Reproducible Research (RR) and *Biostatistics*

Sahir Rai Bhatnagar[1]

January 23, 2014

[1] McGill Biostats Reading Group

Disclaimer

- I will ask you alot of questions
- Your participation is necessary for this to be useful
- Interrupt me often
- This is a ~~reading~~ discussion group

Outline

- Some motivating examples
- The problem
- A solution

What is Science Anyway?

What is Science Anyway?

According to the American Physical Society:

*Science is the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into **testable** laws and theories. The success and credibility of science are anchored in the **willingness** of scientists to **expose their ideas** and results to **independent testing** and **replication** by other scientists*

A Minimum Standard to Verify Scientific Findings

A Minimum Standard to Verify Scientific Findings

Reproducible Research in Computational Sciences

The data and the code used to make a finding are available and they are sufficient for an independent researcher to recreate the finding

For Science

For Science

1. Findings cannot be considered genuine contributions until verified through independent replication (whenever possible)

For Science

1. Findings cannot be considered genuine contributions until verified through independent replication (whenever possible)
    - *"Don't worry, the car runs perfectly... Give me $10k, and I give you my word"*

For Science

1. Findings cannot be considered genuine contributions until verified through independent replication (whenever possible)
   - *"Don't worry, the car runs perfectly... Give me $10k, and I give you my word"*

2. Enables the cumulative growth of future scientific knowledge

For Science

1. Findings cannot be considered genuine contributions until verified through independent replication (whenever possible)
    - *"Don't worry, the car runs perfectly... Give me $10k, and I give you my word"*
2. Enables the cumulative growth of future scientific knowledge
    - *Stop wasting public funds on something that has already been done*

For You

For You

1. **Better work habits**
   - *Who cares if no one else is watching?*

For You

1. **Better work habits**
   - *Who cares if no one else is watching?*

2. **Better teamwork**
   - *Bring current and future collaborators upto speed with ease*

For You

1. **Better work habits**
   - *Who cares if no one else is watching?*
2. **Better teamwork**
   - *Bring current and future collaborators upto speed with ease*
3. **Changes are easier**
   - *No research process is linear*

For You

1. **Better work habits**
   - *Who cares if no one else is watching?*
2. **Better teamwork**
   - *Bring current and future collaborators upto speed with ease*
3. **Changes are easier**
   - *No research process is linear*
4. **Higher research impact**
   - *Others more willing to read, learn, build and cite*

How did they get those numbers?

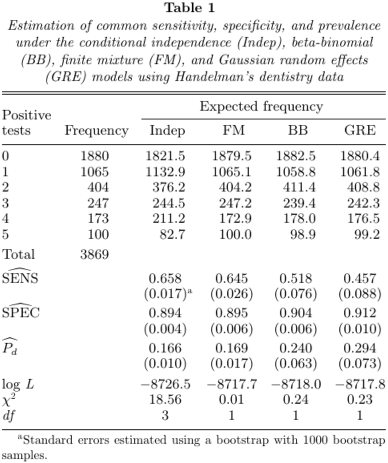


Table 1

Estimation of common sensitivity, specificity, and prevalence under the conditional independence (Indep), beta-binomial (BB), finite mixture (FM), and Gaussian random effects (GRE) models using Handelman's dentistry data

Positive tests	Frequency	Expected frequency			
		Indep	FM	BB	GRE
0	1880	1821.5	1879.5	1882.5	1880.4
1	1065	1132.9	1065.1	1058.8	1061.8
2	404	376.2	404.2	411.4	408.8
3	247	244.5	247.2	239.4	242.3
4	173	211.2	172.9	178.0	176.5
5	100	82.7	100.0	98.9	99.2
Total	3869				
$\widehat{\text{SENS}}$		0.658	0.645	0.518	0.457
		(0.017)[a]	(0.026)	(0.076)	(0.088)
$\widehat{\text{SPEC}}$		0.894	0.895	0.904	0.912
		(0.004)	(0.006)	(0.006)	(0.010)
$\widehat{P_d}$		0.166	0.169	0.240	0.294
		(0.010)	(0.017)	(0.063)	(0.073)
log L		−8726.5	−8717.7	−8718.0	−8717.8
χ^2		18.56	0.01	0.24	0.23
df		3	1	1	1

