REPRODUCIBILITY

Stats and R workshop 23rd-24th November 2016

Susan Jarvis

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NERC SCIENCE OF THE ENVIRONMENT

SCIENCE

A Sharp Rise in Retractions Prompts Calls for Reform

By CARL ZIMMER APRIL 16, 2012

The reproducibility crisis in science

A statistical counterattack

More people have more access to data than ever before. But a comparative lack of analytical skills has resulted in scientific findings that are neither replicable nor reproducible. It is time to invest in statistics education, says **Roger Peng**

nature international weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

News & Comment › News › 2016 › November › Article

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Rights & Permissions

Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.

The studies they took on ranged from whether expressing insecurities perpetuates them to

Brian Nosek's team set out to replicate scores studies.

LETTERS

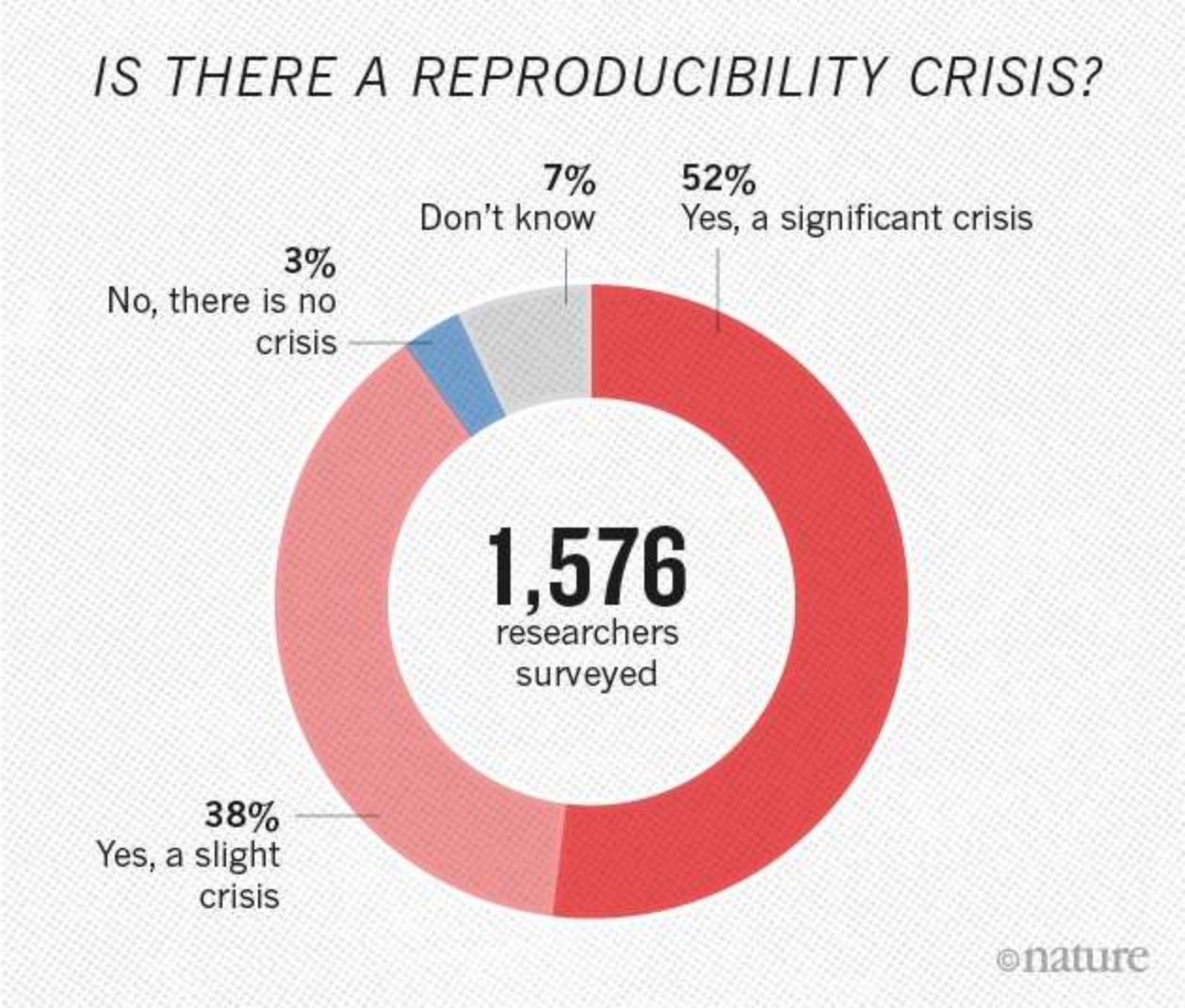
Reproducibility in ecological research

Xiaolei Huang

+ Author Affiliations

E-mail: huangxl@fafu.edu.cn

Science 12 Dec 2014:
Vol. 346, Issue 6215, pp. 1307
DOI: 10.1126/science.346.6215.1307-c


IS THERE A REPRODUCIBILITY CRISIS?
7%
Don't know
52%
Yes, a significant crisis
3%
No, there is no crisis
1,576
researchers surveyed
38%
Yes, a slight crisis
©nature

What is reproducibility?

