and Employers Essential

Teaching and Tools for Reproducible Research

Recent Policy Steps

Experiences in Reproducible Research
Changes in the Scientific Method

- Branch 1: Deductive/Theory: e.g. mathematics; logic,
- Branch 2: Inductive/Empirical: e.g. hypothesis framework; statistical analysis of controlled experiments,
- Branch 3? 4? Large scale extrapolation and prediction, using simulation and other data-intensive methods.
The traditional published article is inadequate for verification of modern computational results:

- powerful computational hardware
- vast data collection and storage capabilities
- complexity of computational analysis
- digital communication mechanisms, the Internet.
Pervasiveness of Computational Methods

- The Journal of the American Statistical Association:

<table>
<thead>
<tr>
<th>June</th>
<th>Computational Articles</th>
<th>Code Publicly Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>9 of 20</td>
<td>0%</td>
</tr>
<tr>
<td>2006</td>
<td>33 of 35</td>
<td>9%</td>
</tr>
<tr>
<td>2009</td>
<td>32 of 32</td>
<td>16%</td>
</tr>
<tr>
<td>2011</td>
<td>29 of 29</td>
<td>21%</td>
</tr>
</tbody>
</table>

- Social network data / quantitative revolution in social science (Lazer et al. 2009);  
- In traditionally nonquantitative fields: e.g. Wordhoard project at Northwestern and the Stanford Literary Lab, both in English departments
Funding Agency Recommendations

- Software and data should be “open by default” and access only restricted if openness conflicts with other considerations such as confidentiality.
- Add ‘Reproducible Research’ to the list of specific examples that proposals could include in their Broader Impact statements.
- Software and dataset curation should be explicitly included in grant proposals and recognized as a scientific contribution by funding agencies, and funds made available to support it.
Editors and reviewers must insist on the full disclosure of computational details.

Rigorous verification and validity testing should be expected by authors.

Authors need to state intended exceptions for full disclosure, such as for proprietary, medical, or other confidentiality issues, upon submission, and reviewers and editors must agree such exceptions are reasonable.

All software and data used in a paper should be explicitly cited.
Employer Recommendations

- Tenure and promotion committees should reward software and dataset contributions.
- Data and code citation practices should be recognized and expected in computational research.
The skills required to carry out and disseminate reproducible research in the computational sciences should be taught as part of scientific methodology.

Modern programming and software engineering techniques should be taught to all data scientists.

Could be done as a graduate seminar course in Reproducible Research.

Many tools for reproducibility developing in many different fields.
Policy Changes to Support Scientific Reliability

- On Feb 22 the OSTP instructed federal funding agencies to develop plans for enabling public access to both journal articles and digital datasets that arise from federal grants (due Aug 22!).
- On March 5, the Research subcommittee of the House Committee on Science, Space, and Technology convened a hearing on Scientific Integrity & Transparency.
Experience in Geophysics and Statistics

- 1991: Stanford Professor Jon Claerbout requires theses to conform to standard of reproducibility,
- reduces "startup time" for new students from years to weeks,
- his vision adopted and adapted by many others, e.g. Sergey Fomel, David Donoho.
The Credibility Crisis
Standards and Scientific Communication
Support from Funders, Journals, and Employers Essential
Teaching and Tools for Reproducible Research
Recent Policy Steps
Experiences in Reproducible Research

Madagascar (Sergey Fomel and collaborators)
The Credibility Crisis
Standards and Scientific Communication
Support from Funders, Journals, and Employers Essential
Teaching and Tools for Reproducible Research
Recent Policy Steps
Experiences in Reproducible Research

Donoho Research Group, Stanford

“WaveLab (1999)”

“Sparselab (2006)”
Biostatistics

▶ In Jul 2010, 33 biostatisticians send letter to the NIH’s IOM director, Harold Varmus.
▶ Letter requests a study of clinical trial approved based on flawed genomics research.

“We strongly urge that the clinical trials in question ... be suspended until a fully independent review is conducted of both the clinical trials and of the evidence and predictive models being used... For this to happen, sufficiently detailed data and annotation must be made available for review. The data should be sufficiently documented for provenance to be assessed... and the computer code used to predict which drugs are suitable for particular patients must be made available to allow ... analysts to assess its validity and reproducibility...”
Recommendations from the Institute of Medicine

- Recommends new standards for omics-based tests, including a fixed version of the software, expressly for verification purposes.
“The fully specified computational procedures are locked down in the discovery phase and should remain unchanged in all subsequent development steps.”
Grassroots Efforts

Independent efforts by researchers:

- ICERM 2012 “Reproducibility in Computational and Experimental Mathematics”
- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- AMP / ICIAM 2011 “Community Forum on Reproducible Research Policies”
- SIAM Geosciences 2011 “Reproducible and Open Source Software in the Geosciences”
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: “The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer”
- SIAM CSE 2011: “Verifiable, Reproducible Computational Science”
- Yale 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences

Policy changes:

- NSF/OCI report on Grand Challenge Communities (Dec 2010)
- NSF report “Changing the Conduct of Science in the Information Age” (Aug 2011)
- IOM “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials” (2012)
- NIH, NSF multiple requests for input on data policies
- Journal policy movement toward code and data requirements (ie. Science Feb 2011)

...