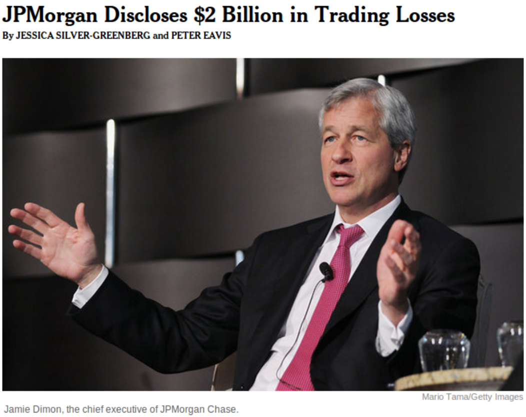
[a] Standard errors estimated using a bootstrap with 1000 bootstrap samples.

Figure 1: Paper presented by Maarten Van Smeden on latent class models.

The Secret Statistical Society

Figure 2: Illustration of Marie-Pierre's dilemma

Blame Copy Paste...Not Greed



JPMorgan Discloses $2 Billion in Trading Losses

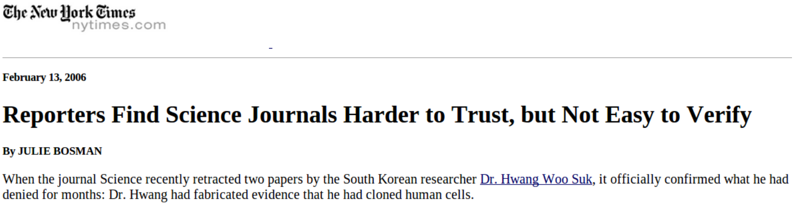
By JESSICA SILVER-GREENBERG and PETER EAVIS

Mario Tama/Getty Images

Jamie Dimon, the chief executive of JPMorgan Chase.

Figure 3: The hedging strategy operated through a series of Excel spreadsheets, which had to be completed manually, by a process of copying and pasting data from one spreadsheet to another

Fabricating data



The New York Times
nytimes.com

February 13, 2006

Reporters Find Science Journals Harder to Trust, but Not Easy to Verify

By JULIE BOSMAN

When the journal Science recently retracted two papers by the South Korean researcher Dr. Hwang Woo Suk, it officially confirmed what he had denied for months: Dr. Hwang had fabricated evidence that he had cloned human cells.

Figure 4: Convicted of falsifying his papers and embezzling government research funds. A judge sentenced him to a suspended two-year prison term.

Recap

What are the issues here?

Recap

What are the issues here?

1. Non-disclosure of ...
2. Not a requirement for journal submission
3. Copy-paste and GUI interaction
4. Lack of tools

Recap

What are the issues here?

1. Non-disclosure of ...
2. Not a requirement for journal submission
3. Copy-paste and GUI interaction
4. Lack of tools

How can we improve the situation?

1. Shift towards open source (e.g. R, LaTeX)
2. New policies on reproducibility requirements
3. User friendly tools

A powerful Typesetting system

```
A \textbf{bold
\textit{Hello \LaTeX}}
to start!
```

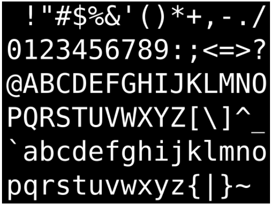
A **bold *Hello***
LaTeX to start !

```
Odds=$\left(\frac{\pi}{1-\pi}
\right)$
```

$$\text{Odds}=\left(\frac{\pi}{1-\pi}\right)$$

1. Input for LaTeX is composed in plain ASCII using a text editor
2. Although Word is useful for writing very short and simple documents, it becomes too complex or even unusable for more complicated tasks
3. Commonly needed features, like user-customized automated numbering or various automated indexes, cannot be created using Word at all
4. LaTeX does require more effort and time to learn to use even for simpler tasks, but once learned, difficult tasks can be accomplished rather easily and straightforwardly

