Reproducible Research, Replicability, and Ethical Practice

Ronald A. Thisted

Departments of Health Studies and Statistics
The University of Chicago

4 August 2014
JSM Boston

I. The roots of statisticians' ethical obligations

- Statisticians produce work that others consume
- The statistician's relationship to these "others" give rise to ethical concerns
- Specifically,
  - what duty do we owe these others?
  - what do we do when those duties conflict? and
  - what practices should we adopt in order to meet these duties and minimize conflicts?
- These duties define what others may expect of us in our professional capacities

To whom do we owe duties (and what do we owe them)?

Good starting point: 1999 ASA Ethical Guidelines for Stat Practice

- Funders, clients, employers
  - Timeliness, correctness, confidentiality, congruence with funders objectives
- Other statisticians
  - Correctness, apply best methods, extend theory and methods, work others can rely on
- The statistics profession
  - "98% of all statistics are made up."
    "Statistics can be made to prove anything"
    —Authors Unknown
  - Duty to prove these wrong by example
  - Promulgate appropriate use of statistical methods

To whom do we owe duties (and what do we owe them)?

. . . Continued

- Collaborators and research colleagues (Our teams)
  - Contributions to joint work are correct, timely
  - Ability to "show our work"—how our conclusions are derived from inputs (clear communication)
  - This applies to statistical *and* non-statistical collaborators
- The general public
  - Public funding $\rightarrow$ public benefit (advanced knowledge, fairness)
  - Protection from direct harm (e.g., research subjects, dose escalation in Phase I cancer trials, credit risk scoring)
  - Broader impacts: environmental standards, drug approval decisions, clinical practice guidelines, use of our methods in substantive research

Interlude toward replicability

Statistical publication emphasizes practices that help ensure aspects of replicability

- Theoretical statistics: Proofs
- Methods research: Arguments, theorems, simulation studies research
- Applications: increasing norms for deposit of data and code

II. Reproducible research practices

- Four practices that can be integrated into statisticians' daily work
- Primarily useful in methods and applications work
- Each depends on computing tools or environments (but free)
- Independently implementable

1. Scripted data management and analysis
2. Version control
3. Integration of exposition and computation
4. Open sources

Scripted data management and analysis I

Problem: Can't get this year's calculations to agree with last year's manuscript

Some data analytic practices risk subsequent uncertainty:

- Drop-down analyses
- Reports cut-and-pasted from interactive sessions
- Statistical output divorced from corresponding inputs

Solution: Structured scripting practices

Scripted data management and analysis II

1. Don't do data management or data cleaning in Excel (use scripts or programs instead)
2. Never modify the original raw data files in any way.
   And make an extra copy of the original data
3. Do all analysis (and data management) from scripts
   Make sure that the final output used for a manuscript is identical to the output from running the master script once from scratch

Product: a series of data files, each derived from its predecessor using a script.
End result: a small, structured collection of files that can reproduce all of my work at any later point in time.

Version control I

Problem: Was that table done before or after outliers were set aside? Was it done with the missing data imputation, or missing items dropped? Was it adjusted for study site, or not? And was that from the data set before or after sub-continent Indians were separated from Native Americans?

- Essential to be able to determine (and return to) the "state of the analysis" at any point in time
- Not enough to just keep a copy of every version of every file
- Need the *relationship* of these files to one another

Solution: Version control of data sets, manuscripts, scripts, and programs

Version control II

Version control options:

- Manual discipline for file naming, saving, and logging
  Example: Scott Long *Workflow* book
- Version control software (with or without GUIs)
  - Subversion
  - Mercurial
  - Git
- Key idea:
  - Work done on files in a *working directory*
  - Recent changes are periodically *committed* to
  - . . . an (invisible) *repository*
  - *WABAC*: the working version can be restored at any time to its state immediately after any commit
- Tools are multi-platform, free software
  (but there is a learning curve)

Integrated exposition and computation I

Problem: "What was I thinking?"

- We need to capture not only the *inputs* and *results* of an analysis, but also the *intended logic*
- There is a need to preserve the thought processes that link sub-analyses and that track further work needed to validate conclusions

Solution: integrate written descriptions of the ideas driving an analysis and the assumptions and intentions that code is intended to implement with the code itself. Simultaneously serves as

- documentation for the computer scripts
- record of the analysis (and/or data cleaning) logic
- explanation of how analysis ideas were implemented

Integrated exposition and computation II

Approaches:

- Manual: disciplined cut-and-paste to assemble
  - narrative
  - code
  - analysis output
- Software: Some examples
  - Sweave (R and LaTeX) $\rightarrow$ pdf
  - statweave (R, Stata, or SAS and LaTeX) $\rightarrow$ pdf
  - knitr (R, SAS, python and LaTeX or markdown) $\rightarrow$ pdf, HTML, Word

Practice: Employ a system that couples intentions with implementation, and that couples statistical results with statistical inputs.

Open sources

Problem: I know what I did, but what did *she* do?
To build on the work of others, we need to know what they actually did, and what data underlie their work.
Making the following freely available increases transparency

- Data set on which analyses are based ("open data")
- Source code used to produce published manuscripts ("open code")
- Packaged software on which analysis relies ("open software")
- Access to the final published manuscript ("open access")

Practice: To the extent possible, maximize the amount of work and work product that is freely and easily available to a broad community.

III. How RR practices relate to our ethical duties

Common themes: correctness, consistency, communication (c^3)
Also: demonstrable reproducibility, long-term efficiency

- Scripted data management and analysis
  - Ensures ability to reproduce an analysis exactly *and* to understand how it was produced. Enhances duties to collaborators and funders, and helps to ensure the general duty of correctness to the profession and the public by permitting critical evaluation and enhanced ability to correct errors or introduce refinements when appropriate.
- Version control
  - Helps to meet our duties of clear communication to collaborators, funders; enhances ability to perform for those to whom a duty of timeliness is owed; enhances effective collaboration. Promotes consistency of reports from the same data and enhances abilities related to the general duty of correctness.

How RR practices relate to our ethical duties II

- Integration of exposition and computation
  - Increases ability to align work with funders needs and expectations, enhance duties to communicate clearly to collaborators, other statisticians, and the public, enhances ability to produce correct results, and for us or others to identify errors and potential improvements.
- Open sources
  - Enhances our duties to the profession and to the general public; enhances collaboration and the public good of making it possible to build on earlier work: data, algorithms, methods, analyses.

Concluding Remarks

- Others rely on the work that we do, so . . .
- . . . we have a duty to them to
  - communicate our work clearly and accurately, and
  - to maximize the likelihood that what we present is correct.
- Reproducible research tools allow for more consistent professional work— not least by helping us communicate (including with our future selves)
- By adopting these practices we also enhance our ability of meet the obligations we owe to employers, clients, collaborators, colleagues, our profession, and the general public.

Thank you!

Ronald Thisted
Department of Health Studies, University of Chicago
email: thisted@health.bsd.uchicago.edu
phone: +1 773.834.1242