Reproducibility: same data, same result

Replication: new data, same result

Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226–1227. doi:10.1126/science.1213847

Isn’t all research reproducible?

Isn’t all research reproducible?

- Couldn’t remember which Excel sheet had the final data

Isn’t all research reproducible?

- Couldn’t remember which Excel sheet had the final data
- Couldn’t remember which script was the one used for the paper

Isn’t all research reproducible?

- Couldn’t remember which Excel sheet had the final data
- Couldn’t remember which script was the one used for the paper
- Couldn’t remember which order to run programs/scripts

Isn’t all research reproducible?

- Couldn’t remember which Excel sheet had the final data
- Couldn’t remember which script was the one used for the paper
- Couldn’t remember which order to run programs/scripts
- Data downloaded from internet seems to have changed

Isn’t all research reproducible?

- Couldn’t remember which Excel sheet had the final data
- Couldn’t remember which script was the one used for the paper
- Couldn’t remember which order to run programs/scripts
- Data downloaded from internet seems to have changed
- Ran the same code, got a different answer

Why be reproducible?

Personal

- Confidence in results
- Faster to repeat
- Return to project in future
- Faster to write
- Easier for reviewers

Why be reproducible?

Projects

- Multiple people can run analyses – efficient, resilient
- Auditability
- Avoid data forensics

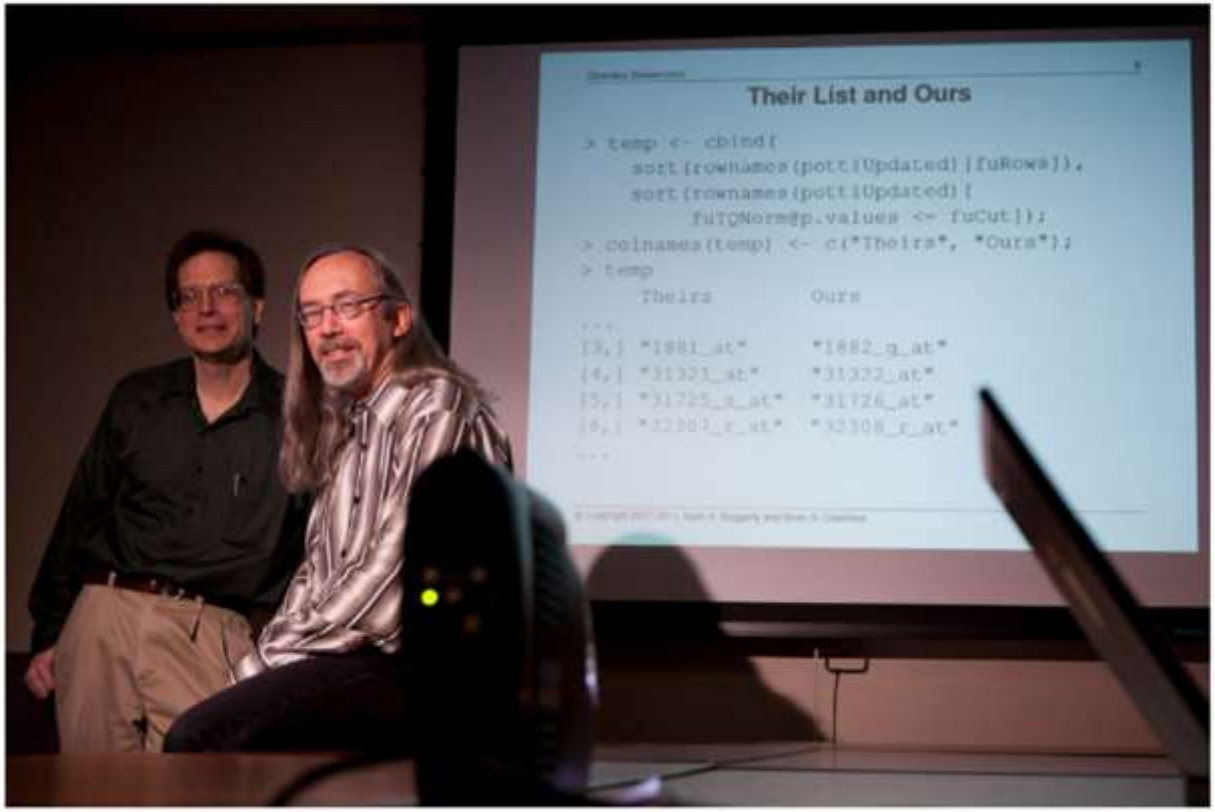
Why be reproducible?