What is ASCII?



```
 !"#$%&'()*+,-./
0123456789:;<=>?
@ABCDEFGHIJKLMNO
PQRSTUVWXYZ[\]^_
`abcdefghijklmno
pqrstuvwxyz{|}~
```

Figure 5: 95 printable ASCII characters, numbered 32 to 126. (0 to 31 & 127 are non-printing control characters)

1. When you save your document, it is saved in the form of plain text i.e in "ASCII" (the American Standard Code for Information Interchange)
2. ASCII is composed of 128 (2^7) characters: 7 binary digits for its encoding (Fig. 5)
3. An ASCII message will be understandable by any computer in the world. If you send such a message, you can be sure that the recipient will see precisely what you typed

Comparison

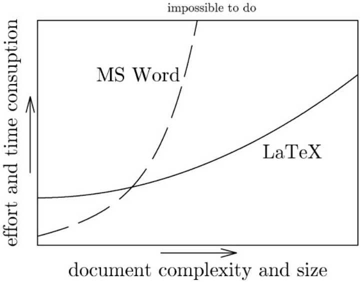


Figure 6: Comparison

- LATEX has a greater learning curve
- Many tasks are very tedious or impossible (most cases) to do in MS Word or Libre Office

The Philosophy behind LATEX

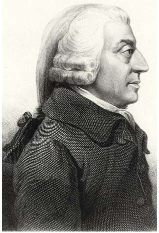


Figure 7: Adam Smith, author of *The Wealth of Nations* (1776), in which he conceptualizes the notion of the division of labour

Division of Labour

Composition and logical structuring of text is the author's specific contribution to the production of a printed text. Matters such as the choice of the font family, should section headings be in bold face or small capitals? Should they be flush left or centered? Should the text be justified or not? Should the notes appear at the foot of the page or at the end? Should the text be set in one column or two? and so on, is the typesetter's business

The Genius Behind LaTeX

Figure 8: Donald TeXproject was started in 1978 by Donald Knuth (Stanford). He planned for 6 months, but it took him nearly 10 years to complete. Coined the term “Literate programming”: mixture of code and text segments that are “human” readable. Recipient of the Turing Award (1974) and the Kyoto Prize (1996).

An Open Source Statistical Software Program

Figure 9: R logo

- You interact with R by explicitly writing down your steps as code
- You cannot run analysis by clicking on dropdown menus
- Promotes reproducibility (CRAN task view)
- **Open Source!**

How to include a Figure in a LATEX document

The Tedious Way

```
in R:
pdf("~/cars.pdf")
plot(mtcars[ , c("disp","mpg")])
fit <- lm(mpg ~ disp , data = mtcars)
abline(fit, lwd=2)
dev.off()

then in LaTeX
\begin{figure}[h!]
\centering
\includegraphics[]{./simple}
\caption{Simple linear regression}
\label{fig:simple}
\end{figure}
```

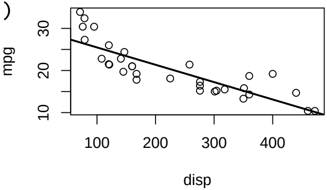


Figure 10: Simple linear regression

How to include a Figure in a LATEX document

- What if the dataset changes?
- What if one observation was wrong?

How to include a Figure in a LATEX document

The Dynamic Way

```
'<<fig.cap='Linear regression'>>=
plot(mtcars[ , c("disp","mpg")])
fit <- lm(mpg ~ disp , data = mtcars)
abline(fit, lwd=2)
'@
```

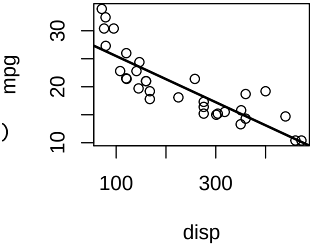


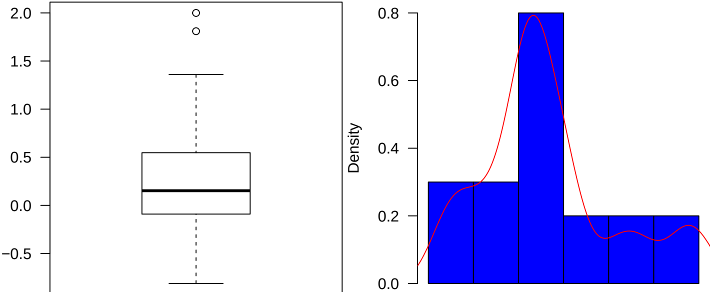
Figure 11: Linear regression

R + LaTeX= knitr (Yihui Xie (2013))

```
(x = rnorm(20))  # create some random numbers

##  [1]  0.14496  0.43832  0.15319  1.08494  1.99954 -0.81188
##  [7]  0.16027  0.58589  0.36009 -0.02531  0.15088  0.11008
## [13]  1.35968 -0.32699 -0.71638  1.80977  0.50840 -0.52746
## [19]  0.13272 -0.15594

boxplot(x)
hist(x, main = "", col = "blue", probability = TRUE)
lines(density(x), col = "red")
```



The possibilities are endless

Pros

- Highly customizable for repetitive tasks
- Easily extendible to Markdown documents (Gruber 2004)
- Interactive presentations via `Slidify` (Vaidyanathan 2013)
- Interactive web applications to present results
- Avoids error prone copy-paste
- Ensures reproducibility
- Allows for caching (think big data)
- You can focus more time on methods and analysis

The possibilities are endless

Pros

- Highly customizable for repetitive tasks
- Easily extendible to Markdown documents (Gruber 2004)
- Interactive presentations via `Slidify` (Vaidyanathan 2013)
- Interactive web applications to present results
- Avoids error prone copy-paste
- Ensures reproducibility
- Allows for caching (think big data)
- You can focus more time on methods and analysis

Cons

- Brute force brings us instant gratification

RR Workflow

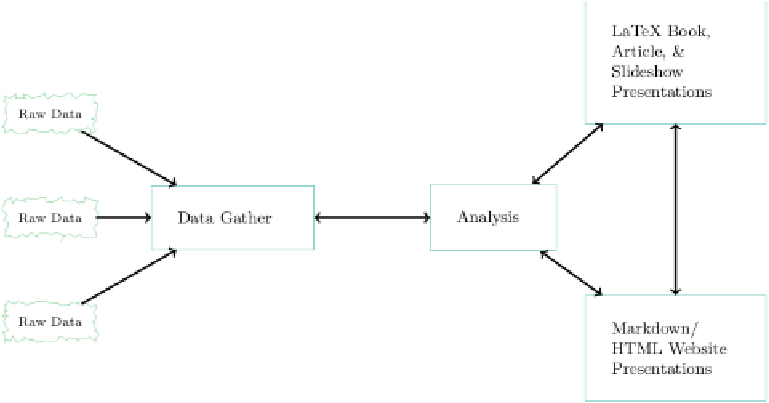

Raw Data
Raw Data
Raw Data
Data Gather
Analysis
LaTeX Book, Article, & Slideshow Presentations
Markdown/ HTML Website Presentations

Figure 12: An example workflow. Notice the direction of the arrows. (*Gandrud 2014*)

A Motivating Quote

"It's week 3... So it must be binomial." - J.A. Hanley

Storing Your Files in the Cloud: GitHub

What is GitHub?

- An interface and a cloud hosting service built on top of the Git version control system
- Git does the version control
- GitHub allows you to store the data remotely

Storing Your Files in the Cloud: GitHub

Why use GitHub?

1. **Storage and Access**
    - Makes projects accessible on a fully featured website
    - Can create and host a website to present results

Storing Your Files in the Cloud: GitHub

Why use GitHub?

1. **Storage and Access**
   - Makes projects accessible on a fully featured website
   - Can create and host a website to present results

2. **Collaboration**
   - Keeps meticulous records of who contributed what to a project
   - "Issues" tracker
   - Each project can host a wiki
   - Anyone can suggest changes to files in a public repository

Storing Your Files in the Cloud: GitHub

Why use GitHub?