Organisation

- Data/code standards
- MacPherson review
- Reputation

How Bright Promise in Cancer Testing Fell Apart

By GINA KOLATA JULY 7, 2011



Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors. Michael Stravato for The New York Times

RECENT COM

texas ta July 9, 2
This story makes
as well. As a gradu
flawed analyses w

tulipsinyard July
A mildly different
mathematics and
America, I remain

Peter Melzer July
The widely-publici
research for novel
the research comm

SEE ALL COMMENTS

CEH
NATURAL ENVIRONMENT RESEARCH COUNCIL

NCE OF THE
RONMENT

Cons (?)

- Data ‘theft’ or misuse
  - Embargos
  - Licensing
- Resources
  - Can be time consuming
  - Pay off can be huge – time invested now vs time saved not having to retract a paper….

Key elements

Analysis can be re-run with same result:

- Data can be identified
- Code is documented
- Workflow documented
- Changes are documented

Key elements

Analysis can be re-run with same result:

- ~~Data can be identified~~ EIDC, data DOIs
- Code is documented
- Workflow documented
- Changes are documented

Methods section not detailed enough

Steps to reproducibility

Steps to reproducibility

THINK BEFORE YOU CLICK!!

1. Script where possible

- Scripting is greatest tool for reproducibility
- Using R? Already doing this!
- Where you can't script, write steps in extreme detail OR save program logs e.g. SAS

2. Script EVERYTHING

- Data preparation MORE important than analysis
- Excel generates errors
  - Manual manipulation
  - No log
  - Untidy data hard to understand (coloured fields, gaps in data, means in the same worksheet)
- Use spreadsheet software for data input only

3. Document your code

- Comment your code – don’t assume you will understand it in a year
- Embed code in Markdown document – write up as you go

4. Document your workflow

- Where are the relevant datasets/scripts?
  - RProjects
- Which scripts are run first?
  - Number scripts in run order
  - Write workflow in readme file
- Can it be automated?
  - Use “source” to call one script from another
  - Use workflow software
  - Integrate programs e.g. call Python from R

5. Use version control

- Which version of the script was used for final results?
- Version control!
  - Encourages keeping files together in a repository
  - Tracks all the changes you make
  - Prevents multiple copies of files with “v1.3” etc
  - Allows sharing with collaborators and even externally (e.g. Github) – no more emailing different versions

Do I need to code to be reproducible?

- No, some software e.g. SAS, SPSS records a log of all operations which can be saved as a record of analyses conducted. But rarely used, hard to document/read.
- But…coding makes it easier and SO MUCH FASTER
- R can be used for most things:
  - Data cleaning
  - SQL
  - Spatial analysis

Does data & code have to be open to be reproducible?

- No – in many cases full openness is not possible due to licensing restrictions and/or confidentiality
- But – data and code should be accessible i.e. someone with the appropriate license should be able to get access
- Access should be long term – could you repeat in 5/10 years?

Steps to reproducibility

1. Script where possible
2. Script everything
3. Document your code
4. Document your workflow
5. Use version control

Steps to reproducibility

1. Script where possible
2. Script everything
3. Document your code
4. Document your workflow
5. Use version control
6. Document your environment
7. Test your code
8. Review and audit

6. Document your environment

- Important to note package versions (packrat), R version, OS etc
- Best practise – virtual environment/container BUT steep learning curve
- Future likely to involve cloud computing solutions to this problem

7. Test your code

- Manual inspection of results
  - Are patterns sensible
  - Are units sensible
- Re-run subsets of data (e.g. cross validation for models)
- Simulated data set – check calculations are correct
- Write checks into code – especially if multiple users/time consuming/contributing to R package

8. Review and audit

- External review – manuscript not sufficient, code rarely checked
- Internal review of code possible
- If project is audited then all scripts/data/workflows should be available

CEH view

- MacPherson review – standards for model QA
- Software development – version control, virtual environments, testing
- Look to future – even if requirements are not there now they may be in 3 years (data/code archiving, version control…)

Journals view

- Journals now tend to require data to be archived before publication – e.g. in EIDC
- Move towards requiring code to be archived as well

Practical session

A reproducible analysis!

1. Script where possible - **R**
2. Script everything - **data import, plots…**
3. Document your code - **RMarkdown**
4. Document your workflow - **RProject**
5. Use version control - **git**
6. Document your environment
7. Test your code
8. Review and audit

Requirements

- RStudio – www.rstudio.com

- git (will be configured in session)