1. **Storage and Access**
    - Makes projects accessible on a fully featured website
    - Can create and host a website to present results

2. **Collaboration**
    - Keeps meticulous records of who contributed what to a project
    - "Issues" tracker
    - Each project can host a wiki
    - Anyone can suggest changes to files in a public repository

3. **Version Control**
    - Can easily revert back to any change you make
    - Previous file versions in Dropbox disappear after 30 days. GitHub stores them indefinetly
    - Identifies difference between two documents and lets you reconcile them

Storing Your Files in the Cloud: GitHub

The main point here is to avoid:

manuscript_v1.2.3_July_2013_sahir.tex

or

data_analysis_and_cleaning_v2.R

Open Source

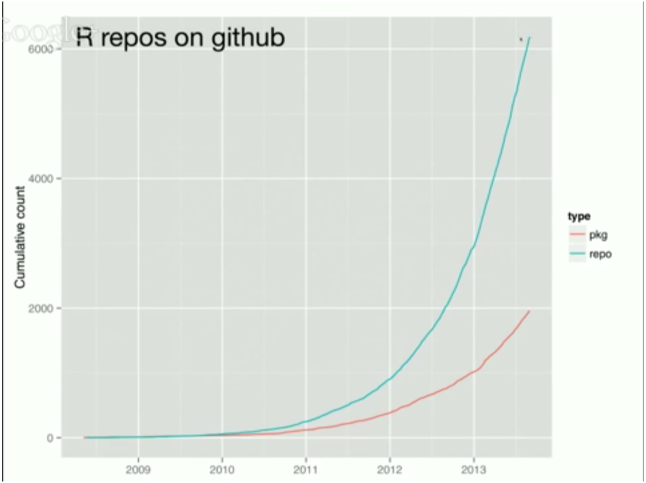
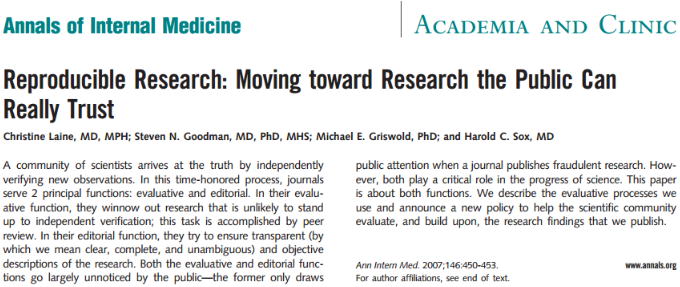


Figure 13: R projects and packages hosted on GitHub (*Wickham 2013*)

Medicine



Annals of Internal Medicine | Academia and Clinic

Reproducible Research: Moving toward Research the Public Can Really Trust

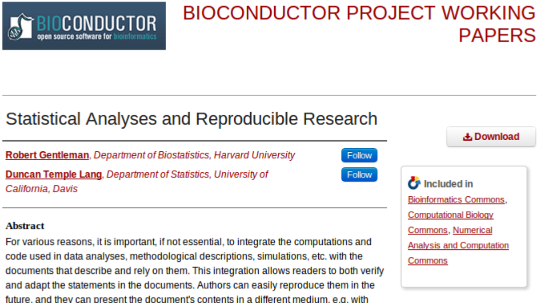
Christine Laine, MD, MPH; Steven N. Goodman, MD, PhD, MHS; Michael E. Griswold, PhD; and Harold C. Sox, MD

A community of scientists arrives at the truth by independently verifying new observations. In this time-honored process, journals serve 2 principal functions: evaluative and editorial. In their evaluative function, they winnow out research that is unlikely to stand up to independent verification; this task is accomplished by peer review. In their editorial function, they try to ensure transparent (by which we mean clear, complete, and unambiguous) and objective descriptions of the research. Both the evaluative and editorial functions go largely unnoticed by the public—the former only draws public attention when a journal publishes fraudulent research. However, both play a critical role in the progress of science. This paper is about both functions. We describe the evaluative processes we use and announce a new policy to help the scientific community evaluate, and build upon, the research findings that we publish.

Ann Intern Med. 2007;146:450-453. www.annals.org
For author affiliations, see end of text.

Figure 14: Annals of Internal Medicine (*Liane et al. 2007*)

Bioconductor



BIOCONDUCTOR PROJECT WORKING PAPERS

Statistical Analyses and Reproducible Research

Robert Gentleman, *Department of Biostatistics, Harvard University*

Duncan Temple Lang, *Department of Statistics, University of California, Davis*

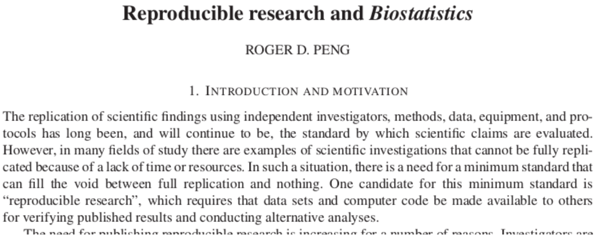
Included in
Bioinformatics Commons, Computational Biology Commons, Numerical Analysis and Computation Commons

Abstract

For various reasons, it is important, if not essential, to integrate the computations and code used in data analyses, methodological descriptions, simulations, etc. with the documents that describe and rely on them. This integration allows readers to both verify and adapt the statements in the documents. Authors can easily reproduce them in the future, and they can present the document's contents in a different medium, e.g. with

Figure 15: Bioconductor (*Gentleman and Lang 2004*)

Biostatistics



Reproducible research and *Biostatistics*

ROGER D. PENG

1. Introduction and motivation

The replication of scientific findings using independent investigators, methods, data, equipment, and protocols has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study there are examples of scientific investigations that cannot be fully replicated because of a lack of time or resources. In such a situation, there is a need for a minimum standard that can fill the void between full replication and nothing. One candidate for this minimum standard is "reproducible research", which requires that data sets and computer code be made available to others for verifying published results and conducting alternative analyses.

The need for publishing reproducible research is increasing for a number of reasons. Investigators are

Figure 16: Biostatistics (*Peng 2009*)

CRAN has a dedicated Task View for RR

CRAN Task Views

Biostatistics requirements for RR

1. data analysis script
2. other code
3. data
4. script for results used in paper
5. `knitr` file (`.Rnw`)
6. resulting `.tex` file from compiling with `knitr`
7. bibTEXfile

The Main Idea

Jon Claerbout, Geophysicist at Stanford, (1995)

*"An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the **complete software development environment** and the **complete set of instructions** which generated the figures"*

If you can only take away one thing from today's discussion…

$$\text{Reproducibility} \propto \frac{1}{\text{copy paste}}$$

References I

- Christopher Gandrud, *Reproducible research with r and rstudio*, Chapman and Hall-CRC The R Series, 2013.
- David Smith, *Did an excel error bring down the london whale?*, http://blog.revolutionanalytics.com/2013/02/did-an-excel-error-bring-down-the-london-whale.html.
- C. Laine, S. N. Goodman, M. E. Griswold, and H. C. Sox, *Reproducible research: moving toward research the public can really trust*, Ann. Intern. Med. **146** (2007), no. 6, 450–453.
- New York Times, *Reporters find science journals harder to trust, but not easy to verify*, http://www.nytimes.com/2006/02/13/business/media/13journal.html?_r=0&adxnnl=1&pagewanted=all&adxnnlx=1390399611-aqm52MhkXkIFF7Azx7irCg.

References II

- R. D. Peng, *Reproducible research and Biostatistics*, Biostatistics **10** (2009), no. 3, 405–408.
- Sergey Fomel and Jon F. Claerbout, *Guest editor's introduction: Reproducible research*, Computing in Science and Engineering (Jan/Feb 2009).
- Yihui Xie, *Dynamic documents with r and knitr*, Chapman and Hall-CRC The R Series, 2013.